

Intelligent Systems Reference Library 260

Witold Pedrycz
Gilberto Rivera
Eduardo Fernández
Gustavo Javier Meschino *Editors*

Artificial Intelligence in Prescriptive Analytics


Innovations in Decision Analysis,
Intelligent Optimization, and
Data-Driven Decisions

 Springer

Intelligent Systems Reference Library

Volume 260

Series Editors

Janusz Kacprzyk , Polish Academy of Sciences, Warsaw, Poland

Lakhmi C. Jain, KES International, Shoreham-by-Sea, UK

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included. The list of topics spans all the areas of modern intelligent systems such as: Ambient intelligence, Computational intelligence, Social intelligence, Computational neuroscience, Artificial life, Virtual society, Cognitive systems, DNA and immunity-based systems, e-Learning and teaching, Human-centred computing and Machine ethics, Intelligent control, Intelligent data analysis, Knowledge-based paradigms, Knowledge management, Intelligent agents, Intelligent decision making, Intelligent network security, Interactive entertainment, Learning paradigms, Recommender systems, Robotics and Mechatronics including human-machine teaming, Self-organizing and adaptive systems, Soft computing including Neural systems, Fuzzy systems, Evolutionary computing and the Fusion of these paradigms, Perception and Vision, Web intelligence and Multimedia.

Indexed by SCOPUS, DBLP, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.


Witold Pedrycz · Gilberto Rivera ·
Eduardo Fernández · Gustavo Javier Meschino
Editors


Artificial Intelligence in Prescriptive Analytics


Innovations in Decision Analysis, Intelligent
Optimization, and Data-Driven Decisions


 Springer

Editors

Witold Pedrycz 
Department of Electrical and Computer
Engineering
Faculty of Engineering
University of Alberta
Edmonton, AB, Canada

Eduardo Fernández 
Centro de Investigación para el Desarrollo
Sostenible e Innovación Empresarial
Universidad Autónoma de Coahuila
Torreón, Coahuila, Mexico

Gilberto Rivera 
División Multidisciplinaria de Ciudad
Universitaria
Universidad Autónoma de Ciudad Juárez
Ciudad Juárez, Chihuahua, Mexico

Gustavo Javier Meschino 
Facultad de Ingeniería
Universidad Nacional de Mar del Plata
Mar del Plata, Buenos Aires, Argentina

ISSN 1868-4394

ISSN 1868-4408 (electronic)

Intelligent Systems Reference Library

ISBN 978-3-031-66730-5

ISBN 978-3-031-66731-2 (eBook)

<https://doi.org/10.1007/978-3-031-66731-2>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

Artificial Intelligence in Prescriptive Analytics: Innovations in Decision Analysis, Intelligent Optimization, and Data-Driven Decisions emerges as a resource for researchers, practitioners, and students interested in the rapidly evolving intersection of artificial intelligence (AI) and prescriptive analytics when data-driven decisions are required. This requirement is becoming more prevalent daily, so this book intends to cover this topic by providing many real-world applications. This volume represents a culmination of cutting-edge research and practical insights, presenting the transformative power of AI in reshaping decision-making processes across industries and disciplines.

This book is the product of a collaborative effort within the Eureka Community, an international and multidisciplinary network of scholars and professionals dedicated to advancing the fields of data and business analytics. Founded in 2008, Eureka has created an ecosystem of collaboration, joining researchers from diverse backgrounds to tackle complex challenges and motivate innovation. With over 60 research groups spanning more than 20 countries, Eureka has been a driving force in building knowledge and expanding research capabilities.

Within these pages, you will encounter a lot of contributions that reflect different impacts of AI on prescriptive analytics, data-driven decisions, optimization strategies, and decision-making frameworks. The chapters include a wide range of approaches, models, and algorithms, including expert systems, metaheuristic algorithms, recommender systems, decision support systems, forecasting techniques, and machine learning. This diversity of perspectives ensures a comprehensive exploration of the field, from both theoretical interests and practical applications, through comprehensive literature reviews, representative applications, and challenging case studies.

The different approaches aim to motivate readers to embrace these technologies and integrate them into their own businesses and industries, contributing to the adoption of Industry 4.0 principles, where AI-driven prescriptive analytics plays a protagonist role in driving efficiency, innovation, and sustainable growth.

Each chapter has been subjected to a double-blind peer-review process, ensuring the quality and validity of the research presented. This evaluation process, conducted

by experts from the Eureka Community and beyond, guarantees that the content within this volume is both reliable and significant.

The book opens with “Part I: Decision Analysis,” analyzing the power of AI to enhance traditional decision-making processes. Novel methodologies like multicriteria ordinal classification are applied to real-world financial data, demonstrating the potential for improved resource management. We also see AI’s impact on infrastructure planning, investment selection, and higher education management, offering readers a quick overview of how widely this is applied.

“Part II: Intelligent Optimization” explores the AI’s optimization capabilities. From internet shopping experiences to searching for optimizing delivery routes with drones, AI algorithms show their ability to tackle intricate problems with precision and adaptability. The automation of complex tasks, such as topic generation for government requests, highlights the potential for AI to enhance efficiency and transparency in administrative processes.

The concluding part, “Part III: Data-Driven Decision,” exemplifies the convergence of AI and data analysis for informed decision-making. In this section, authors examine diverse applications, from predicting the impact of sustainability practices on family businesses to harnessing learning analytics for improved reading comprehension. AI’s potential in financial predictions, garment fit classification, and even stress analysis of engineering structures emphasizes its versatility and adaptability across domains.

These three parts collectively make evident the impact of AI on prescriptive analytics, offering a glimpse into how the current data-driven insights and intelligent optimization converge to empower decision-makers.

We sincerely thank the numerous contributors for finally getting this book. Besides, the Eureka Community deserves recognition for its support and commitment to enhancing collaborative research and knowledge exchange.

Finally, we would like to express our sincere appreciation to all those who have supported this project with encouragement, feedback, and enthusiasm. All of this made it possible to finally bring this book to completion. We hope that *Artificial Intelligence in Prescriptive Analytics* will be a valuable resource and a source of inspiration for further innovation in this dynamic and ever-evolving field.

Edmonton, Canada
Ciudad Juárez, Mexico
Torreón, Mexico
Mar del Plata, Argentina

Dr. Witold Pedrycz
Dr. Gilberto Rivera
Dr. Eduardo Fernández
Dr. Gustavo Javier Meschino

Contents

Part I Decision Analysis

1	A New Methodology Based on Multicriteria Ordinal Classification for the Management of Financial Resources with Application to Real Data from the Stock Market	3
	Efrain Solares, Eduardo Fernández, Eyrán Díaz-Gurrola, Reimundo Moreno-Cepeda, Emmanuel Contreras-Medina, and Edy López Cervantes	
2	M-SALD: A Novel Support Tool for Geo-driven Decision-Making in the Selection of Hydraulic Structures Location	23
	Solangel Rodríguez-Vázquez, Yeleny Zulueta Véliz, and Yamilis Fernández Pérez	
3	Multicriteria Hierarchical Ranking for Investment Selection in Latin American Countries	51
	Manuel Muñoz Palma, Pavel Anselmo Álvarez Carrillo, Eva Luz Miranda Espinoza, Francisco Vargas Serrano, and Ernesto León-Castro	
4	Increasing Performance and Competitiveness in HEIs Applying an ICT Quality KPI Model Analyzed with AHP Method	71
	David Lerma-Ledezma, Georgina Castillo-Valdez, Claudia Gómez-Santillán, and Manuel Paz-Robles	
5	A Multicriteria Model for the Selection of Online Travel Agencies	115
	Jesus Jaime Solano Noriega, Juan Bernal, and Juan Carlos Leyva Lopez	

6 A New Methodology Based on Multicriteria Ordinal Classification for the Management of Financial Resources with Application to Real Data from the Stock Market 137
 Efraín Solares, Eduardo Fernández, Eyrán Díaz-Gurrola, Reimundo Moreno-Cepeda, Emmanuel Contreras-Medina, and Edy López Cervantes

Part II Intelligent Optimization

7 Warm Starting Integer Programming for the Internet SHopping Optimization Problem with Multiple Item Units (ISHOP-U) 159
 Fernando Ornelas, Alejandro Santiago, José Antonio Castan Rocha, Salvador Ibarra Martínez, and Alejandro H. García

8 Hybrid Genetic Algorithm Based on Machine Learning and Fitness Function Estimation Proposal for Ground Vehicle and Drone Cooperative Delivery Problem 177
 Muhammed Mirac Özer

9 Automation of Topic Generation in Government Information Requests in Mexico 217
 Hermelando Cruz-Pérez, Alejandro Molina-Villegas, and Edwin Aldana-Bobadilla

10 Analysis of Accuracy on Data Visualization Techniques for Multi-objective Algorithm Performance Based on Convergence and Diversity Towards the Pareto Frontier 251
 Manuel Paz-Robles, Claudia Gomez-Santillan, Nelson Rangel-Valdez, Ma. Lucila Morales-Rodríguez, and Georgina Castillo-Valdez

11 Enhancing Supply Chain Management: A Hybrid Approach for Smart Decisions and Performance 281
 Sandra Rodríguez-Figueroa, Liliana Ramos-Guerrero, Efraín Solares-Lachica, and Alberto Aguila-Tovar

12 Attribute Weighting Model for Breast Cancer Prediction with the Harmony Search Algorithm 299
 Clara Antonio-Hernández, Jesús D. Terán-Villanueva, José A. Castán-Rocha, Mirna P. Ponce-Flores, and Zurisadai Ponce-Flores

Part III Data-Driven Decision

13 Solidary Family Business, Intellectual Property and Sustainability in Rural Producers in Mexico: A Hybrid SEM-PLS and Fuzzy Approach 327
 Miguel Reyna-Castillo, Alejandro Santiago, Xóchitl Barrios-del-Angel, and Daniel Bucio-Gutierrez

14 Learning Analytics in Reading Comprehension 343
 Maritza Bustos-López, Isaac Machorro-Cano, Giner Alor-Hernández, Jonathan Hernández-Capistran, and José Oscar Olmedo-Aguirre

15 Improving Home Loan Predictions: A Fusion of PCA, Decision Tree and Random Forest Approaches 375
 S. S. Thakur, Soma Bandyopadhyay, Sudip Kumar Bera, and Mahika Thakur

16 Classification of Upper Body Fits Using Fit Models 405
 Juan Carlos Leyva López, Otto Alvarado Guerra, Itzel Juárez Sánchez, and Raúl Oramas Bustillos

17 Regression Models for Estimating the Stress Concentration Factor of Rectangular Plates 435
 J. Alfredo Ramírez Monares and Rogelio Florencia Juárez

18 A Performance Analysis of Technical Indicators on the Indian Stock Market 457
 Hetvi Waghela, Jaydip Sen, and Sneha Rakshit

19 PD-Monitor: A Self-management App for Monitoring Patients with Parkinson’s Disease 503
 Giner Alor-Hernández, Laura-Nely Sánchez-Morales, Francisco-Javier García-Dimas, Nancy-Aracely Cruz-Ramos, and José-Luis Sánchez-Cervantes

20 BootstrapPLS-Fuzzy-Genetic Hybridization to Predict the Effect of the Solidarity Economy and Sustainability on Competitiveness: The Case of a Farmer’s Market in Mexico 527
 Miguel Reyna-Castillo, Alejandro Santiago, Xóchitl Barrios-del-Angel, Lisbeth América Brandt-García, Daniel Bucio-Gutierrez, Yolanda Aranda-Jiménez, and Laura Moreno-Chimely

Part I
Decision Analysis

Chapter 1

A New Methodology Based on Multicriteria Ordinal Classification for the Management of Financial Resources with Application to Real Data from the Stock Market



Efrain Solares , Eduardo Fernández , Eyrán Díaz-Gurrola , Reimundo Moreno-Cepeda, Emmanuel Contreras-Medina , and Edy López Cervantes

Abstract Stock selection is highly complex due to the high heterogeneity of its factors. The determination of the value of a stock depends to a large extent on the investor's preferences towards such factors. This paper describes and evaluates a new methodology that uses investor decision policy to assign stocks to preferentially ordered classes. These classes can be of the “Don't Buy”, “Doubt” or “Buy” type. The classes are identified by limiting profiles at the boundary of each pair of consecutive classes and are given a priori by the investor. A back-testing strategy is used to evaluate the proposal and its results are compared with those of some benchmark approaches. The primary findings highlight that the stocks classified within the best class not only yielded better average returns compared to the broader market but also exhibited significantly lower volatility, suggesting a more favorable risk-reward

E. Solares (✉) · E. Fernández · E. Díaz-Gurrola · R. Moreno-Cepeda · E. Contreras-Medina
Research Center for Sustainable Development and Business Innovation, Universidad Autónoma de Coahuila, Torreón, Mexico
e-mail: efrain.solares@uadec.edu.mx

E. Fernández
e-mail: eduardo.fernandez@uadec.edu.mx

E. Díaz-Gurrola
e-mail: eyran_diaz@uadec.edu.mx

R. Moreno-Cepeda
e-mail: reimundo.moreno@uadec.edu.mx

E. Contreras-Medina
e-mail: emmanuelmedina@uadec.edu.mx

E. López Cervantes
Facultad de Informática, Universidad Autónoma de Sinaloa, Culiacán, Mexico
e-mail: edylopezc@gmail.com

balance and outperforming conventional methods and market benchmarks in terms of both returns and risk management.

Keywords Decision aiding · Outranking approach · Stock selection

1.1 Introduction

Many methodologies for stock selection have been presented in the related literature in recent decades. The interest of the scientific community in this field is due to the important repercussions that it entails, both in the academic ground and for society in general. Its practical and theoretical limitations have been addressed from various perspectives and theories, producing a considerably high number of factors involved in the selection process. However, determining the best way to aggregate the information provided by these factors is an open question.

The so-called multicriteria decision aiding (MCDA) [1] is an important branch of decision theory that can deal with the aggregation of information from multiple factors. In MCDA, a set of actions (decision alternatives) must be assessed based on a set of features called criteria to address choosing, ranking and sorting problems. Sorting problems, also called ordinal classification problems, consist of deciding how to assign actions to a collection of classes or categories that have been ordered (e.g., from worst to best) through the preferences of a decision maker (DM). In multicriteria ordinal classification, or multicriteria sorting, each action is assessed on multiple criteria and, aided by a decision model or decision policy, the actions are compared to reference profiles (reference actions) that characterize or define each class. Depending on the preferences within the decision model, the comparison allows the DM to assign the action to one or more classes.

Within the MCDA theory and in the context of multicriteria sorting, the most widely used method is ELECTRE TRI [2, 3], later called ELECTRE TRI-B by [3]. A recent generalization of ELECTRE TRI-B was introduced in [4]. This method, named INTERCLASS-nB, fulfills all the consistency properties imposed on multicriteria sorting methods and, like ELECTRE TRI-B, uses reference profiles at the limiting boundary of each pair of consecutive classes to perform the classification. Unlike its predecessor, INTERCLASS-nB can exploit more than one reference profile at each boundary. This is a major improvement on the original method, as it can give the sorting process more discerning power. In addition, the new method is versatile with respect to the elicitation of parameter values, since they can be represented by real or interval numbers, which reduces the cognitive effort of the DM to define these values and the time required to define the “best” parameter values. Here, we explore the abilities of INTERCLASS-nB to aggregate the information provided by all common factors and assign stocks to preferentially ordered classes that can lead to the final selection of the most convenient stocks.

In this paper, INTERCLASS-nB is used for first time to address the stock selection problem. Following [5] and, as described in Sect. 1.3, the methodology used to assess the proposal is defined as follows:

1. The set of decision alternatives is composed of the stocks in the S&P500 index [6].
2. The set of criteria is composed of the so-called fundamental factors [7] as well as price forecasting [8].
3. The collection and preparation of the input data is performed through capitaliq.com.
4. Expert investors were simulated to create a unified (although imprecise) decision policy that will provide the parameter values to exploit the multicriteria sorting method.
5. Assign the stocks (actions) to preferentially ordered classes.
6. Only the stocks that have been assigned to the best overall class are bought in a buy-and-hold strategy. A uniform distribution of resources is used to buy the selected stocks.
7. The created portfolio is assessed with a back-testing strategy.

This paper is structured as follows. Section 1.2 provides a review of the related literature and describes the INTERCLASS-nB method. Section 1.3 presents how this method is exploited to address the stock selection problem. The results are shown in Sect. 1.4. Section 1.5 concludes the paper.

1.2 Background

This section begins by mentioning and briefly describing the most outstanding works related to this paper; later, the details of the multicriteria sorting method used are provided.

1.2.1 *Review of the Related Literature*

Determining the most convenient stocks from a large universe of options is defined as stock selection. There are many factors involved in defining this convenience. Common factors in the related literature come from fundamental analysis [7].

Fundamental analysis uses data that is regularly published by the organizations underlying the stocks. This data is used to calculate indicators that investors often use to evaluate stocks. Indicators are both qualitative and quantitative, and typically include information that allows investors to compare the indicators (which represent the actual value of organizations) with current stock prices. Therefore, if one of these indicators shows evidence that the stock is undervalued, then it supports the decision that the investor should support the stock. If a sufficiently large number of indicators

provide such evidence, then the stock should be selected for investment. However, in practice, the decision is not straightforward since indicators do not usually provide evidence to reach the same conclusions.

We found in the literature review that the most used fundamental indicators used during stock selection are the following [9, 10].

- Profitability.
- Leverage.
- Liquidity.
- Efficiency.
- Growth.
- Solvency.
- Operational efficiency.

Several fundamental indicators were used in [7] for stock selection purposes. The indicators used in that work relied mainly on the price of the stocks and their relationship with the financial information published by the organizations underlying the stocks. The indicators used in that work are Debt to Equity, Price to Earnings, and Profit to Earn. Similarly, [11] used the Price to Earnings Ratio and New Loan to Market Capitalization ratios. As many as seventeen indicators were used in [10] for similar purposes.

The literature shows that fundamental indicators are commonly used in combination with other types of information during stock selection. For example, technical analysis is employed to complement fundamental information. In [12, 13], eight technical indicators were used together with eight fundamental indicators. Zarandi et al. [14] combined fundamental indicators with technical indices. On the other hand, price forecasting is also a very common approach used to complement fundamental analysis. Yang et al. [9] used an artificial neural network approach to forecast stock prices and combined them with twelve fundamental indicators.

Various stock selection methods have been employed, with notable ones including artificial neural networks, data envelopment analysis, evolutionary algorithms, sentiment analysis, and support vector machines [15]. In [16], data envelopment analysis is combined with multicriteria decision aiding theory for fund selection. Another study [17] introduces a novel three-stage network model in multiplier data envelopment analysis. In a different approach [18], a hybrid model between a feed-forward neural network and an adaptive neural fuzzy inference system is proposed. Additionally, a study [9] suggests using differential evolution to optimize an objective function based on historical prices, which in turn weights a set of indicators derived from fundamental analysis. Lastly, support vector machines are utilized in two studies [19, 20]

1.2.2 The INTERCLASS-nB Method

INTERCLASS-nB is a multicriteria sorting method that generalizes the most used sorting method within the MCDA theory, the so-called ELECTRE TRI-B. The main characteristics of INTERCLASS-nB are (i) its ability to assign actions to ordered classes through multiple profiles in the limiting boundary of each pair of consecutive classes (as opposed to ELECTRE TRI-B, that can use only one), and (ii) its capacity to cope with imprecision in the definition of its parameter values (through the so-called interval numbers).

Imprecise, vague, or ill-defined values can be assigned to the parameters of INTERCLASS-nB to reflect situations where the decision maker does not want/ is not willing to engage in an arduous process to define these parameters, and/or when the information to define the impacts on the criteria (factors) is not precisely known. This is accomplished in INTERCLASS-nB through the concept of interval number. An interval number can be defined as a variety of values (in the form of interval) that an unknown quantity can achieve. Formally, let's consider an amount r whose value is uncertain. The possible range of this value can be defined by its highest attainable value r^+ and its lowest attainable value r^- . This allows us to capture the variability and potential volatility of r within a bounded range, providing a clear framework to analyze its fluctuations between r^+ and r^- , respectively, $r = [r^-, r^+]$ is called interval number of r . We use bold font to identify interval numbers.

$A = \{a_1, a_2, a_3, \dots\}$ denotes the set of decision actions, while the collection of classes is denoted by C_k , $k = 1, \dots, M$, with C_{k+1} being preferred to C_k . To differentiate the boundary between each pair of classes, a set $B_k = \{b_{k,j}; j = 1, \dots, \text{card}(B_k)\}$ of limiting actions (profiles) $b_{k,j}$ is used, where $\{B_0, B_1, \dots, B_M, B_{M+1}\}$ is the set of all the limiting profiles (B_0 , and B_{M+1} are composed of the anti-ideal and ideal actions, respectively).

Consider δ and β established following the preferences of the decision maker. Given the assertion " $x \in A$ dominates $y \in A$ ", we denote its credibility as $xD(\alpha)y$. The assertion " x dominates y " is accepted when $xD(\alpha)y \geq \delta$. Similarly, the assertion " x is at least as good as y " is denoted by $\eta(x, y)$. The assertion " x is at least as good as y " is accepted when $\eta(x, y) \geq \beta$, and is denoted as $xS(\beta)y$. Furthermore, if $xS(\beta)y$ is hold but $yS(\beta)x$ does not, then it is said that " x is preferred to y ", which is denoted by $xPr(\delta, \lambda)y$. The specific steps to calculate $xD(\alpha)y$ and $\eta(x, y)$ are described in [4]. INTERCLASS-nB also exploits the following conditions to sort the actions in A into classes.

The following conditions are imposed to INTERCLASS-nB for the method to fulfill consistency properties [4]:

1. C_k is defined through a set of reference upper limiting profiles, B_k , and through a set of reference lower limiting profiles, B_{k-1} . It is assumed that all $b_{k,j}$ of B_k are in C_{k+1} .
2. B_0 (respectively, B_M) is composed of the anti-ideal (resp. the ideal) action.
3. For all k , there is no pair $(b_{k,j}, b_{k,i})$ such that $b_{k,j} Pr(\delta, \lambda)b_{k,i}$.
4. For all (k, h) ($h > k$), there is no pair $(b_{k,j}, b_{h,i})$ such that $b_{k,j} Pr(\delta, \lambda)b_{h,i}$.

5. For all k and for each action w in B_k , there is at least one action z in B_{k+1} such that $zD(\alpha)w$ (with $\alpha \geq \delta$).
6. For all k and for each action w in B_{k+1} , there is at least one action z in B_k such that $wD(\alpha)z$ (with $\alpha \geq \delta$).

Preferential relations can also be established between an individual action x and a set of profiles B_k as follows:

$$xS(\delta, \lambda)B_k \iff xS(\delta, \lambda)w \text{ for any } w \in B_k \text{ and there is no } z \in B_k \text{ with } zPr(\delta, \lambda)x.$$

$$B_kS(\delta, \lambda)x \iff wS(\delta, \lambda)x \text{ for any } w \in B_k \text{ and there is no } z \in B_k \text{ with } xPr(\delta, \lambda)z.$$

The assignment of alternatives to classes in INTERCLASS-nB is carried out through two interconnected methods: the pseudo-conjunctive and pseudo-disjunctive procedures. These procedures rely on the sets of profiles meeting specific conditions, as outlined previously. Here's a closer look at how each procedure functions:

- **Pseudo-Conjunctive Procedure:** This approach aims to ensure that an alternative satisfies a certain set of criteria, similar to a conjunctive evaluation. Only if the alternative meets the necessary conditions from the profiles will it be assigned to a corresponding class.
- **Pseudo-Disjunctive Procedure:** In contrast, this method allows for more flexibility by assigning an alternative to a class if it meets any one of the specified conditions from the profiles, resembling a disjunctive evaluation.

Both procedures work together to provide a comprehensive classification framework, enabling accurate and balanced assignments to the appropriate classes.

Pseudo-Conjunctive Procedure

- i. Compare x to B_k for $k = M - 1, \dots, 0$, until the first value, k , such that $xS(\delta, \lambda)B_k$;
- ii. Assign x to class C_{k+1} .

Pseudo-Disjunctive Procedure

- i. Compare x to B_k for $k = 1, \dots, M$, until the first value, k , such that $B_kPr(\delta, \lambda)x$;
- ii. Assign x to class C_k .

For more details of this method, such as general steps, the step-by-step algorithm, an optimization-based inference method for the INTERCLASS-nB, and other considerations, the reader is referred to [4].

1.3 Proposed Methodology

The proposed methodology exploits the INTERCLASS-nB method to assign each element in a set of actions to preferentially ordered classes; then, creating a portfolio with the stocks that were assigned to the best class and supporting each of these with an equal amount of resources; finally, using a back-testing strategy to compare the results of the proposal with those of the benchmark approaches.

This methodology is shown in Fig. 1.1 and described in the rest of this section.

1.3.1 Input Data

The Standard and Poor's 500 (S&P500) index is perhaps the most important stock benchmark worldwide, as it contains five hundred of the most representative companies of the United States of America. To perform the back-testing strategy, we used daily data for the last ninety business days. We use historical data provided by [capitaliq.com](https://www.capitaliq.com) about the stocks in the S&P500 index consisting of daily prices from March 5 to July 14, 2021 (ninety periods). Such a platform allows us to access financial statements and ratios preparing the input data particularly by discarding missing information (as opposed to fulfilling it with artificial information). Sixty periods are used to "train" the model. Then, the multicriteria sorting method assigns the stocks in the period immediately after the sixty training periods to finally select the stocks of the best class. This is done in the form of a sliding window, so that the selection of stocks is performed in thirty different periods that represent different contexts and trends of the market (assessing robustness of the proposal). [Capitaliq.com](https://www.capitaliq.com) is particularly used for downloading and preparing the data, while the rest of the procedure is performed using basic tabular tools such as Excel.

Table 1.1 provides a sample of the organizations in the S&P500.

Among the different options of fundamental factors used in the literature, given the complementarity of the information they provide, here we use the following factors based on the recommendations provided by [5, 9, 10, 13, 18]:

- a. Return on asset (RoA): indicates the profitability of the organization regarding its total assets. Provides an insight about the efficiency of the organization exploiting its assets to generate profits. It is calculated by dividing its net income by its total assets.
- b. Price to Earnings (PtE): measures the stock price of the organization regarding its earnings per share (that is, the profit of the organization divided by the outstanding shares of its common stock). Provides an insight about the relative value of the shares of the organization.
- c. Price to Book (PtB): indicates the price per share in the market regarding its book value. Provides an insight considering if the value of the organization is priced properly by the market.

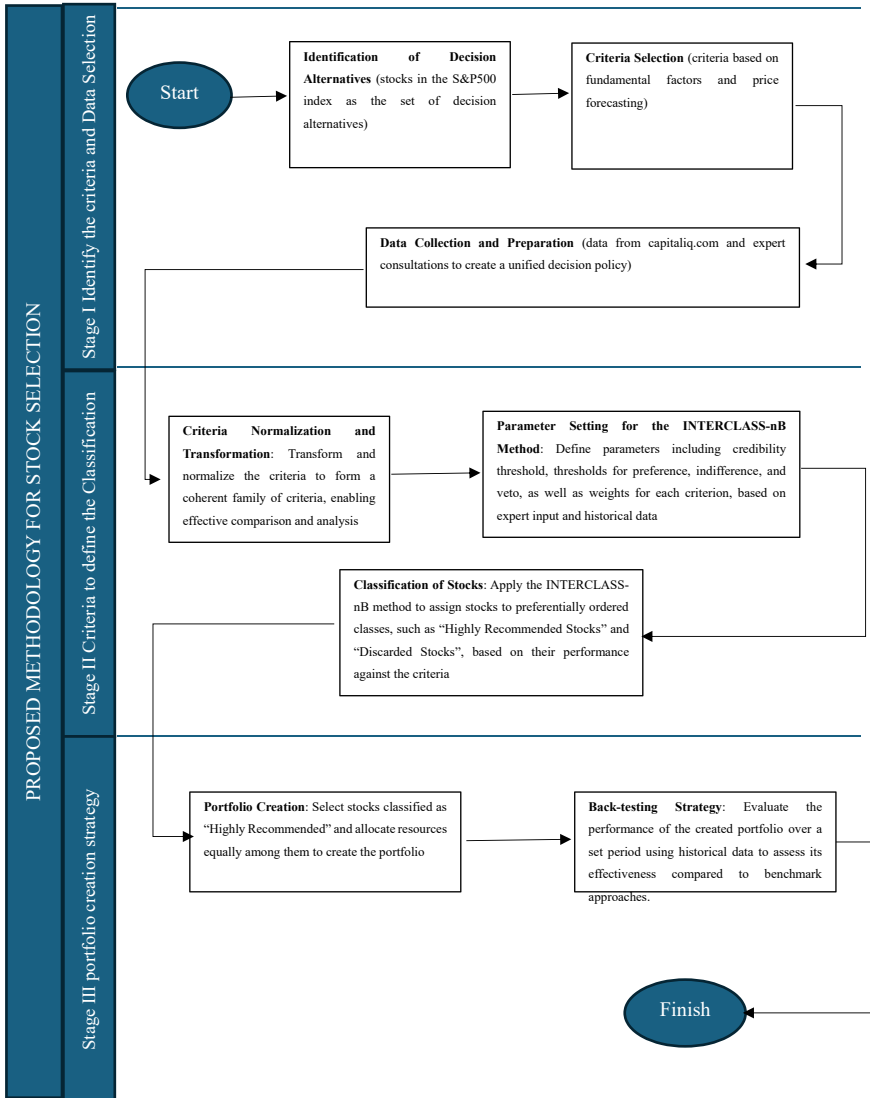


Fig. 1.1 Proposed methodology

- d. Price to Sales (PtS), indicates the price per share in the market regarding the revenue per share. Provides an insight about the stock price of an organization with respect to its revenues; that is, denotes the value that the market has puts on each dollar the organization has sold.
- e. Return on equity (RoE), indicates the efficiency of the organization in generating profits. It can be calculated as the net income of the organization divided by its average shareholder's equity.

Table 1.1 Organizations in the S&P500 index

Constituents name	Constituents	Constituents name	Constituents
3 M Company	NYSE:MMM
A. O. Smith Corporation	NYSE:AOS	Valero Energy Corporation	NYSE:VLO
Abbott Laboratories	NYSE:ABT	Ventas Inc	NYSE:VTR
AbbVie Inc	NYSE:ABBV	VeriSign Inc	NasdaqGS:VRSN
Abiomed Inc	NasdaqGS:ABMD	Verisk Analytics Inc	NasdaqGS:VRSK
Accenture plc	NYSE:ACN	Verizon Communications Inc	NYSE:VZ
Activision Blizzard Inc	NasdaqGS:ATVI	Vertex Pharmaceuticals Incorporated	NasdaqGS:VRTx
Adobe Inc	NasdaqGS:ADBE	ViacomCBS Inc	NasdaqGS:VIAC
Advance Auto Parts Inc	NYSE:AAP	Viatis Inc	NasdaqGS:VTRS
Advanced Micro Devices Inc	NasdaqGS:AMD	Visa Inc	NYSE:V
Aflac Incorporated	NYSE:AFL	Vornado Realty Trust	NYSE:VNO
Agilent Technologies Inc	NYSE:A	Vulcan Materials Company	NYSE:VMC
Air Products and Chemicals Inc	NYSE:APD	W. R. Berkley Corporation	NYSE:WRB
Akamai Technologies Inc	NasdaqGS:AKAM	W.W. Grainger Inc	NYSE:GWW
Alaska Air Group Inc	NYSE:ALK	Walgreens Boots Alliance Inc	NasdaqGS:WBA
Albemarle Corporation	NYSE:ALB	Walmart Inc	NYSE:WMT
AleOandria Real Estate Equities Inc	NYSE:ARE	Waste Management Inc	NYSE:WM
AleOion Pharmaceuticals Inc	NasdaqGS:ALON	Waters Corporation	NYSE:WAT
Align Technology Inc	NasdaqGS:ALGN	WEC Energy Group Inc	NYSE:WEC
Allegion plc	NYSE:ALLE	Wells Fargo & Company	NYSE:WFC
Alliant Energy Corporation	NasdaqGS:LNT	Welltower Inc	NYSE:WELL
Alphabet Inc	NasdaqGS:GOOG	West Pharmaceutical Services Inc	NYSE:WST
Alphabet Inc	NasdaqGS:GOOG.L	Western Digital Corporation	NasdaqGS:WDC
Altria Group Inc	NYSE:MO	Westinghouse Air Brake Technologies Corporation	NYSE:WAB
Amazon.com Inc	NasdaqGS:AMZN	WestRock Company	NYSE:WRK
Amcor plc	NYSE:AMCR	Weyerhaeuser Company	NYSE:WY

(continued)

Table 1.1 (continued)

Constituents name	Constituents	Constituents name	Constituents
Ameren Corporation	NYSE:AEE	Whirlpool Corporation	NYSE:WHR
American Airlines Group Inc	NasdaqGS:AAL	Willis Towers Watson Public Limited Company	NasdaqGS:WLTW
American Electric Power Company Inc	NasdaqGS:AEP	Wynn Resorts Limited	NasdaqGS:WYNN
American Express Company	NYSE:AxP	Xcel Energy Inc	NasdaqGS:XEL
...	...	Xilinx Inc	NasdaqGS:XLNX

Following the trend in the literature to combine multiple types of factors, in addition to these fundamental indicators that provide a wide range of perspectives regarding the financial situation of the organizations underlying the stocks, we use stock price forecasts based on more traditional estimations. We utilize an estimated future return for each stock, calculated as the mean return over the most recent sixty periods, plus or minus three times the standard deviation of those returns. This approach captures the expected return along with its variability, providing a range that reflects potential fluctuations based on historical data. Here, we exploit the ability of INTERCLASS-nB to deal with parameters defined as interval numbers. This is particularly useful in this situation given the high complexity usually involved in forecasting the stock returns.

Some of the factors mentioned above reflect information in various contexts. Following [1], we carry out an assessment of the factors to ensure that they can form a so-called coherent family of criteria. This way, the factors form a set of criteria that fulfill the properties of non-redundancy, completeness, and consistency. First, we confirmed that no criterion duplicates the function of another within the decision-making process. Each criterion should measure a unique aspect of the stocks, preventing any undue influence from overlapping measures. We achieved it by using factors that measure the quality of the stocks in different perspectives: forecasted returns, profitability, and company performance. Second, we ensured that the set of criteria covers all relevant aspects of the decision problem; in the context of this work, completeness refer to two basic aspects, i.e., fundamental analysis and forecasting factors. Finally, criteria are independent, meaning the evaluation of a stock under one criterion does not affect its evaluation under another. These criteria are normalized to $[0,1]$ (except for the forecasted return) considering the last sixty historical periods of stock prices. Table 1.2 shows a sample for the most recent period considered in the experiments, July 14, 2021, which is part of the input used to assess the methodology.

Table 1.2 Sample of the performance matrix for July 14, 2021

Constituents name	Constituents	RoA	PtE	PtB	PtS	RoE
3 M Company	NYSE:MMM	0.015	0.514	0.470	0.796	0.155
A. O. Smith Corporation	NYSE:AOS	0.668	0.393	0.851	0.746	0.730
Abbott Laboratories	NYSE:ABT	0.787	0.212	0.534	0.401	0.808
AbbVie Inc	NYSE:ABBV	0.000	0.736	0.684	0.943	0.000
Abiomed Inc	NasdaqGS:ABMD	0.013	0.374	0.463	0.482	0.000
Accenture plc	NYSE:ACN	0.170	1.000	1.000	1.000	0.194
Activision Blizzard Inc	NasdaqGS:ATVI	0.102	0.259	0.338	0.244	0.000
Adobe Inc	NasdaqGS:ADBE	0.973	1.000	1.000	1.000	0.017
Advance Auto Parts Inc	NYSE:AAP	0.864	0.235	0.919	0.625	0.928
Advanced Micro Devices Inc	NasdaqGS:AMD	0.763	0.626	0.739	0.615	0.264
Aflac Incorporated	NYSE:AFL	0.730	0.115	0.549	0.271	0.784
Agilent Technologies Inc	NYSE:A	0.535	0.619	0.958	0.949	0.894
Air Products and Chemicals Inc	NYSE:APD	0.024	0.579	0.538	0.425	0.000
Akamai Technologies Inc	NasdaqGS:AKAM	0.036	0.823	0.857	0.850	0.000
Alaska Air Group Inc	NYSE:ALK	0.197	0.000	0.000	0.352	0.000
...
Waste Management Inc	NYSE:WM	0.037	0.998	0.998	0.998	0.051
Waters Corporation	NYSE:WAT	0.073	1.000	1.000	1.000	0.000
WEC Energy Group Inc	NYSE:WEC	0.090	0.282	0.331	0.231	0.000
Wells Fargo & Company	NYSE:WFC	1.000	0.000	1.000	0.000	1.000
Welltower Inc	NYSE:WELL	0.068	1.000	1.000	1.000	0.000
West Pharmaceutical Services Inc	NYSE:WST	0.799	0.696	0.961	0.951	0.904
Western Digital Corporation	NasdaqGS:WDC	0.786	0.065	0.440	0.489	0.650
Westinghouse Air Brake Technologies Corporation	NYSE:WAB	0.071	0.719	0.741	0.823	0.000
WestRock Company	NYSE:WRK	0.060	0.000	0.222	0.234	0.000
Weyerhaeuser Company	NYSE:WY	0.855	0.039	0.143	0.124	0.879
Whirlpool Corporation	NYSE:WHR	0.712	0.123	0.201	0.277	0.926
Willis Towers Watson Public Limited Company	NasdaqGS:WLTW	0.092	0.000	0.000	0.003	0.756
Wynn Resorts Limited	NasdaqGS:WYNN	0.874	0.000	0.816	0.120	0.000
Xcel Energy Inc	NasdaqGS:XEL	0.095	0.380	0.564	0.262	0.000
Xilinx Inc	NasdaqGS:XLNX	0.013	0.604	0.536	0.604	0.418

1.3.2 Parameter Values

Here, we describe the parameter values used to operationalize the proposed methodology. Please note that expert knowledge from the investors is simulated by determining parameter values from (i) the historical performances of the stocks, (ii) preliminary experiments, and (iii) the literature.

Given the outstanding feature of INTERCLASS-nB to handle imprecise information in the form of interval numbers, the cognitive effort imposed to the decision maker during the elicitation of parameter values is considerably reduced. Therefore, we assume that the decision maker is able and willing to provide such parameter values and that he/she can do it in a straightforward manner.

1.3.2.1 Representing the Preference Model of the Decision Maker

According to the discussion in Sect. 2.1 and the specific steps described in [4], INTERCLASS-nB requires the following parameters to be defined in order to reproduce the preference model of the decision maker:

- A credibility threshold, δ , to establish clear preference relations.
- A threshold, λ , to define a strong majority in the outranking relation.

n criteria weights, w_i ($i = 1, 2, \dots, n$), to denote the relative importance of the criteria.

n indifference thresholds, q_i , ($i = 1, 2, \dots, n$) to denote the maximum differences between criteria impacts such that the decision maker feels that the impacts are indifferent.

n preference thresholds, p_i , ($i = 1, 2, \dots, n$) to denote the minimum differences between criteria impacts such that the decision maker can discern which of the impact is preferred.

n veto thresholds, v_i , ($i = 1, 2, \dots, n$) to denote the minimum differences between the impacts of actions x and y on the i th criterion, say $g_i(x)$ and $g_i(y)$, such that the decision maker can veto the assertion “ x is at least as good as y according to the i th criterion” when $g_i(y) - g_i(x) \geq v_i$.

To evaluate the proposal, the following parameter values are utilized. The credibility threshold for establishing clear preference relations, δ , is set at 0.51. The threshold λ for defining a strong majority in the outranking relation is set between 0.51 and 0.66. All criteria are assumed to have equal weight, with the threshold values reflecting that the criteria impacts are normalized to the range $[0, 1]$. The indifference thresholds are defined as $[0, 0.1]$, the preference thresholds as $[0.1, 0.2]$, and the veto thresholds as $[0.5, 0.7]$. As stated above, these values were set given the results from the historical performances of the stocks, some preliminary experiments, and the literature. Particularly, we have noticed from previous unstructured experiments on historical data that low values of δ and λ helped to obtain greater returns in the long term, while the (indifference, preference, and veto) thresholds are supported by the

Table 1.3 Profiles used to represent the limiting boundaries between classes

Boundary	Return on asset	Price to earnings	Price to book	Price to sales	Return on equity	Forecasted return
B_0	0	0	0	0	0	-0.08
B_1	0.6	0.6	0.6	0.6	0.6	[0.2, 0.4]
B_2	1	1	1	1	1	0.10

works [13, 21, 22]. Evidently, more structured experiments should be carried out to improve the performance of the proposal. However, this is out of the scope of this work, so the authors will address the topic as a future research line.

1.3.2.2 Representing Limiting Boundaries Between Classes Through Reference Profiles

Two classes were established for stock selection: the Discarded class (C_1) and the Highly Recommended Stocks class (C_2). The defining profiles at the boundaries of these classes are outlined in Table 1.3. The profile values were intuitively assigned given the ranges of criteria scores ([0, 1] for the fundamental indicators and around 0 for the average forecasted returns); particularly, such values denote convenience points characterizing the limiting boundaries between the classes.

1.4 Results and Discussion

As mentioned in Sect. 1.3, we consider daily data (for business days) of stocks in the S&P500 index for the time frame from March 5 to July 14, 2021 (ninety periods). In each of these periods, the fundamental indicators were derived from data for the immediately preceding quarter of the year, while the last sixty daily historical prices were used to forecast the stock's return; later, this information was used to evaluate the stock and assign it to one of the classes. To carry out the buy and hold strategy (with a portfolio with equally distributed weights), only the stocks allocated to the highest class were selected. Below we discuss first the allocations made by the proposal and then the returns the portfolios achieved.

1.4.1 Results of the Multicriteria Sorting Method

The work of the multicriteria sorting method consists of processing the input data to evaluate each stock on the multiple criteria and assign it to a preferential class. Since we are only interested in the most outstanding stocks, we only have two classes, the

Highly Recommended Stocks and the Discarded Stocks classes. Table 1.4 provides a sample of the assignments made by the method. For the period March 5, 2021. The complete set of results can be found in this link.

1.4.2 Returns of the Recommended Portfolios

Table 1.5 presents a summary of the returns produced by the selected stocks (as classified by the INTERCLASS-nB method) for each of the thirty historical periods under consideration. The periods correspond to specific dates, and for each period, the performance of the recommended stocks is compared against a benchmark index. The table is structured to provide insights into the effectiveness of the stock selection methodology by showing the financial outcomes of following the proposed strategy versus market performance as represented by the benchmark.

The entries under both “Recommended Stocks” and “Benchmark Index” columns are percentage changes, reflecting the variation in investment value from the beginning to the end of each period. Positive values indicate a gain, while negative values signify a loss. The comparison between these two columns is crucial for assessing the proposed stock selection methodology’s relative performance. It demonstrates whether the method consistently outperforms, matches, or underperforms against the market benchmark, providing a measure of its effectiveness and potential value to investors.

Given the large number of stocks in the S&P500 index, the third column of Table 1.5 tends to show larger numbers than the second one. However, it can still be clearly appreciated that the selected stocks provide better overall returns. The mean return for the stocks in the index after the thirty periods is -6.21% , while that of the selected stocks is -0.28% .

Another interesting result is that the summatory of the returns produced by the recommended stocks not necessarily follows the tendency (direction) of the market, which may be helpful in periods of general losses for the market. It is important to remark that the market presents high volatility. The concept of volatility (measured, for example, by the standard deviation) is a crucial indicator of the risk that the investor would be taking if he/she participates in a given portfolio of investment. The standard deviation of the summatory of returns of the stocks in the S&P500 index is 373.80% , while that of the recommended stocks is 1.33% .

1.5 Conclusions

A novel way of selecting stocks through a multicriteria sorting method has been described and assessed. The main component of the proposed methodology is the so-called INTERCLASS-nB method. This method exploits the available information about the preferences of the decision maker to assign actions (the stocks) to predefined

Table 1.4 Sample of the assignments of stocks to ordered classes made by INTERCLASS-nB

Constituent	Class	Constituent	Class	Constituent	Class	Constituent	Class	Constituent	Class	Constituent	Class
NYSE:MMM	C1	NasdaqGS:AEP	C1	NYSE:AVY	C1	NYSE:CARR	C2	NasdaqGS:CMCS.A	C1		
NYSE:AOS	C1	NYSE:Axp	C1	NYSE:BKR	C1	NYSE:CTLT	C1	NYSE:CMA	C2		
NYSE:ABT	C1	NYSE:AIG	C1	NYSE:BLL	C1	NYSE:CAT	C1	NYSE:CAG	C1		
NYSE:ABBV	C1	NYSE:AMT	C2	NYSE:BAC	C1	BATS:CBOE	C1	NYSE:COP	C1		
NasdaqGS:ABMD	C1	NYSE:AWK	C1	NYSE:BAx	C1	NYSE:CBRE	C1	NYSE:ED	C1		
NYSE:ACN	C1	NYSE:AMP	C1	NYSE:BDx	C1	NasdaqGS:CDW	C1	NYSE:STZ	C1		
NasdaqGS:ATVI	C1	NYSE:ABC	C1	NYSE:BRK.B	C1	NYSE:CE	C1	NasdaqGS:CPRT	C1		
NasdaqGS:ADBE	C1	NYSE:AME	C1	NYSE:BBY	C2	NYSE:CNC	C1	NYSE:GLW	C2		
NYSE:AAP	C1	NasdaqGS:AMGN	C1	NYSE:BIO	C1	NYSE:CNP	C1	NYSE:CTVA	C1		
NasdaqGS:AMD	C1	NYSE:APH	C1	NasdaqGS:BIIB	C1	NasdaqGS:CERN	C1	NasdaqGS:COST	C2		
NYSE:AFL	C1	NasdaqGS:ADI	C2	NYSE:BLK	C1	NYSE:CF	C1	NYSE:CCI	C1		
NYSE:A	C1	NasdaqGS:ANSS	C1	NasdaqGS:BKNG	C1	NYSE:CRL	C1	NasdaqGS:CSx	C1		
NYSE:APD	C1	NYSE:ANTM	C1	NYSE:BWA	C1	NasdaqGS:CHTR	C1	NYSE:CMi	C1		
NasdaqGS:AKAM	C1	NYSE:AON	C1	NYSE:BxP	C1	NYSE:CVx	C1	NYSE:CVS	C1		
NYSE:ALK	C1	NasdaqGS:APA	C1	NYSE:BSx	C1	NYSE:CMG	C1	NYSE:DHI	C1		
NYSE:ALB	C1	NasdaqGS:AAPL	C1	NYSE:BMY	C1	NYSE:CB	C2	NYSE:DHR	C1		
NYSE:ARE	C1	NasdaqGS:AMAT	C1	NasdaqGS:AVGO	C1	NYSE:CHD	C1	NYSE:DRI	C1		
NasdaqGS:ALxN	C1	NYSE:APTV	C1	NYSE:BR	C1	NYSE:CI	C1	NYSE:DVA	C1		
NasdaqGS:ALGN	C1	NYSE:ADM	C1	NYSE:BF.B	C1	NasdaqGS:CINF	C2	NYSE:DE	C1		
NYSE:ALLE	C2	NYSE:ANET	C1	NasdaqGS:CHRW	C1	NasdaqGS:CTAS	C1	NYSE:DAL	C1		
NasdaqGS:LNT	C1	NYSE:AJG	C1	NYSE:COG	C1	NasdaqGS:CSCO	C1	NasdaqGS:xRAY	C1		
NasdaqGS:GOOG	C1	NYSE:AIZ	C1	NasdaqGS:CDNS	C1	NYSE:C	C1	NYSE:DVN	C1		

(continued)

Table 1.5 Summary of the returns produced by the selected stocks in each of the thirty historical periods

Period	Recommended stocks (%)	Benchmark index (%)
5/3/2021	-1.60	52.83
8/3/2021	0.07	-98.19
9/3/2021	0.47	251.15
10/3/2021	-1.13	-93.48
11/3/2021	0.18	86.98
12/3/2021	-2.40	-229.60
15/3/2021	-0.95	48.69
16/3/2021	1.33	178.53
17/3/2021	-1.26	-208.36
18/3/2021	0.48	16.75
19/3/2021	-0.97	-340.13
22/3/2021	-3.58	-507.96
23/3/2021	-3.06	-833.29
24/3/2021	1.96	971.61
25/3/2021	0.59	69.01
26/3/2021	0.34	-69.39
29/3/2021	0.75	322.24
30/3/2021	1.35	364.07
31/3/2021	-1.09	-219.71
1/4/2021	0.15	-55.93
5/4/2021	0.74	125.37
6/4/2021	0.94	329.89
7/4/2021	0.63	114.34
8/4/2021	-0.63	-436.99
9/4/2021	0.23	103.36
12/4/2021	-1.71	-529.83
13/4/2021	1.23	811.70
14/4/2021	0.21	134.79
15/4/2021	-1.27	-467.06
16/4/2021	-0.26	-77.65

and preferentially ordered classes. The main characteristics of this method are that (i) it uses reference profiles at the limiting boundary between each pair of consecutive classes, (ii) it fulfills all the consistency properties imposed on multicriteria sorting methods, (iii) it is capable of incorporating imprecise data, vague or poorly defined preference information, so that obtaining the values of its parameters is relatively

easy for the decision maker, (iv) it can handle the effects of non-compensation and veto during the decision process.

This work proposes to use the INTERCLASS-nB method to determine outstanding stocks based on a set of factors taken from the fundamental analysis (an approach that is widely used by practitioners) and the forecast of stock prices. According to the literature review, combining different types of information that describe the quality of the stocks is a common practice. Furthermore, we notice a strong tendency to incorporate information from these types of analyses since they tend to make it easier to find undervalued stocks (and, therefore, with high probability of increasing their price). However, considering multiple factors usually increases the complexity of the problem, so addressing this problem is not straightforward. The characteristics of the INTERCLASS-nB method allowed us to cope with the complexity of the problem.

We assessed the proposal by simulating long-position investments with actual historical data (that is, a back-testing strategy) in stocks within the Standard and Poor's 500 (S&P500) index. To perform the back-testing strategy, we used daily data for the last ninety business days. First, sixty periods are used for "training", then the INTERCLASS-nB method assigns the stocks to the classes, so we can select the best classified stocks. This is performed in a sliding-window manner such that the selection of stocks is performed in thirty different periods representing different contexts and trends of the market, thus assessing robustness of the proposal. The results of the experiments showed that around ten percent of the stocks were assigned to the best category of stocks in each of the test periods. This is an important result because the stock selection phase (one of the stages of the overall management of stock portfolios) requires identifying a limited number of the best stocks. The actual return of the selected stocks is then compared to that of the stocks in the S&P500 index. Table 1.5 shows that the proposal outperformed the market in most scenarios in the context of summary of returns and with respect to volatility and cumulative return. Therefore, we conclude that the proposed approach is adequate to be considered by practitioners.

As can be seen in Table 1.4, INTERCLASS-nB only assigned around ten percent of the stocks to the best category (C_2), which is mainly due to the high demand imposed through the profiles in the boundaries between pairs of classes. In general terms, a portfolio with fewer shares is more convenient to exercise better control and achieve lower levels of commissions.

In future works, more experiments will be carried out to investigate the impact that such a requirement produces on the number of supported stocks. In any case, the decision maker can always provide the requirements and number of profiles with which he/she feels most comfortable.

The proposal should be further assessed using other preference models to represent the different behaviors of decision makers. Particularly, different number of profiles per limiting boundary, different sets of fundamental factors, other ways of forecasting stock prices, and diverse factors coming from other types of sources (such as technical and sentimental analyses).

References

1. Roy, B.: *Multicriteria Methodology for Decision Aiding*. Kluwer Academic Publishers, The Netherlands (1996)
2. Roy, B., Bouyssou, D.: *Aide multicritère à la décision: méthodes et cas*. Economica Paris (1993)
3. Almeida-Dias, J., Figueira, J.R., Roy, B.: Electre Tri-C: a multiple criteria sorting method based on characteristic reference actions. *Eur. J. Oper. Res.* 204, 565–580 (2010) <https://doi.org/10.1016/j.ejor.2009.10.018>
4. Fernández, E., Figueira, J.R., Navarro, J.: Interval-based extensions of two outranking methods for multi-criteria ordinal classification. *Omega* 95, 102065 (2020) <https://doi.org/10.1016/j.omega.2019.05.001>
5. Solares, E., de-León-Gómez, V., Fernández, E., Contreras-Medina, E., Lopez, O.: Multicriteria ordinal classification to improve strategic plan-ning in the financial sector of the company. *Int. J. Combinat. Optim. Probl. Inform.* 13, 214 (2022)
6. Frino, A., Gallagher, D.R.: Tracking S&P 500 index funds. *J. Portfolio Manag.* 28, 44–55 (2001)
7. Xidonas, P., Mavrotas, G., Psarras, J.: A multicriteria methodology for equity selection using financial analysis. *Comput. Oper. Res.* 36, 3187–3203 (2009) <https://doi.org/10.1016/j.cor.2009.02.009>
8. Solares, E., Salas, F.G., De-Leon-Gomez, V., Diaz, R.: A comprehensive soft computing-based approach to portfolio management by discarding undesirable stocks. *IEEE Access* 10, 40467–40481 (2022) <https://doi.org/10.1109/ACCESS.2022.3167153>
9. Yang, F., Chen, Z., Li, J., Tang, L.: A novel hybrid stock selection method with stock prediction. *Appl. Soft Comput.* 80, 820–831 (2019) <https://doi.org/10.1016/j.asoc.2019.03.028>
10. Shen, K.-Y., Tzeng, G.-H.: Combined soft computing model for value stock selection based on fundamental analysis. *Appl. Soft Comput.* 37, 142–155 (2015) <https://doi.org/10.1016/j.asoc.2015.07.030>
11. Chai, J., Du, J., Lai, K.K., Lee, Y.P.: A hybrid least square support vector machine model with parameters optimization for stock forecasting. *Math. Probl. Eng.* 2015, 1394 (2015) <https://doi.org/10.1155/2015/231394>
12. Fernandez, E., Navarro, J., Solares, E., Coello, C.C.: A novel approach to select the best portfolio considering the preferences of the decision maker. *Swarm Evol. Comput.* 46, 140–153 (2019) <https://doi.org/10.1016/j.swevo.2019.02.002>
13. Fernandez, E., Navarro, J., Solares, E., Coello, C.C.: Using evolutionary computation to infer the decision maker's preference model in presence of imperfect knowledge: a case study in portfolio optimization. *Swarm Evol. Comput.* 54, 100648 (2020) <https://doi.org/10.1016/j.swevo.2020.100648>
14. Zarandi, M.H.F., Rezaee, B., Turksen, I.B., Neshat, E.: A type-2 fuzzy rule-based expert system model for stock price analysis. *Exp. Syst. Appl.* 36, 139–154 (2009) <https://doi.org/10.1016/j.eswa.2007.09.034>
15. Andriopoulos, D., Doumpos, M., Pardalos, P.M., Zopounidis, C.: Computational approaches and data analytics in financial services: a literature review. *J. Oper. Res. Soc.* 70, 1581–1599 (2019). <https://doi.org/10.1080/01605682.2019.1595193>
16. do Castelo Gouveia, M., Duarte Neves, E., Cândido Dias, L., Henggeler Antunes, C.: Performance evaluation of Portuguese mutual fund portfolios using the value-based DEA method. *J. Oper. Res. Soc.* 69, 1628–1639 (2018) <https://doi.org/10.1057/s41274-017-0259-7>
17. Galagedera, D.U.A., Roshdi, I., Fukuyama, H., Zhu, J.: A new network DEA model for mutual fund performance appraisal: an application to US equity mutual funds. *Omega* 77, 168–179 (2018) <https://doi.org/10.1016/j.omega.2017.06.006>
18. Huang, Y., Capretz, L.F., Ho, D.: Neural network models for stock selection based on fundamental analysis. In: *Proceedings of the 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pp. 1–4 (2019) <https://doi.org/10.1109/CCECE.2019.8861550>

19. Huang, H., Wei, X., Zhou, Y.: A sparse method for least squares twin support vector regression. *Neurocomputing* 211, 150–158 (2016) <https://doi.org/10.1016/j.neucom.2015.12.133>
20. Huang, H., Wei, X., Zhou, Y.: Twin support vector machines: a survey. *Neurocomputing* 300, 34–43 (2018) <https://doi.org/10.1016/j.neucom.2018.01.093>
21. Xidonas, P., Askounis, D., Psarras, J.: Common stock portfolio selection: a multiple criteria decision making methodology and an application to the Athens stock exchange. *Oper. Res.* 9, 55–79 (2009) <https://doi.org/10.1007/s12351-008-0027-1>
22. Xidonas, P., Mavrotas, G., Krintas, T., Psarras, J., Zopounidis, C., Xidonas, P., Mavrotas, G., Krintas, T., Psarras, J., Zopounidis, C.: Stock selection. *Multicrit. Portfolio Manag.* 32, 23–55 (2012) https://doi.org/10.1007/978-1-4614-3670-6_2

Chapter 2

M-SALD: A Novel Support Tool for Geo-driven Decision-Making in the Selection of Hydraulic Structures Location



Solangel Rodríguez-Vázquez , Yeleny Zulueta Véliz ,
and Yamilis Fernández Pérez 

Abstract Expert systems are fundamental tools in the fields of artificial intelligence and decision-making. In the context of the selection of areas for the location of dams, expert systems are especially relevant since they can integrate multiple variables for the detection of optimal locations that minimize environmental impacts and maximize the efficiency of the infrastructure. This research paper presents a new model of multicriteria and geospatial analysis (M-SALD). M-SALD is an expert system structured in three phases that can select from a set of areas the most suitable for the possible location of dams. In addition, it integrates the processes of multicriteria analysis and geospatial analysis, allowing multiple factors to be considered simultaneously, considering the spatial location of the data and the interaction between them. The model is applied to a case study in which a total of 29 watersheds (alternatives) were evaluated, considering 4 criteria and 25 sub-criteria. As a result of the application, it is evident that the model allows for the expansion of the number of possible factors, parameters, and alternatives to be evaluated, reducing the inconsistency from 80 to 20%. It eliminates the subjective evaluation performed by the experts during the weighing of the alternatives and reduces by up to 22% the number of candidate points (12,591) evaluated on the rivers. To obtain the results, the possible scenarios of hydrological development were considered, including promising areas to ensure the balance of water resources.

Keywords Artificial intelligence · Decision-making · Expert systems · Geospatial analysis · Multicriteria analysis · M-SALD · Location of dams

S. Rodríguez-Vázquez (✉) · Y. Z. Véliz · Y. F. Pérez
University of Informatics Sciences, 10800, Havana, Cuba
e-mail: solrusita85@gmail.com

2.1 Introduction

Nowadays, the planning and construction of hydraulic structures, such as dams, requires an approach based on multiple variables and factors to ensure their viability, efficiency, and sustainability. The precise selection for the location of the dam is a significant milestone, as it has a significant impact not only on the management of the water resource, but also on its operation on a social, economic, and environmental level. In this sense, artificial intelligence is a tool capable of revolutionizing the way this topic is addressed using a wide range of techniques, including expert systems, fuzzy logic, neural networks, and applied data mining. The use of expert systems in the selection of areas for the location of dams (SALD) can help to carry out the decision-making process and reduce the risk of errors and negative consequences for the natural and social environment based on environmental, technical, and socio-economic criteria. In this way, expert systems can systematically analyze and weigh these factors and provide recommendations based on specific rules and algorithms.

Consequently, the multicriteria decision methods, far from being considered infallible and precise elements, the use of which allows finding an optimal and definitive solution, are a basis, supported by scientific elements, that provides distinctive improvements to make a decision. As Semenov et al. [1] have studied, in any case, these are decisions based on quantifiable components that allow us to weigh the risk, and, by virtue of this, they are able to choose the “decision” that, at best, turns out to be the most satisfactory and, at worst, the least satisfactory.

With the traditional SALD approach, hydrological information is processed independently, resulting in a large volume of variables (criteria, sub-criteria and alternatives) to be evaluated during decision-making. The analysis is complicated by the high level of imprecision and vagueness related to the initial information, the future conditions, the preferences of the decision-makers at the time of choosing the location of the dam. In addition, the process of selecting areas for the location of dams is carried out manually, being able to then affirm that the predictive power of the areas in terms of the water viability of the watersheds and the accuracy in the selection of the area on the rivers for the location of the dams will be affected.

This paper proposes a new multicriteria and geospatial model (M-SALD) for decision-making on the selection of the location of hydraulic installations for an efficient and sustainable use of water. M-SALD allows the integration of the methods of experimental theoretical study of the characteristics of watersheds, the methods of multicriteria evaluation and geospatial analysis of distances between points. In addition, it reduces the subjectivity in the evaluation of alternatives by using the real values of the sub-criteria (parameters) to obtain the weighting of the same, and not the direct assessment of the experts. It allows to reduce environmental and social risks in the analyzed area for the placement of dams and increase the predictive value and accuracy during the pre-feasibility study of hydraulic projects. It solves the SALD problems and decreases the time during the decision-making in the pre-feasibility analysis stage.

The research paper is divided into four sections, starting with the introduction. The introduction provides an overview of the study, including the motivation, the problem statement, the research contribution, the novelties and impacts and the general structure of the article. Next, in the related studies section, a brief study is presented regarding the evolution of the SALD problems and their solution approaches, as well as a comparative table between the analyzed studies and the proposed model using as indicators the limitations found during the study. The section on methodologies and methods describes in a general way the proposed M-SALD, deepening in the operation of each of its phases. Continuing, the results and discussion section presents several important aspects. It includes the validation of the model through: (1) its application to a real case study, (2) the comparison between the inconsistencies obtained by different authors and M-SALD in terms of the number of criteria, sub-criteria and alternatives used and (3) the use of Spearman's correlation coefficient to compare the correlation between the results of applying M-SALD and other methods to the same case study. Finally, section four presents the general conclusions resulting from the research.

2.2 Related Studies

As mentioned in the previous section, the traditional approach to solving SALD problems is performed manually using geospatial maps that allow areas to be visualized and/or through the analysis of the calculation of the parameters that directly influence the areas. This hinders the speed and quality of decision-making because a bias is introduced as a result of the uncertainty caused by the absence of key information during the process. For some years now, researchers have been including some artificial intelligence techniques that favor the decision-making process and the selection of the most appropriate areas for dam positioning.

Othman et al. [2] present a study to identify suitable locations to build dams within a single basin. They use the fuzzy Analytical Hierarchical Process (AHP) and the Weighted Sum Method (WSM) and compare their results to select suitable prey sites. During the analysis, 7 criteria and 21 alternatives are considered. On the other hand, Rane et al. [3] evaluate the potentiality of the locations by integrating the Spatial Analytical Multiple Influence Factor (MIF) and the Technique in Order of Preference by Similarity with the Ideal Solution (TOPSIS). They considered 3 criteria and 12 sub-criteria during the evaluation. The selection of the potential site was obtained through a suitability map created using the weighted superposition technique. Alrawi et al. [4] use the AHP method in the analysis, combined with Geographic Information System (GIS) and Remote Sensing (RS) techniques. They use 9 criteria during the evaluation among which the environmental factor is not found. As in [3] they used the weighted superposition technique for the final site selection. Zytoon et al. [5] use RS techniques integrated with GIS to evaluate the feasibility of building a dam. In this study, only 4 possible sites are analyzed, and the characteristics of the watershed are considered, among other aspects. Among the analysis criteria, the environmental

factor is not considered. Dortaj et al. [6] evaluate 10 alternatives taking into account 4 criteria. The alternatives were classified using ELECTRE in all its variants and the results were combined applying the grade point average and the classification strategies of Borda and Copeland. Alkhuzai et al. [7] use GIS for the mapping and selection of the dam site basing their analysis on the use of the Digital Elevation Model (DEM) as a study basis. During the research they use 4 sub-criteria for the final obtaining of three evaluation sites. Bihon et al. [8] uses the sensitivity analysis of the weights of the factors based on GIS using the AHP multicriteria analysis method and states that this type of analysis reduces the uncertainty of the final map of potential sites and decreases the subjectivity during the evaluation of the experts. During the analysis it considers 9 criteria among which the environmental factor is not found. Atiq et al. [9] conducted a research on the location of dam sites by using GIS tool and RS technique and it is stated that tectonically active fault regions are considered the worst for dam location. Ajibade et al. [10] identifies potential points on rivers, using geographic information systems and remote sensing techniques integrated with fuzzy logic. Two criteria and four sub-criteria were used during the study. The fuzzy members were combined using the fuzzy overlay technique to obtain the appropriate dam site selection map. Most of the sites obtained in this research are characterized by having a high elevation and a gentle slope. Wang et al. [11] carried out a study of the literature corresponding to prey positioning in the last 20 years considering a total of 136 scientific articles. They analyze the most used methods and techniques focusing on the factors that influence the SALD problems and evaluate the impact of different dam functions on the siting factors. The results of this research show that: (1) the methods and techniques most commonly used in this type of analysis are based on GIS/RS, based on MCDM and MCDM—GIS and based on machine learning, (2) the site selection factors vary greatly, depending on the function of the dam and (3) the insufficiency that still exists in research regarding the integration of different methods of solving SALD problems.

From the bibliographic review carried out and based on the results obtained by the article [11] it is necessary to emphasize that there are inadequacies or limitations in terms of:

- The environmental factor is not considered as one of the most important aspects to achieve a sustainable balance between water resources management and environmental management.
- Analyses are carried out on the potential sites of the rivers that belong to a single watershed.
- The assessment of the water viability of watersheds is not taken into account.
- The investigations are focused on dams with a single function.
- The number of criteria used for the evaluation of sites is low with respect to the number of criteria that intervene in this type of selection.
- According to Wang et al. [11], the percent of criteria that are most used in research are of a topographic and geological nature.

As evidenced by the current, trends for the resolution of SALD problems are directed towards two directions—the use of multicriteria analysis methods to obtain

a hierarchical order of the alternatives that are analyzed and the use of geospatial analysis techniques using the GIS. The M-SALD model proposed in this chapter substantially reduces each of the deficiencies detected during the bibliographic analysis. This is evidenced by its benefits among which are that:

- It allows the integration of the methods of experimental theoretical study of the characteristics of watersheds, the methods of multicriteria evaluation and geospatial analysis of distances between points.
- Presents the environmental factor as a fundamental axis through the distance analysis between spatial objects that is carried out in the geospatial analysis process in phase two of the model.
- During the development of the research, the author considers that the more comprehensive the area of selection, the more precision will exist at the time of selection. That is why it is considered that the evaluation of the water viability of watersheds is an important aspect to consider during the multicriteria analysis process. The better the watershed is evaluated, the greater the water collection that is carried out by the dam to be built.
- All the points about the rivers are considered, being the same discriminated from the geospatial analysis carried out in phase 2 of the model.
- The model allows to evaluate any type of dam without discriminating its function because it is totally independent of the criteria to be evaluated.
- The number of criteria, sub-criteria, and alternatives to evaluate the higher the better. This is possible from the use of the hybrid method proposed for the AHP-TOPSIS multicriteria analysis with its modifications. A more extensive description of this method can be studied in the research paper [12].

Based on the limitations detected during the research, a comparison of the research works analyzed in terms of the above characteristics is made with respect to the M-SALD model proposed in this research work (Table 2.1).

Despite the above, the limitations of the model and future research include the possibility of incorporating the multicriteria analysis on the potential points on the rivers that are obtained by the geospatial analysis method to increase the accuracy in the selection of the location. For a better understanding and deepening of the operation of the proposed model, the doctoral study "*Multicriteria and geospatial model for the problem of selecting areas for the location of dams*" can be used in [13].

2.3 Methodology and Methods

Studies related to the identification of priority areas include the geographic space, the objective and the social aspects as common terms in the definition of this concept [14–16]. For the purposes of this proposed model, the priority areas for the conservation of natural resources are spatial representations of the territory, where specific and optimal environmental, biophysical, socioeconomic, cultural and/or political

Table 2.1 Comparison between the analyzed studies and the proposed model using as indicators the limitations found during the study

Research works	Analysis of watersheds	Number of variables to be evaluated (C—Criteria, SC—Sub-criteria, Alt—Alternatives)	Combined use of techniques, methods and GIS	Use of the environmental factor	Dams with different function	Analysis of rivers in more than one basin
Othman et al. [2]	No	7—C, 21—Alt	No	No	No	No
Rane et al. [3]	No	3—C, 12—SC	Yes	Yes	No	No
Alrawi et al. [4]	No	9—C	Yes	No	No	No
Zytoon et al. [5]	Yes	3—C, 4—Alt	No	No	No	No
Dortaj et al. [6]	No	4—C, 10—Alt	No	Yes	No	No
Alkhuzai et al. [7]	No	4—SC	No	No	Yes	No
Bihon et al. [8]	No	9—C	Yes	No	No	No
Atiq et al. [9]	No	2—C	No	No	No	No
Ajibade et al. [10]	No	2—C, 6—SC	No	Yes	No	No
M-SALD	Yes	AHP-TOPSIS 4-C, 25-SC, 29-Alt Geospatial analysis 12,591 sites analyzed	Yes	Yes	Yes	Yes

attributes converge for a given objective; and whose permanence is in imminent risk due to natural or human causes or both. Priority areas differ from eligible and potential or optimal areas. Those that are classified as suitable correspond to the spaces of the territory that meet the natural characteristics necessary for the provision or development of a certain service; for example, water or ecotourism [17]; or an activity: establishment of commercial forest plantations [18]. If, in addition, it is desired to integrate other types of information, such as socioeconomic aspects, that help or limit the suitability of these spaces, then the potential areas are identified using spatial statistical models and optimization methods. Finally, priority areas arise when those identified as potential turn out to be at risk or be vulnerable to changes that diminish their capacity to provide the service or activity considered. That is, the priority areas are more specific and are circumscribed in the potential areas, which, in turn, are within the suitable areas.

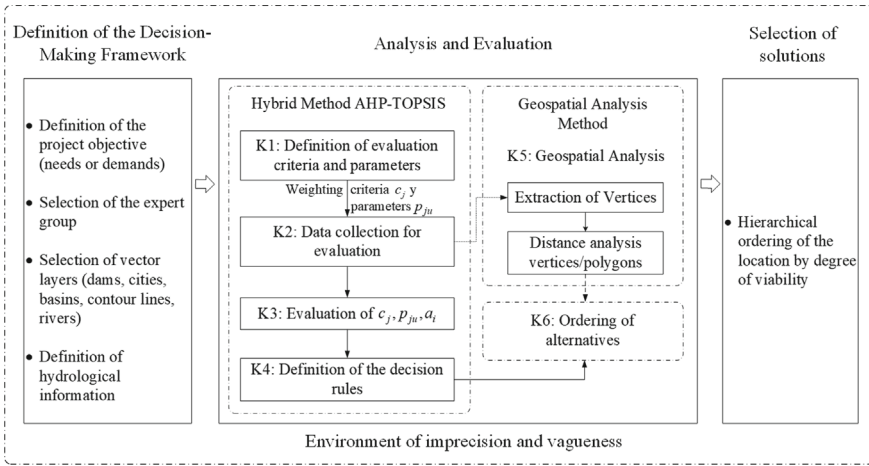


Fig. 2.1 General scheme of the updated M-SALD model. (Update of the schema published by the author in [19])

The techniques of identifying priority areas for conservation have different approaches, from the merely intuitive or qualitative to the quantitative analytical. The model M-SALD, is a novel support tool for geo-driven decision making in the selection of hydraulic structures location. Its main foundation is decision-making in the selection of the best territories in a region to carry out possible dam construction projects. It is structured in three phases as shown in scheme (Fig. 2.1): Definition of the Decision-Making Framework, Analysis and Evaluation, and Selection of solutions.

Phase 1: Definition of the Decision-Making Framework

The first step in carrying out a dam construction project is to define the main objective for which you want to build, that is, what type of dam is the evaluation of the land aimed at and what is the goal you want to achieve with this dam. This enables a correct evaluation by experts towards the criteria and parameters that are subsequently analyzed in Phase 2 of the M-SALD model.

Next, it is necessary to define who or who will be the experts who will evaluate the criteria and parameters established for the final evaluation of the alternatives. In the case in which the evaluation is carried out by a single expert, it will be necessary for them to weigh each of the criteria and parameters that he wishes to consider according to their opinion or according to a collegiate opinion with other experts. In the case in which it is decided to carry out the evaluation by a group of experts, it is proposed that it be carried out by using the geometric mean to obtain a common weighting value for each of the criteria and parameters to be used.

When the analysis of a region is carried out with the objective of selecting which are the best areas for the construction or location of dams, it is necessary to collect as much information as possible so that the model can carry out the analysis as close to

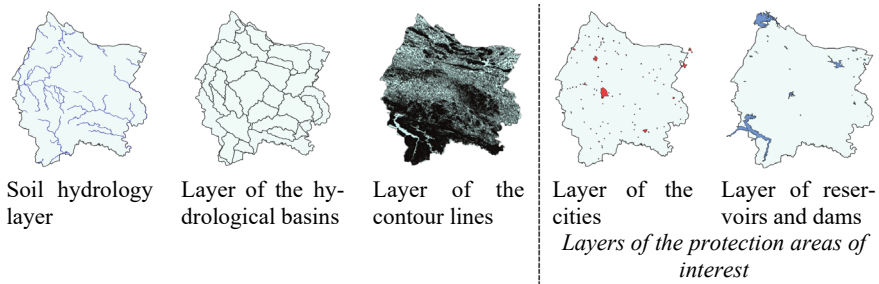


Fig. 2.2 Layers to be used by the M-SALD model

reality as possible. Said information should be composed of the vector layers (.shp) which will be used in the geographic information system (GIS) that is determined to be implemented and the necessary hydrological information regarding the area to be analyzed.

There are several ways to extract these vector layers:

- Making use of raster files such as the Digital Elevation Models (DEM) extracted from any of the existing online databases such as the geographic file download centers of each country, NASA, Geological Prospection USA, ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer), Vertex (Alaska Satellite Facility Data Portal), among others.
- Creating the layers manually, which is the most common form that occurs in countries that do not have internet access or that due to their geographical location do not have access to these online spaces.

It is necessary to point out that to achieve a correct functioning of the model, the experts must provide the following layers (Fig. 2.2):

1. Hydrology layer of the terrain (rivers, streams).
2. Basin layer (the more delimited the basins are in the region, the better the analysis and the more specific the results obtained).
3. Contour lines layer.
4. Layers of the areas of protection interest (Industries, streets, cities and / or towns, forests, dams).

The collection of hydrological parameters through the model proposal is carried out in three different ways:

1. Parameters provided directly by specialists: These parameters are data that are constant already evaluated and analyzed by the expert group and that are equal for all alternatives (i.e. each of the areas to be evaluated), say, for example: return period (years), duration of precipitation, the annual evapotranspiration coefficient, the average runoff. Note that, in this case, these parameters will not belong to the group of parameters that will be evaluated for analysis in the search

for the best alternatives. This is because, as mentioned above, these data are equal for all areas and therefore lose their comparative value, and their value will only be used to obtain parameters if they have evaluative value in the comparison of areas.

2. Internal parameters in vector layers: Each of the vector layers internally contains attributes that characterize the geometric systems that are contained in them, such as: the geometry of the object, the identifier of each object. But also, there are parameters that are previously calculated for each of the alternatives using the GIS tool and that are included as part of the attributes of these vector layers. Examples of this are: the area of the basin, the perimeter of the basin, the length of the basin, the length of the rivers.
3. Internal calculation of other parameters: There are parameters in which its existence depends on the analysis of other parameters such as: the intensity of the rain, the compactness index, the shape index. Each of them is obtained through the formulas established in the bibliography [20–22]. These calculations are proposed to be performed in Phase 2 of the model.

Phase 2: Analysis and Evaluation

Phase 2 of the M-SALD combines two processes that run in parallel: (1) the multi-criteria analysis process using the AHP-TOPSIS hybrid method and (2) the distance analysis process between spatial objects.

Hybrid Method AHP-TOPSIS: According to Jozaghi et al. [23], two of the most widely used multicriteria analysis techniques in solving SALD problems are the Analytical Hierarchy Process (AHP) and the Technique in Order of Preference by Similarity to the Ideal Solution (TOPSIS). As well as in the present research work, Jozaghi et al. [23] advocates the integration of multicriteria analysis methods and data analysis tools, such as GIS, due to their continuous use for the resolution of this type of problems by previous researchers [24–26]. As expressed in [27] and [28–30], GIS is a powerful tool in working with SALD problems; however, its capacity is limited because it requires using several tools to obtain a single result. This causes a delay in the decision-making process and a decrease in the usability of the tool since the specialist must have extensive knowledge in working with the tool. In the application of the AHP method [31–33] to the selection of areas for the location of dams, a process is followed that involves the identification of the key criteria, the construction of a hierarchy that reflects the relationship between these criteria and the alternatives, the assignment of weights to the criteria based on their relative importance, and the comparison of the alternatives in relation to each criterion. Through the comparison and weighting of the criteria and alternatives, the AHP method allows to obtain a structured and systematic evaluation that facilitates informed decision-making and the selection of the optimal location for dam construction. The TOPSIS method [34, 35] is a multicriteria decision analysis technique that is based on comparing the available alternatives with respect to a set of predefined criteria. In the context of the selection of areas for the location of dams, the TOPSIS method [36–38] is used to evaluate and classify the different potential locations based on multiple criteria,

such as geology, hydrology, topography, environmental impact, accessibility. The TOPSIS method is based on the idea that the best alternative is the one closest to the ideal solution and farthest from the worst solution. To do this, the distance of each alternative to the ideal solution and to the worst solution in a multidimensional space defined by the evaluation criteria is calculated. Then, a relative proximity index is determined for each alternative, which allows to establish an order of preference.

Through the analysis carried out in [23] and [39], it is considered that the use of hybrid methods is more interesting and efficient where the advantages that each of these methods can provide are combined. In this way, the use of a hybrid system between the AHP and TOPSIS multicriteria analysis methods is proposed. AHP allows to obtain the weight of the criteria and sub-criteria selected by the experts, as well as the generation of the matrices of comparisons by pairs of criteria, sub-criteria, and of the alternatives corresponding to each of the sub-criteria. In order to reduce human error in the construction of the pairwise comparison matrix and to increase the consistency of said matrices, a value comparison scale described by the author in [12] was created, constructed from the scales proposed by Saaty in [40]. This scale allows to automate the construction of the comparison matrices by pairs eliminating the subjectivity of the experts during the realization of said process and to increase the list of criteria, sub-criteria, and alternatives to be used during the analysis of the areas. For the comparison of priorities between the alternatives with respect to each sub-criteria as one of the fundamental contributions of the proposed hybrid method, an evaluation strategy was created to obtain a final weighting of each of the alternatives with respect to each of the sub-criteria. This strategy is based on the comparison of the values of each of the sub-criteria that directly influence each of the alternatives individually. The description of the “*Strategy of evaluation of the alternatives with respect to the values of the sub-criteria*” was published by the author in [12].

On the other hand, the TOPSIS method, making use of the decision matrix of the alternatives obtained by AHP, first converts the dimensions of the different criteria into non-dimensional criteria to subsequently administer the decision-making rules, determine the positive ideal solution and the negative ideal solution, and calculate the separation distance of each competitive alternative to each positive and negative ideal solution.

The most complete description of the application of the AHP-TOPSIS hybrid system to the proposed M-SALD can be consulted in the publications [19] and [27] of the present author.

Geospatial Analysis Method: The study carried out so far of the hydrographic basins makes it possible to predict which are the areas with the greatest water viability in a region, however, it is not sufficient for determining the potential points in which it is possible to locate dams. For this reason, it was decided to integrate a geospatial analysis of the terrain into the multi-criteria analysis, whose fundamental premise is the environmental care of the ecosystem and the areas of interest around the location point.

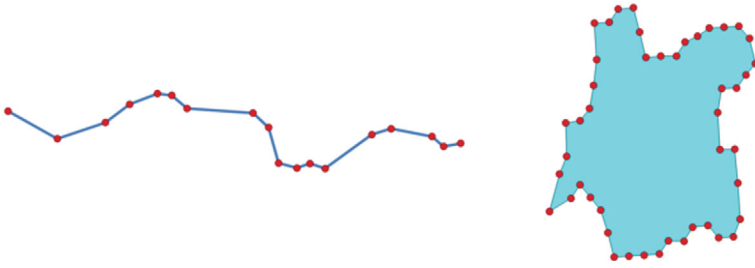
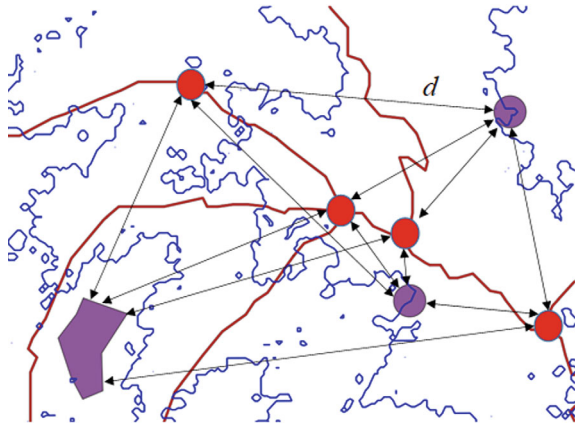


Fig. 2.3 Vertices extracted in line and polygon by the *native:extractvertex* algorithm

Fig. 2.4 Distance analysis between spatial objects



The proposed method of geospatial analysis is based on the studies carried out by Rodríguez-Vázquez and Mokra in [41–43]. This method is based on two fundamental steps: river extraction of vertices (Fig. 2.3) and distance analysis between geospatial objects (Fig. 2.4).

- *Extraction of vertices*: The drainage network of a basin is an independent line for each section, characterized by a set of vertices (Fig. 2.3) to which a unique identifier is assigned. Through these vertices it is possible to perform the distance analysis to nearby polygons, as well as to evaluate possible environmental situations to occur in the analysis area. In the research [41], computational geometry techniques are used for the detection and automated extraction of fluvial vertices corresponding to the rivers of the analysis area. At the conclusion of this analysis, it was defined that the use of the *native:extractvertex* algorithm was the most suitable to be used by the M-SALD.
- *Vertex/Polygon Distance Analysis*: For terrain analysis it is necessary to use vector or raster layers. In each of these layers there are areas that should not be included in the analysis required for the context of this research. It is necessary to consider that there are areas that are of maximum priority according to current legislation

[44] for the process in question. In all cases it is necessary to consider possible legal restrictions to delimit the areas that are suitable for the location of dams.

To achieve this step, the modified *NNjoin* algorithm proposed by Rodríguez-Vázquez and Mokrova in [42] and [43] is used for the distance treatment between objects located geospatially in different vector layers. The modification to the *NNjoin* algorithm was necessary because it only allows the calculation of distance between objects (Fig. 2.4) but does not eliminate the vertices that are within the distance range previously established by the expert to avoid possible damage in the areas surrounding the future dam. This is a necessary aspect for the fulfillment of the final purpose of the new M-SALD model. Said analysis is performed between the vertex layer obtained by the *Extraction of vertices* step and the protected area layers provided by the experts. Protected areas must be in relation to the laws of environmental, social, and economic care and protection.

Phase 3: Selection of Solutions

The overlapping of layers is one of the most important and necessary tools found in a GIS when it comes to decision making procedures/process. This is because it allows combining different sets of geospatial data to analyze the relationship and interaction between them. In addition, it facilitates the visualization and comparison of multiple aspects of the geographical environment on the same map, which facilitates the identification of patterns, trends, and spatial relationships. It allows to evaluate the spatial distribution of phenomena and their relationship with environmental and/or socioeconomic variables.

This tool allows M-SALD to show its output as a thematic map in which it is possible to evaluate not only the hierarchical order of the watersheds regarding their water viability, but also, visually, it is possible to “hierarchically” order the potential points on the rivers according to the hierarchical order corresponding to the basin to which they belong (Fig. 2.5). In addition, it is possible to perform proximity and connectivity analysis in case of looking for areas close to a site of interest (industries, crops, livestock) to place the dam so that it is not only nearby, but that it is located in a highly hydrologically viable area.

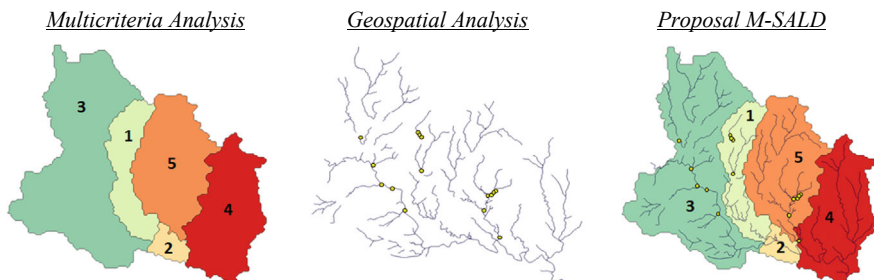


Fig. 2.5 Ranking of watershed alternatives and potential points on rivers

This allows to trace the spatial relationships that exist between the river basins and the potential points on the rivers, providing the specialist with a more precise analysis of the best places in which it is possible to locate the dam.

2.4 Results and Discussion

Case Study: The municipality of Manicaragua is located in the province of Villa Clara, in Cuba. It has various hydrological characteristics due to its mountainous topography and the presence of the *Sierra del Escambray* in the region. It has a network of rivers and streams that run through its territory, providing water sources for the population and agriculture. These watercourses can vary in flow rate and are important for the water supply of the area. It has reservoirs and dams that regulate the flow of water, allowing water to be stored for human, agricultural and industrial use, as well as for the generation of hydroelectric energy. Despite this, the collection of water by dams and reservoirs is insufficient because this municipality is responsible not only for supplying most of the province but also neighboring towns outside the province.

To improve sustainable water management and the development of the local community, it is necessary to find and select areas in which it is possible to build dams. This will increase the rainwater collection capacity of the municipality and bring dams closer to communal areas, crop areas or industrial areas. This approach constitutes an aid in terms of reducing the economic cost of transporting water from dams, as well as an increase in the supply of water resources to these areas.

Phase 1: Definition of the Decision-Making Framework

Main objective of the project: Select areas in which it is possible to build dams to increase the rainwater collection capacity of the municipality and bring dams closer to communal areas, crop areas or industrial areas.

Selection of the panel of experts: The group of experts is made up of four specialists, three from the National Institute of Hydraulic Resources (INRH) and one from the institute of hydraulic engineering of the Moscow State University of Civil Engineering.

Information collected: The main vector layers (.shp) for work in the GIS, created from the raster file (Fig. 2.6) extracted from the SRTM File Download Center “Earth Explorer” with spatial resolution of 28.7 m. Data from parameters that remain constant in the area were collected.

Expert criteria: The criteria and sub-criteria selected by the experts to evaluate the areas are:

- Hydrology C_1 : Main stream length p_{11} , Mainstream slope p_{12} , Time of concentration p_{13} , Maximum flow estimate p_{14} , Runoff coefficient p_{15} , Rainfall intensity p_{16} , Real evapotranspiration p_{17} , Average annual rainfall of the basin p_{18} , Sinuosity of water currents p_{19} , Average annual rainfall volume of the basin p_{110} ,

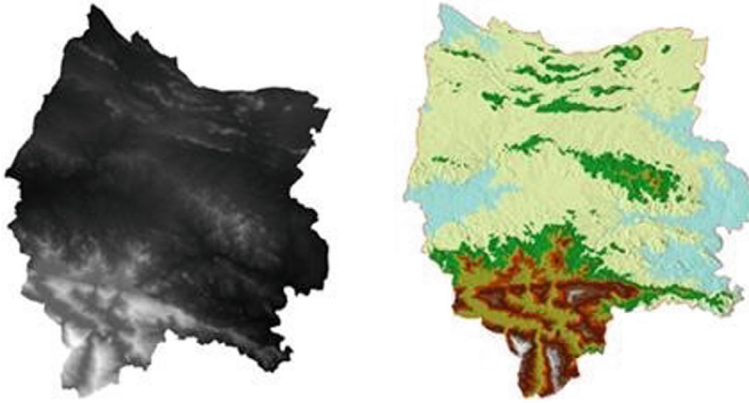


Fig. 2.6 Raster file from the municipality of Manicaragua, Cuba, extracted from the SRTM File Download Center “Earth Explorer”

Constant stability of the river p_{111} , Order of rivers p_{112} , Torrential coefficient p_{113} , Calculation of runoff coefficient p_{114} .

- Topography C_2 : Basin area p_{21} , Compactness index p_{22} , Form Factor p_{23} , Middle slope of the basin p_{24} , Drainage density p_{25} , Average elevation of the basin p_{26} , Watershed width p_{27} , Elongation index p_{28} .
- Geology C_3 : Coefficient of massiveness of the basin p_{31} , Orographic coefficient p_{32} .
- Socio-Economic C_4 : Delimitation of areas suitable for location p_{41} .

The model is completely independent of the parameters, this means that before using the M-SALD it is necessary for the expert group to select which parameters should be considered during the evaluation. It is valid to emphasize that the more parameters are used in the analysis, the more accurate the proposed selection of the areas by the model will be.

One of the limitations currently present in the geological analysis of soils is the absence of databases containing all the information regarding the geotechnical and geomechanical characteristics and properties of the soil foundation. For example, the properties of the soil and the existing rocks in each of the areas of analysis. That is why during the application of the M-SALD to the Manicaragua area in the province of Villa Clara, Cuba, the authors of this article resorted to the estimation of parameters that are easier to measure on site as is the case of the orographic coefficient and the massiveness coefficient. This does not mean that if at some point this type of data were to be available, it could not be included in the analysis (Fig. 2.7).

Phase 2: Components

K1: Definition of Evaluation Criteria and Parameters

Step 1.1: Definition of the objective, the fundamental criteria, parameters, alternatives (watersheds) and hierarchical structure.

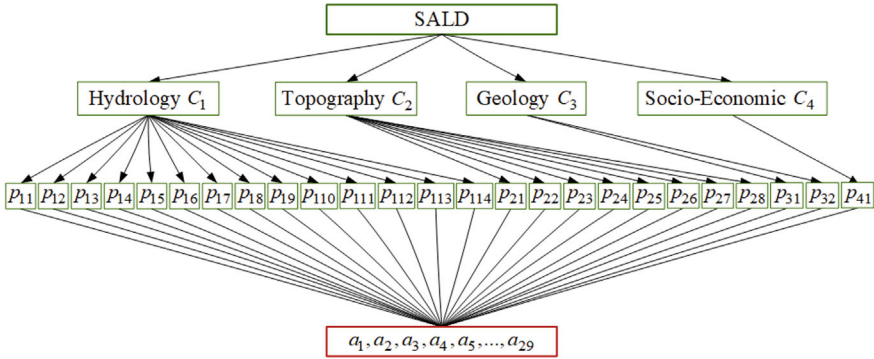


Fig. 2.7 Hierarchical scheme. Model for decision making. Objective, criteria, sub-criteria and alternatives

K2: Data Collection for Evaluation

From a DEM with spatial resolution of 28.7 m corresponding to the municipality of Manicaragua, data from 29 basins were extracted for subsequent evaluation. The model determined that 5 of them (Basin 1, Basin 8, Basin 9, Basin 24, Basin 28) will not be part of the group of alternatives that will finally be evaluated. This is because basins 1, 24 and 28 turned out to be false positives and, in the case of basins 8 and 9, it was detected that their area does not have fluvial tributaries according to the analysis of the fluvial layer provided by the experts.

The parameters that experts consider remaining constant for the region are:

- Run-off coefficient = 0.29;
- Annual potential evapotranspiration coefficient = 0.71;
- Average annual rainfall of the basin = 1375 mm;
- Average runoff = 400 mm;

K3: Evaluation of the Criteria, Parameters, and Alternatives

Step 3.1: Establishing the priority of criteria and parameters

Step 3.2: Establishment of local and global priorities for criteria and parameters

For the paired comparison of the criteria and to obtain the hierarchy between the areas to be evaluated, the hybrid method of multicriteria analysis AHP-TOPSIS is used. Therefore, paired matrices are created between the criteria for obtaining the priority vector (\vec{w}) or weights (Table 2.2). The AHP method proposes to evaluate the consistency of the paired comparison matrix (CR) obtained to know if the evaluations were correctly performed (Fig. 2.8).

As $RC < 0.10$; it is concluded that the paired comparison matrix has an allowable inconsistency (Table 2.3).

As $RC < 0.10$; it is concluded that the paired comparison matrix has an allowable inconsistency.

Table 2.5 Priority vector of each alternative with respect to C_1

C_1	A_2	A_3	A_4	A_5	A_6	A_7	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}
\vec{w}	0.03	0.03	0.04	0.04	0.05	0.05	0.04	0.04	0.04	0.03	0.08	0.04
	A_{16}	A_{17}	A_{18}	A_{19}	A_{20}	A_{21}	A_{22}	A_{23}	A_{25}	A_{26}	A_{27}	A_{29}
\vec{w}	0.03	0.03	0.03	0.05	0.03	0.04	0.05	0.04	0.03	0.03	0.04	0.03

Step 3.4: Establishment of the decision matrix of the alternatives

Once the priority vector matrices have been created for each alternative with respect to the parameters, the next step is to create the priority vector matrix for each alternative with respect to the criteria:

K4: Definition of Decision Rules

Step 4.1: Establishment of the standardized decision matrix

The normalized matrix (Table 2.8) is calculated using vectors obtained in Tables 2.6 and 2.7.

Step 4.2: Establishment of the weighted normalized decision matrix

Step 4.3: Definition of ideal positive and negative solutions.

Table 2.6 Priority vector matrix of each alternative with respect to the criteria

	C_1	C_2	C_3	C_4		C_1	C_2	C_3	C_4
A_2	0.03	0.04	0.06	0.04	A_{13}	0.03	0.02	0.02	0.04
A_3	0.03	0.03	0.02	0.04	A_{14}	0.08	0.04	0.02	0.04
A_4	0.04	0.04	0.02	0.04	A_{15}	0.04	0.07	0.2	0.04
A_5	0.04	0.03	0.02	0.04	A_{16}	0.03	0.05	0.02	0.04
A_6	0.05	0.04	0.02	0.04	A_{17}	0.03	0.04	0.02	0.04
A_7	0.05	0.03	0.02	0.04	A_{18}	0.03	0.02	0.02	0.04
A_{10}	0.04	0.04	0.02	0.04	A_{19}	0.05	0.04	0.02	0.04
A_{11}	0.04	0.03	0.02	0.04	A_{20}	0.03	0.03	0.02	0.04
A_{12}	0.04	0.05	0.11	0.04	A_{21}	0.04	0.04	0.02	0.04
A_{22}	0.05	0.07	0.04	0.04	A_{26}	0.03	0.05	0.02	0.04
A_{23}	0.04	0.04	0.02	0.04	A_{27}	0.04	0.06	0.08	0.04
A_{25}	0.03	0.02	0.02	0.04	A_{29}	0.03	0.07	0.14	0.04

Table 2.7 Matrix of criteria priority vector with respect to the target

	C_1	C_2	C_3	C_4
\vec{w}	0.52	0.20	0.20	0.08

Table 2.8 Weights of the normalized matrix

	C ₁	C ₂	C ₃	C ₄		C ₁	C ₂	C ₃	C ₄
A ₂	0.15	0.18	0.19	0.20	A16	0.15	0.23	0.06	0.20
A ₃	0.15	0.14	0.06	0.20	A17	0.15	0.18	0.06	0.20
A ₄	0.20	0.18	0.06	0.20	A18	0.15	0.09	0.06	0.20
A ₅	0.20	0.14	0.06	0.20	A19	0.25	0.18	0.06	0.20
A ₆	0.25	0.18	0.06	0.20	A20	0.15	0.14	0.06	0.20
A ₇	0.25	0.14	0.06	0.20	A21	0.20	0.18	0.06	0.20
A ₁₀	0.20	0.18	0.06	0.20	A22	0.25	0.32	0.13	0.20
A ₁₁	0.20	0.14	0.06	0.0	A23	0.20	0.18	0.06	0.20
A ₁₂	0.20	0.23	0.36	0.20	A25	0.15	0.09	0.06	0.20
A ₁₃	0.15	0.09	0.06	0.20	A26	0.15	0.23	0.06	0.20
A ₁₄	0.40	0.18	0.06	0.20	A27	0.20	0.28	0.26	0.20
A ₁₅	0.20	0.32	0.66	0.20	A29	0.15	0.32	0.46	0.20

The positive ideal set \bar{A}^+ of values and the negative ideal set \bar{A}^- of values are defined (Tables 2.9 and 2.10).

Table 2.9 Weights of normalized matrix

	C ₁	C ₂	C ₃	C ₄		C ₁	C ₂	C ₃	C ₄
A ₂	0.07	0.03	0.03	0.01	A16	0.07	0.04	0.01	0.01
A ₃	0.07	0.02	0.01	0.01	A17	0.07	0.03	0.01	0.01
A ₄	0.10	0.03	0.01	0.01	A18	0.07	0.01	0.01	0.01
A ₅	0.10	0.02	0.01	0.01	A19	0.13	0.03	0.01	0.01
A ₆	0.13	0.03	0.01	0.01	A20	0.07	0.02	0.01	0.01
A ₇	0.13	0.02	0.01	0.01	A21	0.10	0.03	0.01	0.01
A ₁₀	0.10	0.03	0.01	0.01	A22	0.13	0.06	0.02	0.01
A ₁₁	0.10	0.02	0.01	0.01	A23	0.10	0.03	0.01	0.01
A ₁₂	0.10	0.04	0.07	0.01	A25	0.07	0.01	0.01	0.01
A ₁₃	0.07	0.01	0.01	0.01	A26	0.07	0.04	0.01	0.01
A ₁₄	0.20	0.03	0.01	0.01	A27	0.10	0.05	0.05	0.01
A ₁₅	0.10	0.06	0.13	0.01	A29	0.07	0.06	0.09	0.01

Table 2.10 Positive ideal solution (PIS) and negative ideal solution (NIS)

	C ₁	C ₂	C ₃	C ₄
\bar{A}^+	0.20	0.06	0.13	0.01
\bar{A}^-	0.07	0.01	0.01	0.01

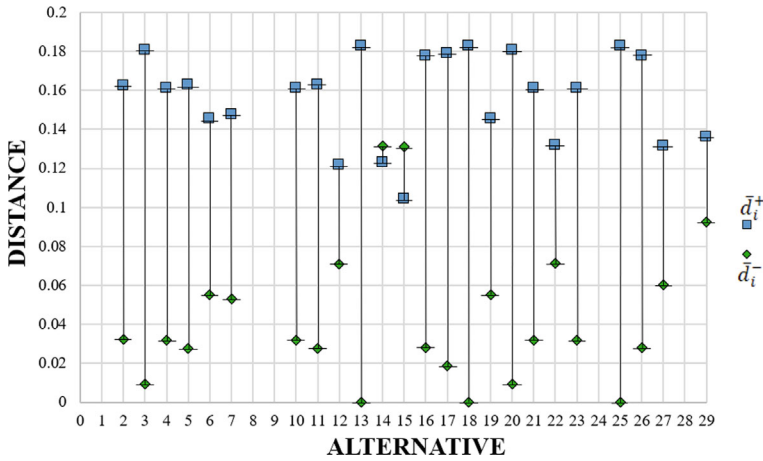


Fig. 2.9 Best ideal value and worst ideal value for each alternative

Step 4.4: Determination of distances between an alternative and an ideal positive/negative solution

Step 4.5: Evaluation of the proximity of the alternative to the ideal solution \bar{R}_i

The best alternatives are ordered according to their \bar{R}_i in descending order (Figs. 2.9, 2.10, 2.11 and 2.12).

K5: Geospatial Analysis.

Phase 3: Selection of Solutions

According to the model, and to the criteria and parameters used, basin № 15 is the best area to analyze the location of dams, while basin № 25, is, in turn, the worst classified for its use. In addition, of the 12,591 sites detected and evaluated as vertices on the river, the M-SALD reduced the areas suitable for the construction of dams by 22% (Fig. 2.13)

The results of the multicriteria analysis of M-SALD were compared with other analysis methods used in the literature in solving SALD problems. The indicators used in the comparison are the distribution of criteria, parameters and alternatives used during the construction of the paired comparison matrices (Table 2.11) and the three highest percentages obtained during the evaluation process of inconsistencies in each of the studied investigations ([14, 30, 45–48]) (Fig. 2.14). The result obtained from the comparison corroborates what is stated in this research work regarding that the increase in the number of variables in use does not influence the result and allows to reduce the inconsistencies obtained by maintaining the CR value below 0.02.

Spearman’s correlation [49] is used in the context of multicriteria analysis because it allows to evaluate the relationship between different criteria or variables in a non-linear way and without assuming a normal distribution of the data. This statistical

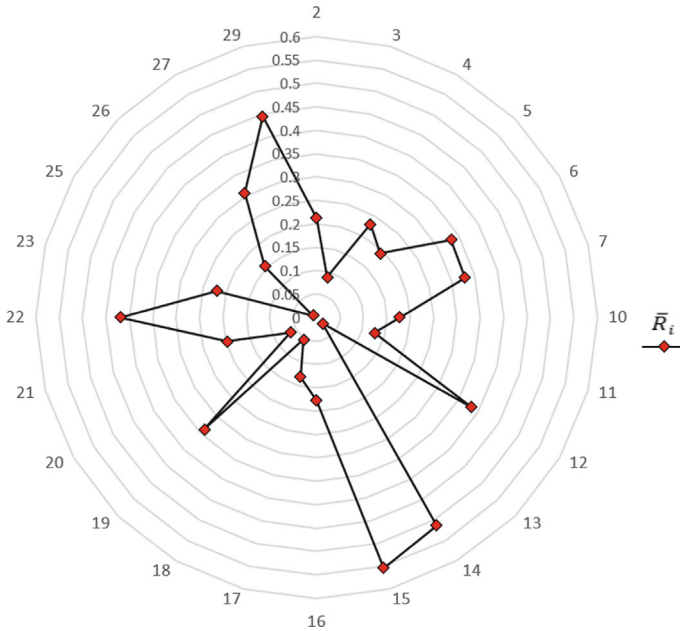


Fig. 2.10 Representation of the relative proximity \bar{R}_i to the ideal solution

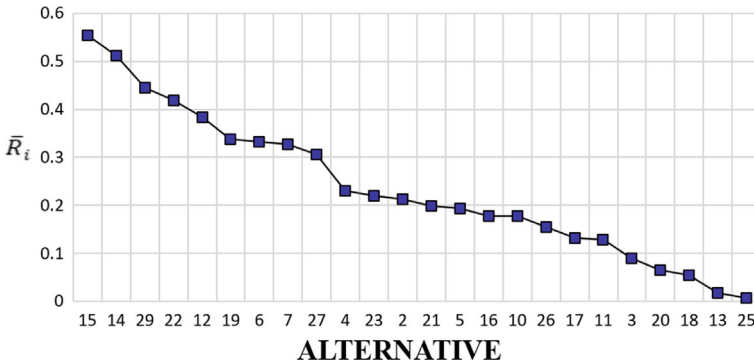


Fig. 2.11 Descending hierarchical order of the evaluated watersheds

correlation is used in the present research because it is useful for identifying the presence of monotonic relationships between the results of multicriteria analysis methods compared (M-SALD_AHP-TOPSIS, AHP_Saaty and TOPSIS using two normalization methods (Vector Normalization and Linear Normalization) [50–52]) and applied to the same case study. This result can be crucial for understanding how they relate to each other and how they affect decision-making. In addition, Spearman’s correlation helps to determine if there is consistency in the results obtained by each method,

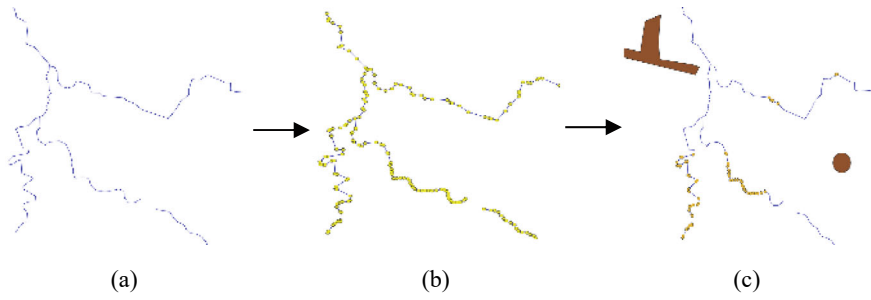


Fig. 2.12 Conversion process from a Line type object to a Vertex type object. **a** Example of a river represented in a vector layer by the Line type object. **b** A vector layer of a Line type object converted to a Vertex type object. **c** Analysis of the distance between the vertices of the rivers of the vector layer and the objects of the remaining vector layers (Cities, Dams)

Fig. 2.13 Non-dammed natural reservoir detected by M-SALD among the potential analysis points as a possible area for the location of dams



Table 2.11 Distribution by number of variables used by some authors of Fig. 2.14

	M-SALD	Zardari et al. [45]	Ahmad et al. [14]	Chezgi et al. [46]	Dai [29]	Hagos et al. [47]	Ghazali et al. [48]
Criteria	4	18	7	4	4	6	4
Parameters	25			12		17	
Alternatives	24		3	31	3	6	5

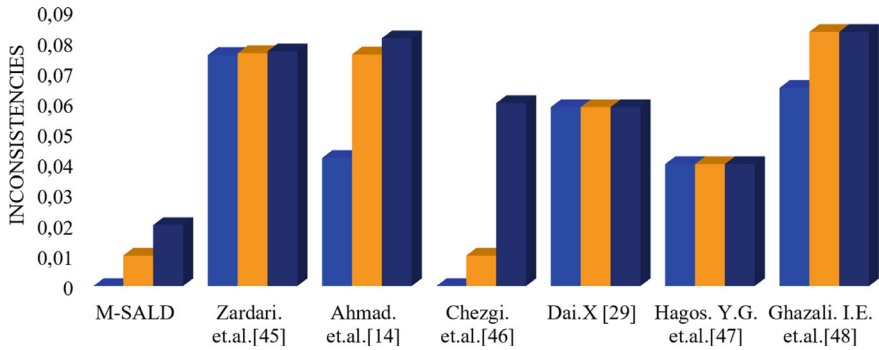


Fig. 2.14 Comparison between the inconsistencies obtained by different authors and M-SALD

identifying possible patterns of association between the evaluated criteria. This may be especially relevant to validate the robustness and reliability of the M-SALD model in the analysis.

The results (Fig. 2.15) show that there is a very high relationship between all the methods compared. It is valid to point out that M-SALD shows a higher correlation in both evaluations with respect to the original AHP_Saaty compared to the rest of the methods used. Its use contributes to a more informed and substantiated decision-making during the application of the model to a real case study.

2.5 Conclusions

The approach of artificial intelligence techniques such as expert systems oriented to solve SALD problems promotes the integration of specialized knowledge and technical criteria in a systematic and efficient way. The new M-SALD model developed allows to eliminate some of the shortcomings found so far in the literature. It uses specific rules and algorithms to evaluate data and generate recommendations based on logical and structured analysis. This helps to optimize decision-making, reduces subjectivity and minimizes the risk of errors during the selection process. M-SALD allows the processing of large amounts of data quickly, guaranteeing an exhaustive analysis in a short time and facilitating the identification of optimal locations for the construction of dams. It integrates the processes of multicriteria analysis and geospatial analysis, allowing to consider multiple factors simultaneously, considering the spatial location of the data and the interaction between them. In addition, the real value of the parameters that directly influence the areas and that are used during the evaluation and weighting of the alternatives is extracted from the raster and vector layers of the area in question. This facilitates the identification of optimal areas for the construction of dams, maximizing their efficiency and minimizing possible environmental impacts which is the most important factor for the model during the analysis process. Having the environmental factor as the fundamental axis of the model, a

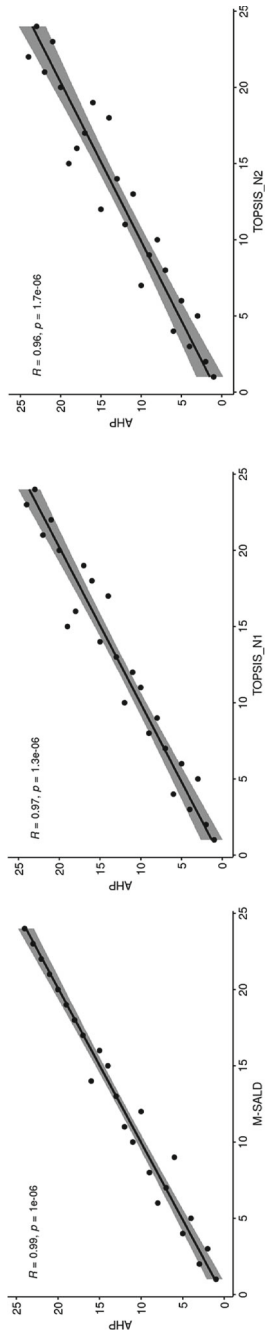


Fig. 2.15 Comparison of results of M-SALD_AHP-TOPSIS, AHP_Saaty, TOPSIS method using two normalization methods

sustainable and balanced development is guaranteed that benefits both the infrastructure and the natural environment, minimizing possible alterations to ecosystems and promoting responsible management of natural resources.

M-SALD has some limitations such as: (1) the efficiency in terms of the solutions provided by the model depends largely on the quality of the vector and raster layers that are used during the analysis, because they are where the real values of each of the parameters that are used during the evaluation and weighting of the resulting alternatives will be obtained. (2) The model is oriented for the study of large areas, that is, areas containing more than 1 watershed or sub-watershed. (3) So far, the multicriteria analysis is not included in the model directly on the points obtained on the rivers by the geospatial analysis, therefore, the hierarchical order obtained from them is directly related to the hierarchical position obtained by the watershed to which they belong. All these limitations are part of future studies to be considered to improve the work with the proposed new model.

References

1. Semenov, S.S., Poltavsky, A.V., Maklakov, V.V., Kryanev, A.V.: Overview of decision-making methods in the development of complex technical systems. *Reliability* **0**, 72–96 (2014). <https://doi.org/10.21683/1729-2646-2014-0-3-72-96>
2. Othman, A.A., Al-Maamar, A.F., Al-Manmi, D.A.M.A., Liesenberg, V., Hasan, S.E., Obaid, A.K., Al-Quraishi, A.M.F.: GIS-based modeling for selection of dam sites in the Kurdistan Region, Iraq. *ISPRS Int. J. Geo-Information*. **9**, 244 (2020). <https://doi.org/10.3390/ijgi9040244>
3. Rane, N.L., Achari, A., Choudhary, S.P., Mallick, S.K., Pande, C.B., Srivastava, A., Moharir, K.N.: A decision framework for potential dam site selection using GIS, MIF and TOPSIS in Ulhas river basin, India. *J. Clean. Prod.* **423** (2023). <https://doi.org/10.1016/j.jclepro.2023.138890>
4. Alrawi, I., Chen, J., Othman, A.A., Ali, S.S., Harash, F.: Insights of dam site selection for rainwater harvesting using GIS: a case study in the Al- Qalamoun Basin Syria. *Heliyon*. **9**, e19795 (2023). <https://doi.org/10.1016/j.heliyon.2023.e19795>
5. Zytoon, A., Gharineiat, Z., Alajarmeh, O.: Supplementary dam site selection using GIS-remote sensing approach: a case study of Wivenhoe dam, Preprints.org (2024). <https://doi.org/10.20944/preprints202404.0244.v1>
6. Dortaj, A., Maghsoudy, S., Doulati Ardejani, F., Eskandari, Z.: A hybrid multi-criteria decision making method for site selection of subsurface dams in semi-arid region of Iran. *Groundw. Sustain. Dev.* **10**, (2020). <https://doi.org/10.1016/j.gsd.2019.100284>
7. Alkhuzai, K.A., Mohammed, N.Z.: Digital elevation model analysis for dam site selection using GIS. *World J. Eng. Res. Technol.* **8**, 73–85 (2022). <https://doi.org/10.33774/coe-2022-gmnb9-v2>
8. Bihon, Y.T., Meshesha, M.A., Melese, D.W., Beyene, T.K., Kifle, T., Mihretu, E.N.: Suitable dam site selection with GIS-based sensitivity analysis of factors weight determination (In Birr River, Upper Blue Nile, Ethiopia). *Int. J. Innov. Sci. Res. Technol.* **7**, (2022). <https://doi.org/10.5281/zenodo.7067750>
9. Atiq, M.Z., Arslan, M., Baig, Z., Ahmad, A., Tanveer, M.U., Akhtar, A., Naeem, K., Mahmood, S.A.: Dam site identification using Remote Sensing and GIS (a case study Diemer Basha dam site). *Int. J. Agric. Sustain. Dev.* **1**, 168–178 (2019). <https://doi.org/10.33411/ijist/2019010412>
10. Ajibade, T.F., Nwogwu, N.A., Ajibade, F.O., Adelodun, B., Idowu, T.E., Ojo, A.O., Iji, J.O., Olajire, O.O., Akinmusere, O.K.: Potential dam sites selection using integrated techniques of

- remote sensing and GIS in Imo State, southeastern, Nigeria. *Sustain. Water Resour. Manag.* **6**, (2020). <https://doi.org/10.1007/s40899-020-00416-5>
11. Wang, Y., Tian, Y., Cao, Y.: Dam siting: a review. *water* **13**, 2080 (2021). <https://doi.org/10.3390/W13152080>
 12. Rodríguez Vázquez, S., Mokrova, N. V.: An M-SALD_AHP and GIS-based approach to watershed analysis during the development of small and large-scale hydraulic structure construction projects. *Lecture Notes Computer Science (including Subser. Lecture in Notes Artificial Intelligence. Lecture Notes in Bioinformatics)*, 14335 LNCS, 48–59 (2024). https://doi.org/10.1007/978-3-031-49552-6_5/FIGURES/5
 13. Rodríguez-Vázquez, S.: Multicriteria and geospatial model for the problem of selecting areas for the location of dams. (2023)
 14. Ahmad, I., Verma, M.K.: Application of analytic hierarchy process in water resources planning: a GIS based approach in the identification of suitable site for water storage. *Water Resour. Manag.* **32**, 5093–5114 (2018). <https://doi.org/10.1007/s11269-018-2135-x>
 15. Funk, A., Martínez-López, J., Borgwardt, F., Trauner, D., Bagstad, K.J., Balbi, S., Magrath, A., Villa, F., Hein, T.: Identification of conservation and restoration priority areas in the Danube River based on the multi-functionality of river-floodplain systems. *Sci. Total. Environ.* **654**, 763–777 (2019). <https://doi.org/10.1016/j.scitotenv.2018.10.322>
 16. Hermoso, V., Filipe, A.F., Segurado, P., Beja, P.: Freshwater conservation in a fragmented world: dealing with barriers in a systematic planning framework. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **28**, 17–25 (2018). <https://doi.org/10.1002/aqc.2826>
 17. Aires, U.R.V., Santos, B.S.M., Coelho, C.D., da Silva, D.D., Calijuri, M.L.: Changes in land use and land cover as a result of the failure of a mining tailings dam in Mariana, MG, Brazil. *Land use policy* **70**, 63–70 (2018). <https://doi.org/10.1016/j.landusepol.2017.10.026>
 18. Singh, L.K., Jha, M.K., Chowdary, V.M.: Multi-criteria analysis and GIS modeling for identifying prospective water harvesting and artificial recharge sites for sustainable water supply. *J. Clean. Prod.* **142**, 1436–1456 (2017). <https://doi.org/10.1016/J.JCLEPRO.2016.11.163>
 19. Rodríguez-Vázquez, S., Mokrova, N.V.: Hybrid system (M-SALD) of multicriterial analysis as a decision support tool for the selection of areas for the construction of hydraulic structures. In: Piñero Pérez, P.Y., Bello Pérez, R.E., Kacprzyk, J. (eds.) *Studies in Computational Intelligence*, pp. 401–415. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97269-1_22
 20. Bezukhov, D.A., Golos, V.N., Panin, A.V.: Assessment of the sediment delivery coefficient of small catchments in forest-steppe and steppe regions of the east European plain. In: *Proceedings of the Russian Academy of Sciences. The series is geographical*, pp. 73–84 (2019). <https://doi.org/10.31857/S2587-55662019473-84>
 21. Shumie, M.C.: River slope and roughness impact on downstream hydraulic structures. *J. Earth Sci. Clim. Change.* **9**, 1–7 (2018). <https://doi.org/10.4172/2157-7617.1000500>
 22. Solodovnikov, A.B.: Methods for determining the time of basin run-up in small catchments. *Designing the development of the regional railway network*, pp. 32–57 (2018)
 23. Jozaghi, A., Alizadeh, B., Hatami, M., Flood, I., Khorrami, M., Khodaei, N., Ghasemi Tousi, E.: A Comparative Study of the AHP and TOPSIS techniques for dam site selection using GIS: a case study of Sistan and Baluchestan province, Iran. *Geosciences*, **8**, 494 (2018). <https://doi.org/10.3390/geosciences8120494>
 24. Mollalo, A., Sadeghian, A., Israel, G.D., Rashidi, P., Sofizadeh, A., Glass, G.E.: Machine learning approaches in GIS-based ecological modeling of the sand fly *Phlebotomus papatasi*, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran. *Acta Trop.* **188**, 187–194 (2018). <https://doi.org/10.1016/j.actatropica.2018.09.004>
 25. Mollalo, A., Alimohammadi, A., Shirzadi, M.R., Malek, M.R.: Geographic information system-based analysis of the spatial and spatio-temporal distribution of zoonotic cutaneous leishmaniasis in Golestan Province, north-east of Iran. *Zoonoses Public Health* **62**, 18–28 (2015). <https://doi.org/10.1111/zph.12109>
 26. Nsanziyera, A., Rhinane, H., Oujaa, A., Mubea, K.: GIS and Remote-Sensing Application in Archaeological Site Mapping in the Awsard Area (Morocco). *Geosciences* **8**, 207 (2018). <https://doi.org/10.3390/geosciences8060207>

27. Rodríguez-Vázquez, S., Mokrova, N.V.: AHP-TOPSIS hybrid decision support system for dam site selection. *Mag. Civ. Eng.* **114**, 11405–11405 (2022). <https://doi.org/10.34910/MCE.114.5>
28. Koohbanani, H., Barati, R., Yazdani, M., Sakhdari, S., Jomemanzari, R.: Groundwater recharge by selection of suitable sites for underground dams using a GIS-based fuzzy approach in semi-arid regions. In: *Progress in River Engineering & Hydraulic Structures*, pp. 11–38. International Energy and Environment Foundation, Najaf, Iraq (2018)
29. Jamali, A.A., Randhir, T.O., Nosrati, J.: Site Suitability Analysis for Subsurface Dams Using Boolean and Fuzzy Logic in Arid Watersheds. *J. Water Resour. Plan. Manag.* **144**, 04018047 (2018). [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000947](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000947)
30. Dai, X.: Dam site selection using an integrated method of AHP and GIS for decision making support in Bortala, northwest China. https://webapps.itc.utwente.nl/librarywww/papers_2016/msc/gem/dai.pdf (2016)
31. Saaty, T.L.: *The analytic hierarchy process: planning, priority setting, resource allocation*, McGraw-Hill International Book Co (1980)
32. Pathan, A.L., Agnihotri, P.G., Patel, D.: Integrated approach of AHP and TOPSIS (MCDM) techniques with GIS for dam site suitability mapping: a case study of Navsari city, Gujarat, India. *Environ. earth Sci.* **81**, 443 (2022). <https://doi.org/10.1007/s12665-022-10568-6>
33. Abdullah, T., Ali, S., Al-Ansari, N., Knutsson, S.: Possibility of groundwater pollution in Halabja Sidsadiq hydrogeological basin, Iraq Using Modified DRASTIC Model Based on AHP and Tritium Isotopes. *Geosciences* **8**, 236 (2018). <https://doi.org/10.3390/geosciences8070236>
34. Noori, A., Bonakdari, H., Morovati, K., Gharabaghi, B.: The optimal dam site selection using a group decision-making method through fuzzy TOPSIS model. *Environ. Syst. Decis.* **38**, 471–488 (2018). <https://doi.org/10.1007/s10669-018-9673-x>
35. Kharazi, P., Yazdani, M.R., Khazaelpour, P.: Suitable identification of underground dam locations, using decision-making methods in a semi-arid region of Iranian Semnan Plain. *Groundw. Sustain. Dev.* **9**, 100240 (2019). <https://doi.org/10.1016/J.GSD.2019.100240>
36. Tsolaki-Fiaka, S., Bathrellos, G., Skilodimou, H.: Multi-criteria decision analysis for an abandoned quarry in the Evros region (NE Greece). *Land* **7**, 43 (2018). <https://doi.org/10.3390/land7020043>
37. Afshari, A., Vatanparast, M., Cockalo, D.: Application of multi criteria decision making to urban planning: a review. *J. Eng. Manag. Compet.* **6**, 46–53 (2016). <https://doi.org/10.5937/jemc1601046A>
38. Rahman, M.A., Jaumann, L., Lerche, N., Renatus, F., Buchs, A.K., Gade, R., Geldermann, J., Sauter, M.: Selection of the best inland waterway structure: a multicriteria decision analysis approach. *Water Resour. Manag.* **29**, 2733–2749 (2015). <https://doi.org/10.1007/s11269-015-0967-1>
39. Rodríguez-Vázquez, S., Mokrova, N.V.: Multi-criteria assessment of territorial planning alternatives using geographic information system technologies for dam construction. *Russ. J. Resour. Conserv. Recycl.* **7**, 1–17 (2020). <https://doi.org/10.15862/11INOR120>
40. Saaty, R.W.: The analytic hierarchy process—what it is and how it is used. *Math. Model.* **9**, 161–176 (1987). [https://doi.org/10.1016/0270-0255\(87\)90473-8](https://doi.org/10.1016/0270-0255(87)90473-8)
41. Rodríguez-Vázquez, S., Mokrova, N.V.: Comparison of applicability of different computational geometry algorithms for the detection of vertices in river layers in GIS systems. In: *2020 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon)*, pp. 1–4. IEEE (2020). <https://doi.org/10.1109/FarEastCon50210.2020.9271660>
42. Rodríguez-Vázquez, S., Mokrova, N.V.: Hybrid algorithms for geospatial analysis of dam location points in protective tasks for protected areas. *Her. Daghestan State Tech. Univ. Tech. Sci.* **48**, 40–49 (2021). <https://doi.org/10.21822/2073-6185-2021-48-2-40-49>
43. Rodríguez-Vázquez, S., Mokrova, N.V.: Vector-spatial analysis of GIS application layers for placing strategic points in dam design. *Constr. Ind. Saf.* **20**, 43–51 (2021). <https://doi.org/10.37279/2413-1873-2021-20-43-51>
44. Asamblea Nacional del Poder Popular: Ley No.124/17 (GOC-2017-715-EX51), <https://www.gacetaoficial.gob.cu/sites/default/files/goc-2017-ex51.pdf> (2017)

45. Zardari, N.H., Ahmed, K., Shirazi, S.M., Yusop, Z.Bin: weighting methods and their effects on multi-criteria decision making model outcomes in water resources management. Springer Cham, Malaysia (2015). <https://doi.org/10.1007/978-3-319-12586-2>
46. Chezgi, J., Pourghasemi, H.R., Naghibi, S.A., Moradi, H.R., Zarkesh, M.K.: Assessment of a spatial multi-criteria evaluation to site selection underground dams in the Alborz Province, Iran. *Geocarto Int.* **31**, 628–646 (2016). <https://doi.org/10.1080/10106049.2015.1073366>
47. Hagos, Y.G., Andualem, T.G., Mengie, M.A., Ayele, W.T., Malede, D.A.: Suitable dam site identification using GIS-based MCDA: a case study of Chemoga watershed. Ethiopia. *Appl. Water Sci.* **12**, (2022). <https://doi.org/10.1007/s13201-022-01592-9>
48. Ghazali, I.E., Yusof, N.M., Hamzah, N.: Evaluation of physical and mechanical properties of intact sandstones through the Analytic Hierarchy Process (AHP). In: IOP Conference Series: Earth Environmental Science, vol. 1151, (2023). <https://doi.org/10.1088/1755-1315/1151/1/012048>
49. Schober, P., Boer, C., Schwarte, L.A.: Correlation coefficients: appropriate use and interpretation. *Anesth. Analg.* **126**, 1763–1768 (2018). <https://doi.org/10.1213/ANE.0000000000002864>
50. Pomerol, J.-C., Barba-Romero, S.: *Multicriterion decision in management: principles and practice*. Kluwer Academic Publishers, Boston/Dordrecht/London (2000)
51. Kerr, C.S., Ibne Hossain, N.U., Jaradat, R.M.: Application of method for non-linear scaling of multi-criteria decision making attribute values. In: SYSCON 2020—14th Annual IEEE International System Conference Proceedings (2020). <https://doi.org/10.1109/SYSCON47679.2020.9275918>
52. Hsu-Shih Shih., Huan-Jyh Shyur, E., Stanley Lee: An extension of TOPSIS for group decision making. *Math. Comput. Model* **45**, 801–813 (2007). <https://doi.org/10.1016/j.mcm.2006.03.023>

Chapter 3

Multicriteria Hierarchical Ranking for Investment Selection in Latin American Countries



Manuel Muñoz Palma , Pavel Anselmo Álvarez Carrillo ,
Eva Luz Miranda Espinoza , Francisco Vargas Serrano ,
and Ernesto León-Castro 

Abstract The article aims to evaluate a set of Latin American countries to identify the level of risk as an investment option. Financial indicators and macroeconomic variables that present two grouping levels are considered to generate a ranking of prospective investment countries using the multicriteria hierarchical approach based on the ELECTRE III method. This approach allows countries to be sorted by dimensions in a group of indicators applied as decision criteria. The main contribution is identifying countries with investment viability in specific evaluation dimensions. As a result, an ordering is obtained at a global level, and a subgroup ordering of criteria the preferences of investors.

Keywords Multicriteria hierarchical process · ELECTRE III · Investment risk

M. M. Palma (✉) · F. V. Serrano
Universidad de Sonora, Hermosillo, Mexico
e-mail: manuel.munoz@unison.mx

F. V. Serrano
e-mail: francisco.vargas@unison.mx

P. A. Á. Carrillo · E. L. M. Espinoza
Universidad Autonoma de Occidente, Los Mochis, Sinaloa, Mexico
e-mail: pavel.alvarez@uadeo.mx

E. L. M. Espinoza
e-mail: mirandaeva@hotmail.com

E. León-Castro
Universidad Catolica de la Santisima Concepcion, Bío Bío, Chile
e-mail: eleon@ucsc.cl

3.1 Introduction

In countries with emerging economies in Latin America, evaluating the degree of risk for institutional investment integrates the different economic indicators that affect the development of the economies due to a fundamental factor in decision-making by national and foreign institutional investors. Sovereign bonds issued by countries represent a nodal element for risk rating; these are rated by the rating agencies Moody's, Fitch Ratings, and Standard and Poor's (S&P), a reference used by investment decision-makers of national and international brokerages.

Risk as a specific indicator of emerging countries is a critical decision reference for foreign investment [1]. The economic effect of COVID-19 worldwide is unique due to the speed and damage to the global economy. In 2021, the reopening and operating production activities of a more significant percentage of companies and the economic reactivation of the countries will restart. According to the World Bank, growth of 2.3% in 2022 and 2.2% in 2023 is expected in regional GDP in Latin American countries.

The decision-making process is complicated for institutional investors, particularly in today's interconnected economic environment. Investors must consider various criteria to maximize the value of their assets while minimizing the risk. However, maximizing the value of assets cannot be the sole objective of investors. Other goals that unite the participants in the organization should also be considered. Unfortunately, most investment models do not consider the multidimensional nature of the problem, making it challenging to make informed investment decisions. In this way, investors must adopt more sophisticated methods to include more criteria in their decisions.

Multicriteria decision-making (MCDM) is a useful alternative that offers a range of techniques and methods to classify and choose the best investment options based on their degree of risk and the investor's risk profile. MCDM assists decision-makers in identifying the most satisfactory solution to their decision-making problem and the best solution among alternatives [2]. Investors often consider various economic and financial indicators to determine the most attractive countries for investment. However, the current scenario is uncertain, with increasing economic, social, and environmental factors. Moreover, there are now more conflicting criteria to consider when ranking countries for investment. Despite this complexity, it is important to consider the preferences of investors when generating such rankings [3].

The research is approached from quantitative and qualitative approaches; it focuses on the factors that influence decision-making to establish strategies in selecting an institutional investment portfolio in the face of the reopening of the countries' economies and its effects on its leading macroeconomic indicators due to the COVID-19 crisis. The importance of selecting an institutional investment portfolio in countries is addressed in various investigations. However, current approaches do not consider the investor profile and the existence of contradictory criteria. In this sense, it is necessary to use analytical tools in decision-making processes to make better decisions.

This work addresses the problem of selecting prospective investment countries in a multicriteria ordering of the problem, adapting to the hierarchical multicriteria process to include considerations of economic and financial criteria specific to each country. An analysis of the financial and macroeconomic indicators classified into various subgroups of criteria and comprehensive ordering is carried out. The findings may aid investors in making a more informed decision by considering individual preferences and potential locations for consideration in an investment decision-making process.

Finally, the work is organized as follows: Sect. 3.2 presents the literary review. Section 3.3 addresses the methodology of the multicriteria hierarchical process where the hierarchical version of the ELECTRE III method is considered. Section 3.4 analyzes the performance of the countries and the results obtained. Conclusions are shown in Sect. 3.5.

3.2 Theoretical Framework

The main challenge for those responsible for implementing countries' economic and financial policies is attracting national and foreign investment. Latin American countries are considered emerging economies due to the high volatility of their financial markets and more significant risk in their investments. Aizenman et al. [4] establishes the responsibility of carrying out a healthy and sustainable macroeconomic policy to achieve balance. For example, the growth of credit to obtain financing through the issuance of sovereign bonds has caused a crisis in the payment of its debt, as in the particular case of Argentina, due to non-payment having national and global macroeconomic implications in its systemic nature [5]. The above has caused the implementation of structural reforms due to increased unemployment, inequality, and poverty in the regions [6]. Institutional investors consider risk, particularly the risk rating on 10-year sovereign bonds, when making investment decisions. This risk rating correlates more with the S&P 500 stock market index [7]. However, risk rating agencies' role in foreign investment is prone to social and political instability and macroeconomic variables that may increase a country's risk indicator [8]. Therefore, it is important to manage risks and adopt sustainable corporate strategies for investment [9]. International investors with long-term positions are generally at higher risk of significant losses in the Mexican market due to financial crises and currency depreciation [10].

Investors looking to invest their financial resources in different countries face a complex decision-making process, especially in today's globalized economic environment. There are multiple factors to consider, and COVID-19 has made things even more complicated due to its rapid spread and impact on the global economy [11]. The interests of shareholders cannot be ignored, as it is essential to maximize the value of a company [12]. However, most existing investment models do not consider the multidimensional nature of the investment decision problem. To make the right decisions, investors must adopt more sophisticated methods incorporating

multiple criteria. Multicriteria decision-making (MCDM) methods are helpful tools for solving complex problems with high uncertainty, conflicting objectives, different forms of data and information, and multiple interests and perspectives [13]. These methods can also account for complex and evolving biophysical and socioeconomic systems [14]. Institutional investors seeking to allocate their financial resources for investment in different countries must navigate this complicated decision-making process.

In this context, traditional evaluation methods appear to have limitations as they only consider expected return and risk parameters, neglecting other relevant information. Moreover, most models fail to incorporate the multidimensional nature of investment decision-making, including the role of investor intuition. Experts' decision-making relies on intuition, which plays a significant role in the process, as mentioned by [15]. Core intuition, as defined by [16], refers to affectively charged judgments that arise through rapid, non-conscious, and holistic associations between different elements, including complex experience-based patterns that financial specialists use to arrive at decisions.

Multidimensional methods for analyzing and evaluating institutional investments in countries are gaining interest due to their ability to consider all factors involved in risk reduction, such as those evaluated during investment in projects [17]. Academics highly discuss the stock market, financial variables, and macroeconomic news sensitivity, as investors closely monitor economic data, announcements, political events, and regulatory mandates. There are several theoretical justifications for a relationship between the monetary and financial policies adopted by those responsible for their implementation in countries. The perception of an economic slowdown is enough to generate significant changes in stock market prices [18].

International and national brokerages propose a series of intelligent system techniques to solve the problem of institutional investment selection in countries. Among these, reinforcement learning [19–21], neural networks [22, 23], genetic algorithms [24–26], decision trees [27], support vector machines [28–30], and empowerment and weighting of experts [31, 32]. Although these researches attempt to interpret the state of the market and predict the future trend of the market, they are not beneficial for small investors because these techniques require a certain degree of experience. Furthermore, these techniques cannot help investors compare businesses on multiple ambiguity criteria [33].

3.3 Methodology

One of the essential characteristics of multicriteria analysis is to compare alternatives based on a series of criteria. Therefore, multicriteria ranking methods are designed to build a recommendation on a set of alternatives according to the preferences of the expert or decision maker.

To generate the ordering of the leading Latin American countries for investment, the multicriteria hierarchical process is applied to the economic and financial indicators for each of the main economic sectors of the countries: real sector, external sector, financial and monetary sector, and public. The data from the indicators of the Economic Commission for Latin America [34] and the Latin American Study Circle [35] were considered.

Figure 3.1 presents the framework of this research; five work stages are defined here. Stage 1 describes the investment problem. Here, decision criteria and alternatives are identified. In Stage 2, the primary data, which is the result of the leading economic and financial indicators of the main sectors of the economy, is presented. The expert's preferences are elicited to define parameter values. Stage 3 generates an outranking model from the expert's preferences and economic and financial indicators. Stage 4 corresponds to the exploitation of the preferential model. For this step, a distillation process is used to order the countries. In Stage 5, the result of the organization and analysis of the information by decision-makers to establish investment policies in the countries is presented. In this sense, the process and method consider the investor's profile and the level of risk that he is willing to accept. Next, the multicriteria hierarchical strategy and the ELECTRE III multicriteria method are described to establish the ordering of the countries in Latin America.

3.3.1 *Multicriteria Hierarchical Process*

In the MCDA process, the definition of a set of alternatives is developed $A = \{a_1, a_2, \dots, a_m\}$ and a coherent family of criteria $G = \{g_1, g_2, \dots, g_m\}$. Any MCDA method develops a comprehensive preference method as an aggregation procedure. The method generates a recommendation by ranking alternatives from best to worst. The first stage of the investment selection problem consists of developing a country evaluation ranking. For this problem, it is easy to observe the hierarchical structure of the decision criteria. Therefore, it is often the case that a practical application imposes a hierarchical structure [36]. For this reason, the multicriteria ranking of countries is generated with a new method, the multicriteria hierarchical process (MCHP).

A multicriteria analysis method in the classical approach analyzes the leading Latin American countries at the same level, evaluating all criteria simultaneously. This way, you can find which countries have the best and worst investment prospects. Still, you cannot understand how some subcriteria (subgroups of economic sectors) interact to evaluate the economic and financial indicators that impact the selection of the countries for investment. In this sense, a different method would be valuable to assess countries by a subset of criteria at different levels following the MCHP methodology to solve the problem of selecting prospective countries for investment in the Latin American stock markets.

It is often the case that a practical application imposes a hierarchical criteria structure [36]. In the problem of investment selection of Latin American countries, there are many decision criteria; in fact, evaluating the selection of countries requires

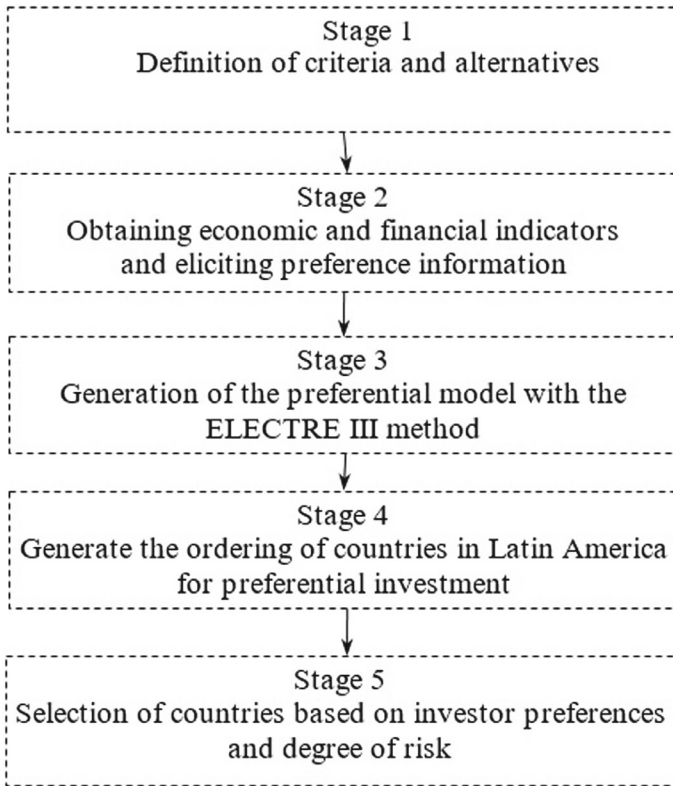


Fig. 3.1 Research model of the ordering of the leading Latin American countries

various types of information commonly addressed from the economic sectors and financial indicators. Considering these characteristics, the MCHP approach allows for decomposing the investment selection problem of the ten countries in Latin America into subproblems, considering a hierarchy of criteria to facilitate an analysis of the countries' economic policies.

To address decision-making problems where evaluation criteria are considered at the same level, a hierarchical structure is instead used to organize them into a part of the problem. The basic idea of MCHP is based on considering preference relationships at each node of the hierarchical criteria tree. These preference relationships refer to both the phase of obtaining preference information and the phase of analysis of a final recommendation by the decision maker [36].

A hierarchical structure of criteria can be viewed as a criteria tree. The tree structure takes a particular interest on the part of the expert or decision maker and agglomerates a subset of criteria into leaves. The sheets decompose the problem into minor problems to understand the interaction in elementary criteria. However, the same problem can be analyzed into smaller problems, such as a hierarchy. In the tree

criteria structure, some leaves contain branches with more leaves, making a sub-issue tree. Corrente et al. [37] integrate the MCHP with the ELECTRE III method. To explain the ELECTRE III hierarchy, the notation of [38] is followed.

G	It is a comprehensive set of all criteria at all levels considered in the hierarchy.
G_0	is the root of the criteria.
l_G	is the set of indices of the criteria in G .
$E_G \subseteq l_G$	is the set of indices of the elementary criteria.
g	is the generic criterion (where r is a vector with length equal to the level of the criterion).
$g_{(r,1)}, \dots, g_{(r,n(r))}$	are the immediate subcriteria of the criterion g_r (located at the level below g_r).
$E(g_r)$,	is the subset of indices of all the elementary criteria descending from g_r .
$E(F)$	is the set of indices of an elementary criterion that descend from at least one criterion of the subfamily $F \subseteq G$ (that is, $E(F) = \bigcup_{g_r \in F} E(g_r)$).
G_r	is the set of subcriteria g_r located at the level l in the hierarchy (below g_r).

To better understand the previous notation, Level 1 contains the macro criteria in the hierarchical structure, and the elementary criteria that descend from these are decomposing the subproblem. The entire set of elementary criteria is included in E_g . A different approach to the multicriteria decision support problem can be implemented when a hierarchical structure is generated concerning the criteria of interest at a particular hierarchy level.

The problem of selecting investment prospects for countries in Latin America to integrate an investment portfolio can be addressed as a hierarchical problem, where some macro criteria can combine elementary criteria from a deeper level of the hierarchy. Figure 3.2 illustrates a summarized structure (two macro criteria) of the complete hierarchical problem of selected Latin American countries. The Real Sector macro criterion (g_1) integrates five elementary criteria, External Sector (g_2) integrates five elementary criteria, Financial and Monetary Sector (g_3) that integrates five elementary criteria and finally, the Public Sector (g_4) that integrate two elementary criteria. The evaluation of prospective investment countries in Latin America includes 17 elementary criteria and is structured in a two-level hierarchy. Four macro criteria (non-elementary criteria) are defined at the first level. At Level 2, 17 elementary criteria constitute the macro criteria of Level 1.

3.3.2 ELECTRE III Hierarchical Method

The adapted version of the ELECTRE III hierarchy was first introduced by [37]. The ELECTRE method is developed in two steps. The first step is the aggregation

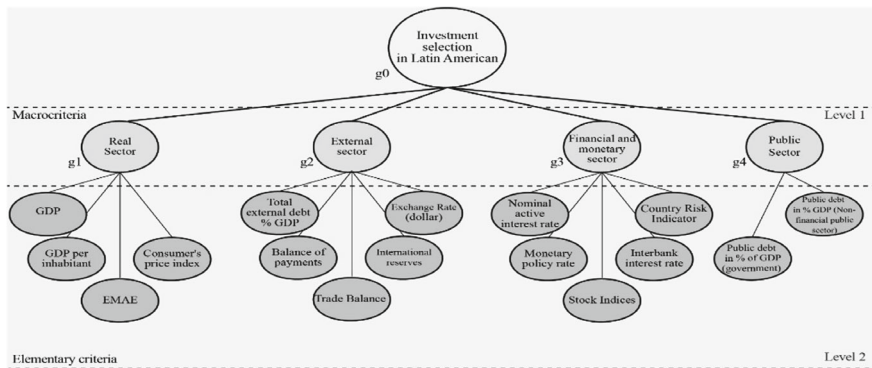


Fig. 3.2 Simplified MCHP structure for selecting countries for investment

of preferences, and the information is created by building a model in the valued improvement relationship. In the second step, the distillation process exploits the valued outperformance ratio, generating a partial or complete ranking of alternatives. For each elementary criterion $g_t, \in E_g$.

The elementary agreement index for each elementary criterion g_t

$$\phi_t(a, b) = \begin{cases} 1 & \text{if } g_t(b) - g_t(a) \leq q_t, (aS_t b) \\ \frac{p_t - [g_t(b) - g_t(a)]}{p_t - q_t} & \text{if } q_t < g_t(b) - g_t(a) < p_t, (bQ_t a) \\ 0 & \text{if } g_t(b) - g_t(a) \geq p_t, (bP_t a) \end{cases} \quad (3.1)$$

The elementary discordant index for each elementary criterion $g_t g_t$

$$d_t(a, b) = \begin{cases} 1, & \text{if } g_t(b) - g_t(a) \geq v_t, \\ \frac{[g_t(b) - g_t(a)] - p_t}{v_t - p_t} & \text{if } p_t < g_t(b) - g_t(a) < v_t, \\ 0, & \text{if } g_t(b) - g_t(a) \leq p_t. \end{cases} \quad (3.2)$$

The partial agreement index for each non-elementary criterion $g_i g_i$

$$C_r(a, b) = \frac{\sum_{t \in E(g_r)} w_t \phi_t(a, b)}{\sum_{t \in E(g_r)} w_t} \quad (3.3)$$

Partial credibility index

$$\sigma_r(a, b) = \begin{cases} C(a, b) x \prod_{g_t \in E(g_r)} \frac{1 - d_t(a, b)}{1 - C_r(a, b)} & \text{if } d_t(a, b) > C_r(a, b) \\ C(a, b) & \end{cases} \quad (3.4)$$

The valued outperformance relationship generated in the previous step corresponds to the decision maker's preferential model. The distillation method is used

to exploit the preferential model. The distillation occurs descending and ascending; therefore, the final preorder is obtained as the intersection of the two distillations. [39] describe an overview of the distillation method.

For the pair A , $a, b \in A$ in the hierarchical process, the alternatives are ordered in a partial or complete preorder for each non-elementary criterion g_r as follows:

$aP_r b$: a is strictly preferred to b in the macro criterion g_r if in at least one of the orderings, a is placed before b , and if in the other ordering a is at least as good as b .

$aI_r b$: a is indifferent to b in the macrocriterion g_r if the two stocks belong to the same position in the two preorders.

$aR_r b$: a is incomparable to b in the macrocriterion g_r if a is ordered better than b in the ascending distillation and b is better ranked than a in the descending distillation or vice versa.

Readers can find applications of the MCHP in different fields like competitiveness [40], innovation [41], and portfolio (stock evaluation) [42, 43].

3.4 Analysis of Prospective Investment Countries with the Multicriteria Hierarchical Process

The analysis is based on the economic and financial indicators corresponding to 2021. The economic and financial indicators were considered to select the macro criteria in the evaluation of the performance of each country [34, 35]. These give indications about the economic situation and prospects for its performance, as well as the evaluation of the position of a country compared to others (see Table 3.1). The data obtained is grouped into four dimensions to evaluate the countries with the best economic and financial performance. Each dimension comprises a subgroup of different indicators (elementary criteria); there are 17 indicators to assess the leading Latin American countries. The economic and financial indicators are used in this work with the multicriteria hierarchical approach to analyze the countries with the best economic and financial performances for investment concerning the interaction of subgroups of criteria at different levels in a hierarchy through the organization of Latin American countries. A similar study analyzing a country's situation for investment location is found in [44] using specific financial and macroeconomic criteria.

The macro criteria for the problem of selecting prospective investment countries, elementary criteria, and their corresponding weights are shown in Table 3.2. It reflects the adjacent preference of the expert. It should be stated the values are derived from the elicitation process supported by an extended version of the SRF developed by Corrente et al. [45] to support the definition of the weight in a hierarchy of criteria. Simos–Roy–Figueira (SRF) is the computational tool of Simos' Revised Procedure. We will use the term hierarchical deck of cards method (HDCM) in [41] to refer to this extended version addressing the investment analysis problem.

Table 3.1 Main Latin American countries

Label	Country
A1	Argentina
A2	Bolivia
A3	Brazil
A4	Chile
A5	Colombia
A6	Costa Rica
A7	Ecuador
A8	México
A9	Perú
A10	Uruguay

Source Own elaboration with data from ECLAC and CESLA

To the methodology proposed in Sect. 3.1, the MCHP is applied to solve the problem of selecting the countries with the most significant economic and financial viability for investment. In the first step, the problem is structured in a multicriteria hierarchy, decomposing the problem into four macro criteria as investment sub-problems of the countries. As shown in the hierarchical structure of Table 3.2, the investment prospect countries are structured in a hierarchy for the four macro criteria and the 17 elementary criteria. The new hierarchical structure for the country selection problem allows the analysis to move closer to the MCHP. This approach implemented in this research evaluates each macrocriterion, allowing the interaction between immediate descending subcriteria directly related to the macrocriterion to be analyzed. It is carried out by generating preferential models and arrangements for each macro criterion to understand how one country works compared to another and, at the same time, how the problem impacts the investment decision by institutional investors.

The ELECTRE III hierarchical and distillation methods of Sect. 3.2 were applied to solve each subproblem g_i (macro criterion) and the comprehensive level. Table 3.3 illustrates the comprehensive ranking g_0 which generates nine positions of the analyzed countries and assigns the countries of Chile in position one (A4); in position two are Peru (A9) and Mexico (A8); in position three Colombia (A5). Uruguay is located in the last position of the ranking for investment (A10) in the last position, position eight Argentina (A1), and position 7 Bolivia (A2). Figure 3.3 shows the levels of investment risk in Latin American countries (g_0).

Each macro-criterion is evaluated by a subset of sub-criteria (elementary criteria that belong to the last level of the hierarchy). Table 3.4 contains the orderings of each macro criterion ($g_1 \dots g_4$). The ordering results from the interaction of elementary criteria that evaluate the corresponding macro criteria. For the investment selection problem, we analyzed how the interaction of the subset of elementary criteria influences the macro criteria (Level 2 of the hierarchy) and then the interaction of the

Table 3.2 Macrocriteria and elementary criteria of the countries

Index	Macrocriterion	Subindex	Criteria	Wt
g ₁	Real sector	g _{1,1}	Total annual gross domestic product (GDP) at current prices in dollars	0.050
		g _{1,2}	Total annual gross domestic product (GDP) per inhabitant at current prices in dollars	0.050
		g _{1,3}	Monthly activity estimator (EMAE)	0.060
		g _{1,4}	Consumer's price index	0.030
g ₂	External sector	g _{2,1}	Total external debt as a percentage of gross domestic product	0.050
		g _{2,2}	Balance of payments	0.050
		g _{2,3}	Trade balance balance (Millions US\$)	0.060
		g _{2,4}	International reserves	0.080
		g _{2,5}	Exchange rate regarding the dollar	0.060
g ₃	Financial and monetary sector	g _{3,1}	Nominal active interest rate	0.045
		g _{3,2}	Monetary policy rate	0.035
		g _{3,3}	Stock Indices	0.090
		g _{3,4}	Interbank interest rate	0.050
		g _{3,5}	Country risk indicator	0.080
g ₄	Public sector	g _{4,1}	Public debt balance in percentages of GDP	0.075
		g _{4,2}	Public debt balance in percentages of GDP	0.075

Table 3.3 A comprehensive ranking of investment prospect countries

Rank	g ₀
1	A4
2	A8, A9
3	A5
4	A6
5	A7
6	A3
7	A2
8	A1
9	A10



Fig. 3.3 Investment risk map in the countries (g_0)

impact of macro criteria for the selection problem of investment prospect countries (Level 1).

The relative importance of the macro criteria is $g_3 = g_2 > g_1 > g_4$, with the weights 0.300; 0.300; 0.250; 0.150 respectively. The financial and monetary sector (g_3) shows the first positions for $A_4 > A_5 > A_9 > A_6$. The external sector (g_2) shows $A_8 > A_9 > A_4 > A_1 > A_6$. The real sector (g_1) shows $A_4 > A_7 > A_6 > A_8$; the public sector (g_4) shows $A_9 > A_4 > A_8 > A_2 > A_{10} = A_7 = A_5$.

In Tables 3.5, 3.6 and Figs. 3.4, 3.5, the results for each country are based on the macro criterion g_2 and g_3 are presented, which are the most important macro criteria. With this information, it is possible to visualize that a different ranking can be made based on each criterion. Still, considering the investor's weight and preference, this can change drastically. For example, in Table 3.5, considering C_6 the worst country is Uruguay, which is consistent with the rank in Table 3.3, but for this specific case, the best country is Mexico, with 34.31%, in the case of C_7 the best country is Chile and the worst is Colombia. The same analysis can be done to C_8 , C_9 and C_{10} . Finally,

Table 3.4 Individual ranking of investment prospect countries

Rank	g_1	g_2	g_3	g_4
1	A4	A8	A4	A9
2	A7	A9	A5	A4
3	A6	A4	A9	A8
4	A8	A1	A6	A2
5	A10	A6	A8	A10, A7, A5
6	A2	A2	A2	A6
7	A1	A3	A7	A3
8	A5	A7	A3	A1
9	A3	A5	A10	–
10	A9	A10	A1	–

in Fig. 3.4, it is possible to visualize with a lighter tone the best countries regarding the macro criterion g_2 and with a darker tone, the worst one.

In the case of Table 3.6, it is possible to visualize that according to C_{11} the most attractive country is Argentina, and the worst country is Bolivia. Something important about this specific indicator is that the interest rate is usually related to inflation and C_{15} , that Argentina has a higher interest rate and country risk. Continuing with the analysis of C_{15} the best countries are Chile and Uruguay, both with 131. Figure 3.5 shows the aggregated financial and monetary sector results: the lighter tone, the better ones, and the darker tone, the worse ones.

Table 3.5 Economic indicators of the external sector

Label	Country	Total external debt % of gross domestic product (C_6)	Balance of payments (C_7)	Balance of trades (Millions US\$) (C_8)	International reserves (C_9)	Exchange rate respect USD (C_{10})
A1	Argentina	69.67	3,312.75	13.99	41.53	6.15
A2	Bolivia	38.84	–188.62	1.67	4.76	1.25
A3	Brazil	44.13	–2,592.34	14.43	367.77	4.06
A4	Chile	82.62	3369.64	10.77	53.31	10.33
A5	Colombia	56.92	–9,926.64	–14.72	58.33	11.36
A6	Costa Rica	51.54	–1349.06	–3.62	6.92	4.72
A7	Ecuador	57.57	2564.53	2.87	7.57	0.00
A8	Mexico	34.31	2612.24	–5.91	202.40	0.02
A9	Peru	44.00	1582.71	13.15	78.32	4.11
A10	Uruguay	86.59	–315.78	–0.97	16.95	1.11

Table 3.6 Financial indicators of the financial and monetary sector

Label	Country	Nominal active interest rate (C_{11})	Monetary policy rate (C_{12})	Stock index (C_{13})	Interbank interest rate (C_{14})	Country risk indicator (C_{15})
A1	Argentina	36.70	39.66	-3.78	26.39	1684.00
A2	Bolivia	6.40	2.50	0.61	3.52	475.00
A3	Brazil	33.70	2.72	-3.56	1.90	253.00
A4	Chile	8.00	0.75	4.01	1.60	131.00
A5	Colombia	9.90	2.77	10.53	4.49	232.00
A6	Costa Rica	11.00	1.10	54.36	3.50	534.00
A7	Ecuador	8.90	8.58	-7.31	0.20	1263.00
A8	Mexico	30.20	5.25	23.88	4.26	378.00
A9	Peru	12.90	0.67	6.78	0.25	164.00
A10	Uruguay	12.60	7.89	-8.73	4.50	131.00

**Fig. 3.4** Investment risk map in countries of the external sector of the economy (g_2)



Fig. 3.5 Investment risk map in countries of the financial and monetary sector (g_3)

Table 3.7 shows the debt classification of countries in Latin America (sovereign bonds), published by the central rating agencies in the world, where the best option to invest in Chile (A_4); Likewise, within the global ranking carried out by the raters, Chile is in position 59 as the best qualified compared to other countries; Mexico is placed 60; Costa Rica 74; and Peru 76; in general, there are similarities. Therefore, considering that the criteria for evaluation are different, even within the rating agencies, as references in decision-making by institutional investors.



These variations are important to consider because it is possible to identify how much the ordering can change if different parameters are used with the same information. In this sense, the orderings are not absolute, but the preference and many other elements related to different quantitative parameters can change. Therefore, it is important to use methodologies that can be adapted to the reality of the decision maker for the integration of an investment portfolio based on the financial indicators and economic performance of the countries and by the profile and preferences of the countries.

Table 3.7 Evaluation by rating agencies in Latin American countries (2021)

Label	Country	Moody's	SP	Fitch	Global opportunity index	Doing business index
A ₁	Argentina	Ca ⁵	CCC+ ⁵	CCC ⁵	89	126
A ₂	Bolivia	B2 ⁴	ND	B ⁴	107	150
A ₃	Brazil	Ba2 ³	BB- ³	BB- ³	69	124
A ₄	Chile	A1 ¹	A+ ¹	A- ¹	39	59
A ₅	Colombia	Baa2 ³	BB- ³	BB- ³	67	67
A ₆	Costa Rica	B2 ⁴	B ⁴	B ⁴	55	74
A ₇	Ecuador	Caa3 ⁵	B- ⁴	B- ⁴	111	129
A ₈	Mexico	Baa1 ²	BBB ²	BBB- ²	54	60
A ₉	Peru	Baa ¹	BBB	BBB+ ²	62	76
A ₁₀	Uruguay	Baa2- ³	BBB	BBB- ²	48	101

Source <https://datosmacro.expansion.com/ratings>

Note 1 1 upper secondary degree; 2 medium–low grade; 3 non-speculative investment grade; 4 highly speculative degree; 5th degree extremely speculative

-  —Best
-  —Good
-  —Average
-  —Worst

3.5 Conclusions

This research analyzes the economic and financial performance of the leading Latin American countries in 2021. It evaluates the variables that affect the main economic sectors of the countries, with four macro-criteria and 17 elementary criteria. From a methodological perspective, the multicriteria hierarchical process was used to analyze the economic and financial policies of those responsible for monetary policies in each country. Subgroups of elementary criteria are evaluated to understand their interaction and the impact of each macro-criterion at the top level of the hierarchy. With this, the analysis process was applied, generating a preferential model and an ordering for each macro criterion, and a comprehensive ordering for the selection problem by institutional investors in the location of prospective investment countries for the integration of the portfolios after going through a period of economic paralysis due to the effects of the COVID-19 pandemic.

From the perspective of an increasingly interconnected economic environment and a growing number of conflicting criteria, in addition to the traditional objectives of maximizing value for investors and minimizing business risks, several different criteria must be considered in the location of institutional investment. In recent decades, Latin America has undergone a significant transformation in trade, labor market, tourism, technology, and financial markets, making it more attractive as an investment destination. However, the choice of brokerage investment is a complex issue, mainly because Latin American countries have experienced hyperinflation and political instability over the years, among other multidimensional aspects that must

be considered when looking at the problem through the lens of decision-maker's preferences.

The MCHP allows for the evaluation of subcriteria at all levels of the hierarchy to analyze the country's situation. The problem of institutional investment selection by national and international brokerages shows the opportunities and needs of companies and allows for more robust and reliable decision-making. MCHP is used to evaluate the formulation of more assertive policies and decisions in Latin American countries. Consequently, it would achieve favorable conditions to encourage the investor. In this sense, the ELECTRE III method provides decision support for real-world problems with a non-compensatory approach. The results of this analysis can help an investor to include individual preferences when making a shortlist of countries to consider in the investment decision-making problem. The research presents several considerations that have not been used in traditional methods and uses a hierarchical approach to analyze the performance of countries' financial and economic parameters.

A limitation of the current work is the arbitrary definition of the categories of risk derived from a ranking and not for a formal class definition in a sorting problem, where the categories are defined a priori.

For future lines of research, it is recommended to use a more complex method to aggregate the information, such as the weighted average operator (OWA) and some of its extensions, such as the induced heavy OWA operator or the prioritized OWA operator, as well as applications to different fields such as organizational innovation, circular economy, and finance, among others.

References

1. Limas Suárez, S. J., Franco Ávila, J. A.: El riesgo país para Colombia: interpretación e implicaciones para la economía y la inversión extranjera 2012–2017 *Revista Finanzas y Política Económica* **10**(1), 153–171 (2018). <https://doi.org/10.14718/revfinanzpolitecon.2018.10.1.6>
2. Stewart, T.: A critical survey on the status of multiple criteria decision-making theory and practice. *Omega* **20**, 569–586 (1992). [https://doi.org/10.1016/0305-0483\(92\)90003-P](https://doi.org/10.1016/0305-0483(92)90003-P)
3. Guerrero-Baena, D. D., Gómez-Limón, J. A., & Fruct Cardozo, V. V.: Are multicriteria decision making techniques useful for solving corporate finance problems? A bibliometric analysis. *Revista de Metodos Cuantitativos Para La Economía y La Empresa*. **17**(1), 60–79 (2014)
4. Aizenman, J., Glick, R.: Sterilization, monetary policy, and global financial integration. Federal Reserve Bank of San Francisco. Working Paper Series (2008). <https://doi.org/10.1111/j.1467-9396.2009.00848.x>
5. Nudelsman, S.: ¿Es posible mejorar la reestructuración de las deudas soberanas? *Problemas Del Desarrollo* **47**(184), 163–185 (2016). <https://doi.org/10.1016/j.rpd.2016.01.008>
6. Soto, R.: América Latina. Entre la Financiarización y el Financiamiento Productivo. *Problemas Del Desarrollo* **44**(173), 57–78 (2013). [https://doi.org/10.1016/S0301-7036\(13\)71875-3](https://doi.org/10.1016/S0301-7036(13)71875-3)
7. Piffaut, P. V., Rey Miró, D.: Integración, contagio financiero y riesgo bursátil: ¿qué nos dice la evidencia empírica para el periodo 1995–2016? *Cuadernos de Economía* **39**(111) (2016). <https://doi.org/10.1016/j.cesjef.2016.09.003>
8. Calahorrano, L., Tigse, S., Caicedo, F.: Variación del indicador riesgo-país en el flujo de inversión extranjera del ecuador. *Universidad Ciencia y Tecnología* **24**(107), (2020). <https://doi.org/10.47460/uct.v24i107.416>

9. Spulbar, C., Ejaz, A., Birau, R., Trivedi, J.: Sustainable investing based on momentum strategies in emerging stock markets: a case study for Bombay Stock Exchange (BSE) of India. *Sci. Ann. Eco. Bus.* **66**(3), (2019). <https://doi.org/10.2478/saeb-2019-0029>
10. Gutiérrez, R. de J., Ortiz, E.: El efecto de la volatilidad del peso mexicano en los rendimientos y riesgo de la Bolsa Mexicana de Valores. *Contaduría y Administración* **58**(3) (2013). [https://doi.org/10.1016/s0186-1042\(13\)71223-3](https://doi.org/10.1016/s0186-1042(13)71223-3)
11. Sharma, P., Leung, T.Y., Kingshott, R.P.J., Davcik, N.S., Cardinali, S.: Managing uncertainty during a global pandemic: An international business perspective. *J. Bus. Res.* **116**, 188–192 (2020). <https://doi.org/10.1016/j.JBUSRES.2020.05.026>
12. Jensen, M.: Value maximization, stakeholder theory, and the corporate objective function. *J. Appl. Corp. Financ.* **22**(1), 32–42 (2010). <https://doi.org/10.1111/j.1745-6622.2010.00259.x>
13. Castellanos, A., Cruz-Reyes, L., Fernández, E., Rivera, G., Gomez-Santillan, C., Rangel-Valdez, N.: Hybridisation of swarm intelligence algorithms with multi-criteria ordinal classification: a strategy to address many-objective optimisation. *Mathematics* **10**(3), 322 (2022). <https://doi.org/10.3390/math10030322>
14. Wang, J.J., Jing, Y.Y., Zhang, C.F., Zhao, J.H.: Review on multicriteria decision analysis aid in sustainable energy decision-making. *Renew. Sustain. Energy Rev.* **13**(9), 2263–2278 (2009). <https://doi.org/10.1016/j.rser.2009.06.021>
15. Salas, E., Rosen, M. A., Diaz Granados, D.: Expertise-based intuition and decision making in organizations. *J. Manag.* **36**(4), 941–973 (2010). <https://doi.org/10.1177/0149206309350084>
16. Dane, E., Pratt, M.G.: Exploring intuition and its role in managerial decision making. *Acad. Manag. Rev.* **32**(1), 33–54 (2007). <https://doi.org/10.5465/amr.2007.23463682>
17. Shvovetsa, O., Rodionova, E., Epstein, M.: Evaluation of investment projects under uncertainty: multicriteria approach using interval data. *Entrepreneurship Sustain. Issues* **5**(4), 914–928 (2018). [https://doi.org/10.9770/jesi.2018.5.4\(15\)](https://doi.org/10.9770/jesi.2018.5.4(15))
18. Peralta-Alva, A.: New technology may cause stock volatility. *The Reg. Econ.* (2012)
19. Moody, J., Wu, L., Liao, Y., Saffell, M.: Performance functions and reinforcement learning for trading systems and portfolios. *J. Forecast.* **17**(5), 441–471 (1998) [https://doi.org/10.1002/\(SICI\)1099-131X\(1998090\)17:5<6%3C441::AID-FOR707%3E3.0.CO;2-%23](https://doi.org/10.1002/(SICI)1099-131X(1998090)17:5<6%3C441::AID-FOR707%3E3.0.CO;2-%23)
20. Moody, J., Saffell, M.: Learning to trade via direct reinforcement. *IEEE Trans. Neural Networks* **12**(4), 875–889 (2001). <https://doi.org/10.1109/72.935097>
21. Jangmin, O., Lee, J.W., Zhang, B. T.: Stock trading system using reinforcement learning with cooperative agents. In: *Proceedings of the 19th International Conference on Machine Learning*, pp. 451–458 (2002)
22. Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M.: Stock market prediction system with modular neural networks. *Neural Networks in Finance Investing*, 343–357 (1993). <https://doi.org/10.1109/IJCNN.1990.137535>
23. Dempster, M.A.H., Payne, T.W., Romahi, Y., Thompson, G.W.T.: Computational learning techniques for intraday FX trading using popular technical indicators. *IEEE Trans. on Neural Networks* **12**(4), 744–754 (2001). <https://doi.org/10.1109/72.935088>
24. Mahfoud, S., Mani, G.: Financial forecasting using genetic algorithms. *Appl. Artif. Intell.* **10**(6), 543–565 (1996)
25. Allen, F., Karjalainen, R.: Using genetic algorithms to find technical trading rules. *J. Financ. Econ.* **51**(2), 245–271 (1999). [https://doi.org/10.1016/S0304-405X\(98\)00052-X](https://doi.org/10.1016/S0304-405X(98)00052-X)
26. Mandziuk, J., Jaruszewicz, M.: Neuro-genetic system for stock index prediction. *J. Intell. Fuzzy Syst.* **22**(2–3), 93–123 (2011). <https://doi.org/10.3233/IFS-2011-0479>
27. Tsang, E., Yung, P., Li, J.: ‘EDDIE-automation’, a decision support tool for financial forecasting. *Decis. Support Syst. Periodical Style* **37**, 559–565 (2004). [https://doi.org/10.1016/S0167-9236\(03\)00087-3](https://doi.org/10.1016/S0167-9236(03)00087-3)
28. Tay, F.E.H., Cao, L.J.: Modified support vector machines in financial time series forecasting. *Neurocomputing* **48**(1–4), 559–565 (2002). [https://doi.org/10.1016/S0925-2312\(01\)00676-2](https://doi.org/10.1016/S0925-2312(01)00676-2)
29. Cao, L.J., Tay, F.: Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Trans. Neural Networks* **14**(6), 1506–1518 (2003). <https://doi.org/10.1109/TNN.2003.820556>

30. Lu, C.J., Lee, T.S., Chiu, C.C.: Financial time series forecasting using independent component analysis and support vector regression. *Decis. Support. Syst.* **47**(2), 115–125 (2009). <https://doi.org/10.1016/j.dss.2009.02.001>
31. Creamer, G., Freund, Y.: A boosting approach for automated trading. *J. Trading* **2**(3), 84–96 (2007)
32. Creamer, G.: Model calibration and automated trading agent for euro futures. *Quant. Financ.* **12**(4), 531–545 (2012). <https://doi.org/10.1080/14697688.2012.664921>
33. Boonjing, V., Boongasame, L.: Combinatorial portfolio selection with the ELECTRE III method: case study of the stock exchange of Thailand (SET). In: Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, vol. 8, issue 4, pp. 719–724 (2016). <https://doi.org/10.1504/AAJFA.2017.087506>
34. CEPAL: Bases de datos y publicaciones estadísticas. Recuperado el 08 de abril de 2022, de <https://statistics.cepal.org/portal/cepalstat/dashboard.html?theme=2&lang=es> (2022).
35. Circulo de estudios de Latinoamérica: Estadísticas económicas 2021. Recuperado el 24 de marzo de 2022, de <https://www.cesla.com/estadisticas-economia.php> (2022)
36. Corrente, S., Greco, S., Słowiński, R.: Multiple criteria hierarchy process in robust ordinal regression. *Decis. Support. Syst.* **53**(3), 660–674 (2012). <https://doi.org/10.1016/j.dss.2012.03.004>
37. Corrente, S., Figueira, J.R., Greco, S., Słowiński, R.: A robust ranking method extending ELECTRE III to hierarchy of interacting criteria, imprecise weights and stochastic analysis. *Omega (United Kingdom)* **73**, 1–17 (2017). <https://doi.org/10.1016/j.omega.2016.11.008>
38. Angilella, S., Catalfo, P., Corrente, S., Giarlotta, A., Greco, S., Rizzo, M.: Robust sustainable development assessment with composite indices aggregating interacting dimensions: the hierarchical-SMAA-Choquet integral approach. *Knowl.-Based Syst.* **158**, 136–153 (2018). <https://doi.org/10.1016/j.knosys.2018.05.041>
39. Giannoulis, C., Ishizaka, A.: A web-based decision support system with ELECTRE III for a personalised ranking of British universities. *Decis. Support. Syst.* **48**(3), 488–497 (2010). <https://doi.org/10.1016/j.dss.2009.06.008>
40. Alvarez, P., Muñoz-Palma, M., Miranda-Espinoza, L., Lopez-Parra, P., León-Castro, E.: Enfoque multicriterio jerárquico para el análisis de la competitividad de las regiones en México. *Inquietud Empres* **20**(2), 29–51 (2020). <https://doi.org/10.19053/01211048.11408>
41. Alvarez, P.A.; Valdez C.; Dutta, B.: Analysis of the innovation capacity of Mexican regions with the multiple criteria hierarchy process. *Socioeconomic Planning Sci.* **84**, (2022). <https://doi.org/10.1016/j.seps.2022.101418>
42. Bernal, M., Velázquez, D., Alvarez, P.A., Muñoz-Palma, M., León-Castro, E.: The financial portfolio selection using the multiple criteria hierarchical process and the Markowitz model. In: Sahni, M., Merigó, J.M., Hussain, W. (eds) *Novel Developments in Futuristic AI-based Technologies. Algorithms for Intelligent Systems*. Springer, Singapore (2023). https://doi.org/10.1007/978-981-99-3076-0_11
43. Muñoz-Palma M., Miranda, E.L., Alvarez, P. A., Bernal, M. León-Castro, E.: Stock selection using a multiple criteria hierarchical process in the Dow Jones index. *Int. J. Innovation Sustain. Dev.* **17**(1-2), 104–122 (2023). <https://doi.org/10.1504/IJISD.2023.127977>
44. Arenas, L., Muñoz Palma, M., Álvarez Carrillo, P. A., León-Castro, E., Gil Lafuente, M. A.: Multicriteria hierarchical approach to investment location choice. In Yuriy, P., Kondratenko, Y. P., Kreinovich, V., Pedrycz, W., Chikrii, A., Gil Lafuente, A. M. (eds.) *Artificial Intelligence in Control and Decision-making System*, Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25759-9_15
45. Corrente, S., Greco, S., Słowiński, R.: Multiple criteria hierarchy process for electre tri methods. *Eur. J. Oper. Res.* **252**(1), 191–203 (2016). <https://doi.org/10.1016/j.ejor.2015.12.053>

Chapter 4

Increasing Performance and Competitiveness in HEIs Applying an ICT Quality KPI Model Analyzed with AHP Method



David Lerma-Ledezma , Georgina Castillo-Valdez ,
Claudia Gómez-Santillán , and Manuel Paz-Robles 

Abstract This document identifies, classifies, and prioritizes the Information and Communication Technologies indicators focused on higher education derived from the consultation of one hundred ninety-seven bibliographical references with the aim of identifying those that directly impact the quality and educational competitiveness of these institutions. A survey was applied to Information and Communication Technologies managers of Higher Education Institutions to know the importance they give to the identified Information and Communication Technologies categories and the Key Performance Indicators that comprise them. After that, these categories and indicators are submitted to a hierarchical process based on standardized values to find the correct prioritization of them. It was found that the best evaluated category in the present research was the quality in Information and Communication Technologies; therefore, the Key Performance Indicators that integrate it were sent to the Analytic Hierarchy Process model where a minimum value of consistency coefficient = 0.047308 was found, which provides us with the following prioritization of quality indicators in Information and Communication Technologies: services, products, competitiveness, impact, electronic learning, blended learning and Personal Learning Environments. This study suggests that the application of the proposed

D. Lerma-Ledezma · G. Castillo-Valdez (✉)
Universidad Politécnica de Altamira, Altamira, Tamaulipas, México
e-mail: georgina.castillo@upalt.edu.mx

D. Lerma-Ledezma
e-mail: david.lerma@upalt.edu.mx

C. Gómez-Santillán · M. Paz-Robles
Tecnológico Nacional de México, Instituto Tecnológico de Ciudad Madero, Ciudad Madero,
Tamaulipas, México
e-mail: claudia.gs@cdmadero.tecnm.mx

M. Paz-Robles
e-mail: g97071322@cdmadero.tecnm.mx

model based on Key Performance Indicators and analyzed through Analytic Hierarchy Process in the technological infrastructure of the Higher Education Institutions will considerably increase both the competitiveness and the quality in higher education.

Keywords Quality in higher education · KPIs · Indicators prioritization · ICT · AHP

4.1 Introduction

Many Higher Education Institutions (HEIs) currently face major problems and challenges, one of them being educational quality. Since these institutions must compete with each other for the recruitment of students, who each day arrive at these institutions with greater knowledge and with great expectations regarding the educational quality that these institutions will be able to provide.

That is why it is important to measure, and above all, prioritize these problems within HEIs. The Key Performance Indicators (KPIs) are undoubtedly very important elements that orient one way or the other in terms of guidelines on where the efforts and decisions should be directed by the managers of these institutions.

Both Information and Communication Technologies (ICT) and ICT indexes become strong allies of HEIs to achieve high levels of quality and competitiveness in higher education; that is why this research related to these issues should be considered by these institutions when establishing their strategic plans.

It is important to start this research referring to what should be understood as a concept of quality in higher education according to different authors. In the same way, the concept of key indicators of performance should be understood in a general way to be able to focus on those KPIs whose purpose is to measure the quality of education in HEIs based on ICT.

Once these indicators have been selected, the relative importance of each one of them must be defined in order to establish a hierarchy through a prioritization mechanism within a model for assessing the quality of ICT in HEIs. Finally submit them to a decision process that make it possible to affirm that the proposed ordering is appropriate at the moment that it is required to implement this model in some HEIs.

Finally, a decision process is done to verify that the prioritization of the KPIs is adequate at the time of implementing the proposed model in the HEIs.

This study is organized as follows: Sect. 4.2 presents the background. Section 4.3 presents the proposed methodology. Section 4.4 shows the experimentation and results. Section 4.5, shows the conclusions.

4.2 Background

A general overall about the quality involved in the HEIs is discussed in this chapter. Topics related with this as KPIs applied to rank universities are mentioned in this chapter too, which is structured as follows: quality in HEIs, ranking KPI in HEIs and the importance of its measuring.

4.2.1 *Quality in Higher Education*

Quality in higher education is a topic that has been considered by universities that claim to be leaders in terms of the services they offer [1]. That is why this item should be analyzed in its different conceptualizations. This term was coined in the decade of the 80's and is widely determined depending on the governmental measures that are applied according to the current government.

There are several ways to define the concept of quality in higher education depending largely on the indicators used in its measurement, such as: Mysticism, reputation, resources, results and the added value that each institution wishes to highlight according to their visions and institutional missions.

One of the most successful notions of quality in higher education is the one proposed in 1993 by Harvey & Green in which five meanings are contemplated, which visualize quality in higher education as:

1. Exceptional condition (something unique to each university as well as constant improvement).
2. Perfection or consistency (the quality of reducing or eliminating defects).
3. Adaptation to a purpose (meets expectations).
4. Delivery of value for money (efficiency in the use of resources used).
5. Transformation (personal maturation process of an integral nature).

According to the authors, two research lines arise regarding the notions of the concept of quality in Higher Education. The first one focuses on finding empirical evidence that supports or refutes the hypothesis of the political content of this term. The second line is associated with the knowledge of the factors that condition the preferences of teachers for some or other notions of educational quality at higher levels. Likewise, they establish a third idea on this concept that emerges as the combination of the previous ones and that considers the opinions, attitudes and preferences of the teachers.

The studies of [2] showed that the most important elements for quality in higher education are:

- The consolidation of the academic groups.
- The institutional organization.
- The available resources.
- The interest of the students.

- The combination of all these elements.

Other collateral results of this research show that in the opinion of teachers, the main functions of the university are the following:

- The development of science
- The training of professionals
- Problem solving.

It is also important to consider the new perspectives for quality assurance in higher education [3] which is closely linked to the following four purposes:

1. Preparation for sustainable employment.
2. Preparation for life as citizens in democratic societies.
3. Personal development.
4. Development and maintenance, through the teaching, learning and research of a broad and advanced knowledge base.

At this point it is important to consider other views on the concept of quality in higher education and the spectrum of activities or functions that this concept can include. For example, the Council of Europe highlights the importance of comprehensively understanding educational quality, which includes both the quality of the system and the institutional quality. It also emphasizes the social dimension of quality: an education system cannot be of high quality unless it provides adequate opportunities for all students (Tables 4.1, 4.2, 4.3, 4.4, 4.5, 4.6).

An analysis of the literature suggests that crystallizing the definition of quality is difficult for two reasons [4]. These are: (1) quality is a relative concept; and (2) quality is used in various contexts. In this case, suggested three different paradigms of Quality Assurance, (QA) in education: “Internal”, “Interface” and “Future” quality. The “Internal QA” focused on improving the internal environment and processes, so the effectiveness of learning and teaching can be ensured to achieve the planned goals. The “Interface QA” is ensuring that education services satisfy the needs of stakeholders and are accountable to the public. The “Future QA” stresses ensuring the relevance of aims, content, practice and outcomes of education to the future of new generations. The concept of Quality Enhancement (QE) is similar to the QA

Table 4.1 Final prioritization of categories with standardized values

Category	1st Zi	2nd Zi	3rd Zi	Position
Quality in ICT	1.70	0.0017	− 0.71	1
ICT infrastructure	1.51	1.08	0.77	2
ICT scalability	0.95	0.58	− 1.08	3
ICT integration	0.62	0.43	− 0.41	4
ICT management	0.098	− 0.21	− 1.34	5
Strategic planning	− 1.08	− 1.17	− 1.72	6

Table 4.2 Application of AHP for the categories of KPIs model of ICT

ICT category	Model categories							1.24
	Quality ICT	ICT infrastructure	ICT scalability	ICT integration	ICT management	Strategic planning		
Quality in ICT	1	0.333333	0.166666667	0.111111	0.125	0.111111		
ICT infrastructure	3	1	0.5	0.5	0.333333	0.25		
ICT scalability	6	2	1	0.5	0.333333	0.5		
ICT integration	9	2	2	1	0.333333	0.5		
ICT management	8	3	3	3	1	0.5		
Strategic planning	9	4	2	2	2	1		
Totals	36	12.33333	8.666666667	7.111111	4.125	2.861111		
							Average	
Standardized	0.027778	0.027027	0.019230769	0.015625	0.030303	0.038835	0.0265	
	0.083333	0.081081	0.057692308	0.070313	0.080808	0.087379	0.0768	
	0.166667	0.162162	0.115384615	0.070313	0.080808	0.174757	0.1283	
	0.25	0.162162	0.230769231	0.140625	0.080808	0.174757	0.1732	
	0.222222	0.243243	0.346153846	0.421875	0.242424	0.174757	0.2751	
	0.25	0.324324	0.230769231	0.28125	0.484848	0.349515	0.3201	
Vector	0.162648					Consistency	6.1454	
Sum	0.478668					Vector	6.2353	
	0.779039						6.0697	
	1.07338						6.1978	
	1.781812						6.4767	
	2.018683						6.3061	
		$\lambda = 6.238496$		CI = 0.047699	CR = 0.0385	CR < 0.1		

Table 4.3 Comparison matrix of the categories of the proposed model

Category	Equally to moderately preferred over	Moderately preferred over	Moderately to strongly preferred over	Strongly to very strongly preferred over	Very strong to extremely preferred over	Extremely preferred over
Quality in ICT		ICT infrastructure		ICT scalability	ICT management	ICT integration Strategic planning
ICT infrastructure	ICT scalability ICT integration	ICT management	Strategic planning			
ICT scalability	ICT integration Strategic planning	ICT management				
ICT integration	Strategic planning	ICT management				
ICT management	Strategic planning					
Strategic planning	None	None	None	None	None	None

concept and the major differences between them applied to the education sector are shown on Table 4.7 (Annexes).

Finally, quality in higher education can be perceived as the sum of many elements, both, internal and external, in universities environments in order to match missions, visions and educational policies with the reality they offer to their students.

4.2.2 Key Performance Indicators (KPIs)

Every company or organization carries out multiple activities to achieve its objectives. Many of them are done sequentially until the general vision is achieved. Organizations are obliged to discover why sometimes their objectives are not met, either totally or partially.

Any activity or management that is carried out must be measured or evaluated with the objective of establishing whether the intended objectives were met. When situations are not favorable, it is necessary to check the activities in a separate or isolated way to identify specific ones in which the particular goals established were not met.

Table 4.4 Final prioritization of quality indicators in ICT

ID	KPI	Description	1st Zi	2nd Zi	Pos
QU-02	Services	ICT services of the HEI are continuously evaluated by users in terms of quality improvements	2.135	0.99	1
QU-03	Products	ICT products of the HEI are continuously evaluated by users in terms of quality improvements	0.907	0.519	2
QU-04	Competitiveness	The variables to be measured at this point are the institutional mission, resources and capabilities, design and implementation of the strategy, analysis of the external and competitive sector	0.907	0.519	3
QU-01	Impact	Effects or social processes that cause profound changes and transformations of a social and cultural nature, in addition to the economic one	0.519	0.1721	4
QU-07	Quantitative indicators	They measure the penetration and use of computers in schools; Studies on the effects of computers on the performance and learning of students and studies on the practices of computer use	0.519	– 0.809	5
QU-05	Perspectives and attitudes	They measure the degree of acceptance and what real possibilities exist to continue in the trend of the use of ICT within the HEI	0.519	– 0.809	6
QU-06	Technological study	Situational analysis on technological environments that may influence the use of ICT within the HEI	0.519	– 1.056	7

(continued)

Table 4.4 (continued)

ID	KPI	Description	1st Zi	2nd Zi	Pos
QU-08	e-learning	It is based on vitality and uses processes to transmit, produce, exchange information and knowledge by electronic means in the teaching–learning processes supported by ICT of the HEI	0.1721	– 2.31	8
QU-09	b-Learning	Methodologies are mixed to achieve better learning results. Integration of elements common to face-to-face teaching with elements of distance education over the Internet	– 0.424	– 0.809	9
QU-10	PLE	Students learn informally and personally at their own place	– 0.809	– 1.369	10

KPIs are instruments that provide quantitative information on the development and achievements of an institution, program, activity or projects in favor of the population or object of its intervention, within the framework of its strategic objectives and mission.

The information obtained from the key performance indicators will allow institutions to compare themselves or with other institutions, to promote processes of educational or academic innovation, distribution of resources according to priorities, accountability to the communities, public institutions, agencies, government and society in general.

In its general form, a KPI can be defined “When you can measure what you are talking about and measure it in numbers, you already know something about it, when you cannot express it in numbers, your knowledge is deficient and unsatisfactory; it can be the beginning of knowledge and you have gone forward in your thoughts to the stage of science” [5].

These authors also differentiate KPIs from other similar concepts that exist and that is why some organizations use them incorrectly. These concepts are listed below:

1. Key Result Indicators (KRI): How critical success factor has been achieved.
2. Results Indicators (RI): they tell you what you have done.
3. Performance Indicators (PI): they tell you what you should do.
4. Key Performance Indicators (KPIs): what to do to increase performance.

In a chronological order of events or processes within an organization, KPIs can be classified, according with [6], on three types of KPIs:

Table 4.5 Model AHP to quality indicators in ICT

KPI	Quality Indicators In ICT										1.41
	QU-02	QU-03	QU-04	QU-01	QU-08	QU-09	QU-10	QU-07			
QU-02	1	0.5	0.5	0.5	0.25	0.5	0.111111	0.111111	0.111111	0.111111	
QU-03	2	1	0.5	0.333333	0.2	0.16667	0.111111	0.125	0.125	0.125	
QU-04	2	2	1	0.333333	0.2	0.5	0.166667	0.125	0.125	0.125	
QU-01	2	3	3	1	0.5	0.33333	0.25	0.25	0.25	0.25	
QU-08	4	5	5	2	1	0.5	0.333333	0.333333	0.333333	0.333333	
QU-09	2	6	2	3	2	1	0.5	0.5	0.5	0.5	
QU-10	9	9	6	4	3	2	1	0.5	0.5	0.5	
QU-07	9	8	8	4	3	2	2	1	1	1	
Totals	31	34.5	26	15.16667	10.15	7	4.472222	2.94444	2.94444	Average	
Standardized	0.032258	0.01449	0.01923077	0.032967	0.024631	0.07143	0.024845	0.03774	0.03774	0.0322	
	0.064516	0.02899	0.01923077	0.021978	0.019704	0.02381	0.024845	0.04245	0.04245	0.03069	
	0.064516	0.05797	0.03846154	0.021978	0.019704	0.07143	0.037267	0.04245	0.04245	0.04422	
	0.064516	0.08696	0.11538462	0.065934	0.049261	0.04762	0.055901	0.08491	0.08491	0.07131	
	0.129032	0.14493	0.19230769	0.131868	0.098522	0.07143	0.074534	0.11321	0.11321	0.11948	
	0.064516	0.17391	0.07692308	0.197802	0.197044	0.14286	0.111801	0.16981	0.16981	0.14183	
	0.290323	0.26087	0.23076923	0.263736	0.295567	0.28571	0.223602	0.16981	0.16981	0.25255	
	0.290323	0.23188	0.30769231	0.263736	0.295567	0.28571	0.447205	0.33962	0.33962	0.30772	
Vector	0.268348		Consistency	8.334168							
Sum	0.255029		Vector	8.309769							
	0.369139			8.347313							
	0.607529			8.519578							

(continued)

Table 4.5 (continued)

KPI	Quality Indicators In ICT										1.41	
	QU-02	QU-03	QU-04	QU-01	QU-08	QU-09	QU-10	QU-07				
	1.023128			8.563281								
	1.211837			8.544076								
	2.165083			8.572923								
	2.629246			8.544337								
$\lambda =$	8.466931		CI =	0.066704		CR =	0.047308					CR < 0.1

Table 4.6 Comparison matrix of the KPIs quality in ICT

KPI	Equally to moderately preferred over	Moderately preferred over	Moderately to strongly preferred over	Strongly preferred over	Strongly to very strongly preferred over	Very strongly to extremely preferred over	Extremely preferred over
QU-02	QU-03 QU-04 QU-01 QU-09		QU-08				QU-10 QU-07
QU-03	QU-04	QU-01		QU-08	QU-09	QU-07	QU-10
QU-04	QU-09	QU-01		QU-08	QU-10	QU-07	
QU-01	QU-08	QU-09	QU-10 QU-07				
QU-08	QU-09	QU-10 QU-07					
QU-09	QU-10 QU-07						
QU-10	QU-07						
QU-07	QU-06 QU-05						

1. Leading indicator. A KPI that measures activities that have a significant effect on future performance.
2. Lagging indicator. A KPI that measures the production of past activities.
3. Diagnostic measure. A KPI that indicates the state of the processes or activities.

In the same research, it was found the following KPIs features:

1. Scarce, that is, the less KPIs exist, the better.
2. Drillable, users can delve into details.
3. Simple, users understand the KPI. Clearly indicates the required action.
4. Action, users know how they affect the results.
5. Ownership, KPIs have an owner, can be served by the CEO.
6. Referenced, users can see origins and context.
7. Correlated, the KPIs drive the desired results.
8. Balanced, KPIs consist of financial and non-financial indicators.
9. Aligned, KPIs do not get in the way of each other.
10. Validated, workers cannot avoid KPIs.
11. Regulated, they are measured frequently (24/7, daily, weekly).
12. Distributed, they are measures that match responsibilities to a team.

4.2.2.1 Most Common KPI Used in Ranking Universities

The world ranking universities began in 2003, and from that time several international methodologies such as U-Multiranking, ARWU (Shanghai Ranking), Times HE, Leiden Ranking, QS (Quacquarelli Symonds), Webometrics (CSIC), PERSPEK-TYWY (based on United States system ranking), HEQAM (Saudi Arabia) have recently been established according to geographical areas, among others, and whose purpose is to improve current ranking systems [7].

The following list shows the Key Performance Indicators that matches in the metrics of the different methodologies described above:

- Quality of Education
- Teaching and Learning
- Quality of Faculties
- Research
- Knowledge Transfer
- International orientation or Internationalization
- Regional engagement
- Productivity.

4.2.2.2 Quality Indicators for Higher Education

There are several ways to measure educational quality in HEIs since many of them are linked to different evaluating bodies, both governmental and private. In other cases, some of these institutions are concerned with highlighting some aspects of others, which vary according to their missions and institutional visions.

An example of this can be seen with [8] where 12 KPIs that measure the quality in higher education are recognized: Ratio of academic staff to disciplinary, ratio of students to academic staff, ratio of academic staff satisfaction of education level, ratio of students' staff satisfaction of education level, number of classrooms to the number of students, number of laps found to factor required for each department, ratio of students satisfaction of teaching aids, number of graduate programs that need to be learned after graduation, number of books in library, time cycle for up-to-dating the library, ratio of students satisfaction of library service, time cycle for up-to-dating the computer and IT equipment's (teaching aids) of the faculty.

Currently, HEIs that wish to join the select group of institutions that offer quality education should have within their infrastructures both physical and logical mechanisms to measure the degree of implementation of their own quality indicators. Do not forget that each HEI can generate or select their own indicators according to their interests to measure their educational quality (self-perception) and also to compare themselves with other institutions (competitiveness). There is specialized software in the market to measure these KPIs within institutions such as the KPI-MS [9].

As a sample, some aspects that can and should be considered when creating quality indicators or systems of quality indicators to increase competitiveness in higher education institutions are presented in the managerialism. The managerialism,

mentioned by [10], in a context of higher education, has the following characteristics that can each be considered as quality indicators.

- A greater separation of academic work and management activity.
- Increased control and regulation of academic work by managers.
- A perceived shift in authority from academics to managers and consequent weakening of the professional status of academics.
- An ethos of enterprise and emphasis on income generation.
- Government policy focused on universities meeting socio-economic needs.
- More market orientation, with increased competition for resources.

Davis et al. [11] state managerialism in universities supports evidence of academics measurement to those who receive educational services including performance management, teaching, quality of research, inspection, performance indicators and goal setting.

On the other hand, indicators of the academy can be considered as indicators that also impact on the educational quality of HEIs, as mentioned by [12] when affirming that an overview of the current mission statements of the highest ranked universities reveals that both knowledge production (research) and knowledge dissemination (teaching/education/learning) are strongly embedded as these universities priorities.

Another indicator of quality in the HEIs is the Entrepreneurship Education (EE) that according to the conclusions of [13] it (EE) can measure the accumulation of positive learning experiences, the development of practical business skills and the combined knowledge of theory and practical exercises.

Finally, there are the KPIs that measure the quality of higher education according to the opinions of the students, which in some countries can be perceived as “clients” according to [14]. This consumer identity appears to be increasingly recognised by students, who demand more from the higher education sector than ever before. But, while a rich tradition of research has investigated how we can predict academic performance there remains a paucity of research on the extent to which today’s students express a consumer orientation and how this may affect academic performance. The traditional factors predicting academic performance, namely learner identity and grade goal, and the interplay with consumer orientation and gives evidence that consumer orientation mediates or influences traditional predictors of academic performance: the more the students expressed a consumer orientation, the poorer their academic performance.

4.2.2.3 Quality KPI Prioritization in HEIs

Currently many companies or organizations involved in projects with quality impact resort to specialized training tools on statistical methods and quality to improve processes that enable them to function as leaders, facilitators and problem solvers according to a correct prioritization of indicators [15].

The process of selecting KPIs to measure educational quality in HEIs is extremely complex due to the inherent subjectivity of the process itself. This subjectivity

is nurtured by the fact that many of the institutions compete with each other for the recruitment of enrolment, offering their educational services according to the guidelines of their missions and institutional visions.

Subjectivity must be replaced by objectivity at the moment of establishing the correct and priority order of the KPIs that must be considered when including them in a model. It is important to consider that for the selection and prioritization of the indicators, the staff that manages the HEIs must be directly involved as well as all the areas and/or dependencies that the institution has [16] applying questionnaires with likert scales which contain items such as “Not important”, “Important”, “Very important”.

According to this structure, a three-level KPI tree can be built, where the first level will be the objective (that is, the total percentage of the institution’s performance), the second level (the criteria, such as teaching, research and support) and a third level (the rating scale, which contains the KPIs related to each criterion).

Once an order of importance of the KPIs has been established, a mechanism must be found to verify that such ordering is suitable and an appropriate technique for this is the decision analysis model AHP TOPSIS [8] which, in general, feeds on parameters or numbers, which are processed mathematically and algorithmically (supported by a spreadsheet such as Excel) to provide objective and concrete results regarding the correct prioritization of said indicators.

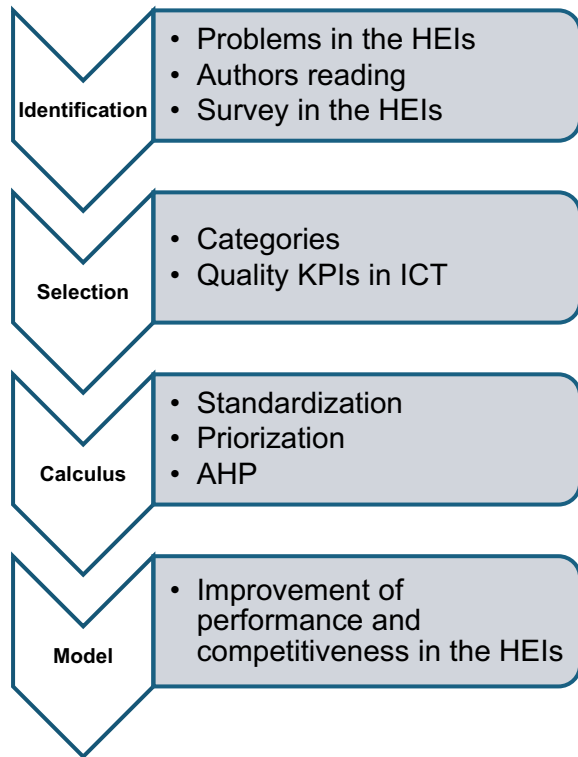
In line with the above and to establish the appropriate measures of the KPIs, the following steps suggested by [17] might be followed:

1. A review of the institution’s situation in accordance with the mission and directions of the country’s education strategy (Political guideline).
2. Analysis of strengths and weaknesses.
3. Establish a synthesis between the requirements of the environment and the competitive profile.
4. Define the vision of the institution.
5. Determine the objectives.
6. Determine the business objective as a variation of the strategic objectives in operational objectives.
7. Identify the indicators that allow monitoring the achievement of the objectives.
8. Definition of acceptable thresholds of the indicators and their validation.
9. Operate the indicators.

4.3 Proposed Methodology

The proposed methodology used in this research process was developed in four stages, which are (1) Identification of the problems, (2) Identification of ICT KPI categories in HEIs and quality ICT KPIs, (3) Calculations of categories in HEIs and quality ICT KPIs, (4) Model implementation. Figure 4.1 summarizes the proposed methodology in this work.

Fig. 4.1 Stages to generate the proposed model



Stage 1: Identification of the Problems

The problems identified in the HEIs are the lack of educational quality based on ICT, the incongruence between institutional policies and ICT policies, the lack of tele-education infrastructure, the lack of algorithmic or procedural prioritization for the distribution of ICT, and the lack of ICT index measurement systems.

That is why we proceeded to search for KPIs that measure ICT within these institutions by consulting 197 references and reaching the conclusion that the indicators found can be classified into six major categories of ICT indexes for HEIs.

A survey was also applied to ICT managers of HEIs to know the prioritization they give to these categories and to the KPIs that conform them, finding that the category called Quality in ICT turned out to be the best positioned reason for which it was selected to perform a more in-depth study of the indicators that constitute it.

Stage 2: Identification of ICT KPI Categories in HEIs and Quality ICT KPIs

According to the 197 consulted references during the present investigation, six categories related to ICT were identified, which compose the ICT quality assessment model based on KPIs. These categories are shown with a brief description of each of them.

1. ICT Management (MA) Related to projects, control and ICT maintenance.

2. ICT Infrastructure (IF) Related to information systems, telecommunications and equipment
3. ICT Integration (IT) Related to integrated services and the way to access them.
4. ICT Scalability (SC) Related to measures, information speed, growth, and ICT innovation.
5. Quality in ICT (QU) Related to impact, services, products, and competitiveness.
6. Strategic Planning (SP) Related to teaching innovation, suppliers, customers, and position.

In quality in ICT category were found the indicators: services, products, competitiveness, impact, quantitative indicators, perspective and attitudes, technological study, e-Learning, b-Learning and PLE. Tables 4.8, 4.9, 4.10, 4.11, 4.12 and 4.13 (Annexes) show these categories and its respectively indicators.

Stage 3: Calculations of Categories in HEIs and Quality ICT KPIs

After selecting this category, it is necessary to submit their indicators to a hierarchical process based on standardized values in order to find the prioritization of these KPIs, which is submitted to the AHP model to check its prioritization to finally establish a model of ICT quality assessment based on KPIs that is analyzed by a process decision model to increase performance and competitiveness in HEIs.

Stage 4: Model Implementation

Finally, once the results are obtained in Excel, it is necessary to create a final model based on the preferences established by the AHP on the ICT quality KPIs in table form.

4.4 Experimentation and Results

This section shows the calculations that were carried out to obtain the results of the study and is structured as follows: The configuration of the equipment used is shown, the proposed methodology is applied, the results obtained between Excel and the AHPy Python library are compared and finally, concludes with some success cases.

4.4.1 Settings

The software used for the experimentation was Excel 2016 and Python version 3.10.12. The tests were run on a DELL Inspiron 15" 5559 Laptop, INTEL Core i5, 8 Ghz, HD 1TB SATA. The indicator to be measured is the CR consistency coefficient and the instances were taken from the random numbers generated by Excel as a comparison of preferences.

4.4.2 Application of the Proposed Methodology

Weighting of the Categories

Table 4.14 (Annexes) shows the concentration of authors who have pronounced in favor of each of the categories of the model and at the end of it shows the total accumulated (197) of the citations made by authors on the indicators that integrate each one of these categories and that will serve as a basis for subsequent calculations.

Table 4.15 (Annexes) shows the analysis of the ICT infrastructure elements of the HEIs surveyed based on interviews with their ICT managers and the following results were obtained.

From Table 4.15, Table 4.16 is generated showing two different orders according to the HEIs, one is according to the results generated by the measuring instrument and another one is according to the analysis of the KPIs of each category.

Prioritization of Categories

Tables 4.14 and 4.16 (Annexes) prioritize the six categories of the model from two points of view (one of authors and the other of the HEIs which is subdivided in one according to the measurement instrument and another with respect to the KPIs); therefore, it is necessary to establish a final prioritization of them using the method of data standardization [18].

For the standardization of the categories ordered according to the criteria of the authors, Table 4.17 (Annexes) is considered. The average (μ) in this case is $\mu = \frac{\sum_i^n X_i}{n} = 16.66$ and the standard deviation (σ) is calculated as $\sigma = \sqrt{\frac{\sum_i^n (X_i - \mu)^2}{n}} = 6.02$. The last column shows the standardized data (Z_i) which are calculated as follows: $Z_i = \frac{X_i - \mu}{\sigma}$ and once obtained these will be used to obtain the final prioritization. In Table 4.17, the categories are also standardized according to the measurement instrument ($\mu = 77.11$, $\sigma = 5.13$ and $n = 6$) and to the indicators ($\mu = 76.52$, $\sigma = 5.82$ and $n = 6$).

Once the standardization table is obtained, a final table is constructed based on the three columns of standardized values (Z_i) which contains the final prioritization of the categories of the proposed model Table 4.1.

As can be seen, the category of Quality in ICT was the best positioned (the most important) according to the authors, the measurement instrument and the KPIs specified by the HEIs. The indicators of this category will be submitted in the same way to the processes of standardization, ranking and application of the AHP model in the following sections.

Application of the AHP Model to the Categories

The AHP (Analytic Hierarchy Process) model developed by Saaty is a popular approach to rank alternatives. Through the ratio-scaled assessment of pairwise preferences between alternatives, the ranks of alternatives are found by computing the eigenvalues of the preference matrix and its very useful visual tool for detecting the cardinal and ordinal inconsistencies [19].

The AHP decision analysis model is the one that will be used because it allows you to enter values in an Excel file with pre-recorded formulas and whose calculations can

be performed iteratively until you get the smallest possible coefficient of consistency (*CR*) value. A *CR* value of less than 10% (0.1) is considered acceptable, trying to obtain even smaller values to ensure the best possible combination.

This method works by comparison of pair values which are assigned in the range of 2 and 9 (value 1 is allowed, but only used when comparing an element against itself) to indicate the degree of preference of an element over the other. Based on these combinations of numbers the corresponding calculations are made. Once the lowest possible value of *CR* is obtained, the correct decision making is done. The final prioritization of categories Table 4.1 is submitted to the AHP model, which consists of pre-recorded formulas in Excel. After performing 50 tests with different values, the combination shown in Table 4.2 is obtained. It shows the lowest value of the coefficient of consistency ($CR = 0.0385 < 0.1$).

Once the AHP model is applied, the comparison matrix of the categories in Table 4.3 is generated. This table shows the preference degree of one category over the others.

Up to this moment, the design of a model of six categories of KPIs has been established. The category called “Quality in ICT” was selected because it was the best positioned and it is also related in a natural way to the concept of educational quality in higher education. This category consists of 10 indicators which will be submitted to the calculation, standardization, prioritization, and application processes of the AHP model.

Prioritization and Application of the AHP to the KPIs of the ICT Quality Category

The calculations for the prioritization of the key performance indicators that integrate the ICT Quality category are shown in Table 4.18 (Annexes) and whose Pareto’s graph is shown in Fig. 4.2.

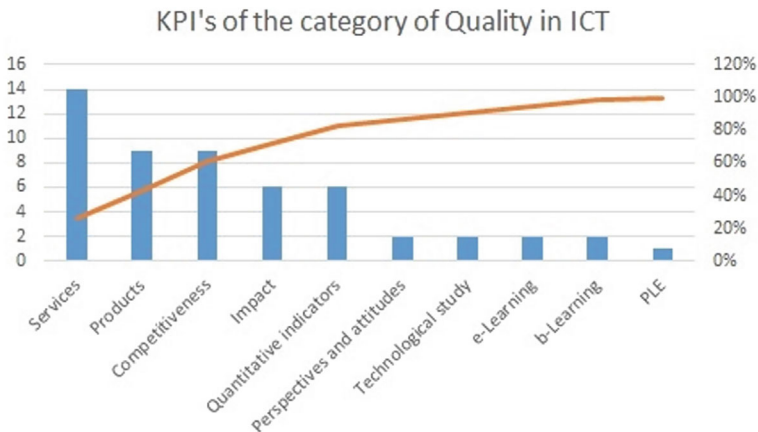


Fig. 4.2 Pareto’s graph of the KPIs of the category of Quality in ICT

Based on Table 4.15, the average of each of the 10 KPIs that integrate the ICT Quality category is calculated and ordered in descending order according to the average as shown in Table 4.19 (Annexes).

According to authors' criteria, standardized values of ICT Quality category are shown in Table 4.20 (Annexes).

Table 4.20 (Annexes) is ordered according to the standardized values (Z_i) and Table 4.4 is generated, which shows the final prioritization of the KPIs of this category and also shows the ID and a description of each KPI.

Finally, the AHP model is applied to the values of Table 4.4 and after 52 iterations the value of $CR = 4.73\%$ shown in Table 4.5 was found.

From Tables 4.5 and 4.6 is generated, showing the comparison matrix of the key performance indicators that integrate the ICT Quality category.

As part of the process of checking the research and based on a sampling focused on the area and nature of the research, three different types of analysis were applied within the HEI. The first of these was a survey applied to ICT managers of the HEI of the study, which revealed that 77.33% of these institutions take advantage of their ICT resources as a driving force of educational quality.

A second analysis based on the analysis of the measurement instrument showed that 77.11% of HEIs apply their ICT in order to increase their educational quality. The third analysis applied to the KPIs indicated that 76.52% of these institutions take advantage of their technological infrastructures to increase their educational quality.

4.4.3 Comparison Between the Results of Excel and AHPy Python Library

AHPy [20] is an implementation of AHP, a method used to structure, synthesize and evaluate the elements of a decision problem. Developed by Thomas Saaty in the 1970s, AHP's broad use in fields well beyond that of operational research is a testament to its simple yet powerful combination of psychology and mathematics.

AHPy attempts to provide a library that is not only simple to use, but also capable of intuitively working within the numerous conceptual frameworks to which the AHP can be applied. For this reason, general terms have been preferred to more specific ones within the programming interface.

A comparison in similar conditions was made between Excel software and the AHPy Python library. Using Excel, it was found a $CR = 0.047308$. This value is better than the value of 0.0862 obtained with Python after 50 executions.

In the order hand, when the number of iterations was increased to 1000, the CR gotten by Python (0.0387) was better than Excel CR value. This value means a

performance increment of 34% and at the same time it is necessary to mention that the execution time to creation tables in Python is significantly upper than Excel (0.6702 s).

Algorithm 1 is shown below, and Table 4.21 (Annexes) contains the results supplied by Python AHPy library after 1000 iterations.

Algorithm 1. Creating comparison tables and calculating consistency coefficients.

Input

$N = 8$.

$consistency_coefficients_vector = []$.

$values_vector = [2-5, 9, 10]$

$tags_factors = \{ tags_factors = \{0:'QU-02', 1:'QU-03', 2:'QU-04', 3:'QU-01', 4:'QU-08', 5:'QU-09', 6:'QU-10', 7:'QU07'\} \}$

Output

Min_value_CR

1. Start.
 2. For *cycle* from 1 until 50 Do:
 3. $data = [N \times N]$.
 4. For *i* from 1 until *N* Do:
 5. For *j* from 1 until *N* Do:
 6. If $i = j$ then:
 7. $data[i, j] = 1$.
 8. If not then:
 9. $value = random(2, values_vector[i])$.
 10. $datos[i, j] = value$.
 11. For *k* from 1 until *N* Do:
 12. For *m* from *k* until *N* Do:
 13. If *k* is different from *m* then:
 14. $data[k, m] = 1 / data[m, k]$.
 15. For *n* from 1 until number of items in *data* Do:
 - 16.. For *p* from 1 to number of items in *data* Do:
 - 17.. $comparisons[(tags_factors [p], tags_factors [n])] = data[n][p]$.
 18. ICTs = `ahpy.Compare(name = 'ICTs', comparisons = ict_comparisons, precision = 4, random_index = 'saaty')`.
 19. $consistency_coefficient = ICTs.consistency_ratio$.
 20. $consistency_coefficients_vector \leftarrow consistency_coefficient$.
 21. $min_value_CR = minimum(consistency_coefficients_vector)$.
 22. Print min_value_CR .
 23. End
-

The code begins by creating a square matrix named 'data' of size $N \times N$ filled with zeros, setting the groundwork for subsequent operations. It then populates the

main diagonal of this matrix with ones and randomly assigns values between 2 and 9 below the diagonal, simulating preferences or relative importance among different factors. Inverses are then calculated and assigned to elements above the main diagonal to ensure matrix symmetry. Subsequently, a dictionary called ‘comparisons’ is created to store pairwise comparisons between factors based on the values in the ‘data’ matrix. The code utilizes the ‘ahpy’ library to calculate the consistency coefficient for each pairwise comparison stored in the ‘comparisons’ dictionary, storing these coefficients in a vector named ‘consistency_coefficients_vector’. Finally, the minimum value of the ‘consistency_coefficients_vector’ is calculated to determine the minimum consistency coefficient obtained during the iteration, which is then printed as the output. This process provides insight into the relative importance of factors and ensures the consistency of pairwise comparisons.

Next each line is explained.

Line 1: A square matrix of 0’s size $N \times N$ is created.

Lines 5 to 10: Assign 1’s to the main diagonal and random values between 2 and 9 below the main diagonal.

Lines 11–14: Inverses are assigned to elements above the main diagonal.

Lines 15–17: Create a dictionary of pairwise comparisons with the preference value taken in the data matrix.

Lines 18–20: Python’s ahpy library is used to calculate the consistency coefficient for each data array and is assigned to a vector *consistency_coefficients_vector*.

Line 21: Take the minimum value from the vector *consistency_coefficients_vector* and assign it to *min_value_CR*.

Line 22: The output displays the minimum value for the coefficient of consistency.

4.4.4 Application of the AHP Method in Combination with Others: Success Cases

The following four cases are presented as cases of practicality, feasibility and validation of the application of AHP model and other mathematical decision analysis models, both to higher education institutions and to other areas:

An example where the models of analysis of decisions applied to HEI is proposed by [21] using classification trees. In their study four intangible resources of Higher Education Institutions with its respectively KPIs were identified: Knowledge management (8 KPIs), brand (5 KPIs), institutional reputation (8 KPIs) and corporate social responsibility (6 KPI), so which the proposed model consists of 27 KPIs.

As a result of the application of classification trees, institutional reputation KPI is prioritized first. Secondly, according to said tree, the training and development of human resources KPI is identified (KPI belonging to the intangible resource of

knowledge management), which strengthens the knowledge that is created, stored and transferred from functions and formation and training programs in different cycles to achieve the analytic thinking promoted in students, professors and administrative staff. Thirdly, is found the KPI corresponding for innovation and learning as a priority function of the HEI.

Likewise, [22] apply the AHP method in Generic Competences (GC) according to the type of sciences (agricultural, natural, and exact, social and administrative and engineering and technology) considered by the National Association of Universities and Institutions of Higher Education (ANUIES).

For this study 30 GC were considered, which are grouped into the four types of sciences specified above.

The following order of priority was obtained once the AHP method was applied, which coincided with the weighted products method, prioritizing the GC as follows: Ability to apply knowledge in practice, knowledge about the area of study and profession, commitment ethics, identification, planning and solving problems, capacity to make decisions, commitment to quality, ability to work in teams and skills in the use of CIT.

The advantage of AHP over the weighted product technique is the speed and simplicity with which the relevance level of GC can be obtained for different subject areas.

Below are described two cases of application of the AHP decision analysis method in other areas where significant advantages have been demonstrated in the application of this mathematical model of decision analysis in combination with other models, demonstrating once again its practicality and viability.

A case of application of the AHP is the one proposed by [23] in an automotive company. These authors combine this method with the TOPSIS method to generate comparison pair matrices in which KPIs are involved (10 in this case, with 19 decision alternatives) and where the usefulness of the combination of both methods is observed. The objective of this combination of methods is the minimization of operational waste. In this study, three comparative tables are offered (two from other authors) and the one proposed by [24] where the improvements were demonstrated in terms of the proximity of the distances towards the optimal solution in the ordering and prioritization of the 19 alternatives.

In the three comparative tables it is concluded that the JIT/continuous flow production is the most important. The alternatives of 5S and focused factory production also had relevant positions in the final ordering ranking.

On the other hand, [23] present us with an example of the application of these decision analysis models by comparing the results of applying the IFDA (Intuitionistic Fuzzy Dimensional Analysis) and AIFDA (Aggregated Intuitionistic Fuzzy Dimensional Analysis) and comparing them with the results obtained by applying the MCDM AHP and TOPSIS multi-criteria methods when selecting a milling machine from among three alternatives (W, X y Z) and considering six criteria (original cost of the machine, power, number of cylinders, displacement, safety of operators and service).

Regarding the milling machine alternatives, all the studies agreed that the best alternative was the Z milling machine. Likewise, all the studies prioritize the safety of operators as the most important criterion. Service was also identified as an important criterion to consider.

4.5 Conclusions

According to categories considered in the model (Table 4.3) it is concluded that HEIs should pay special attention to the quality of ICT specialized in teaching and knowledge generation. For this purpose, it is necessary that these institutions have specialized teaching staff for the evaluation of these aspects, especially with research trends, which are distinctive features of this type of institutions that aim to raise their level in the quality of higher education.

Related to the categories of infrastructure and ICT scalability, it is concluded that both are of similar relevance and intimately linked. Both categories evaluate hardware, software, technologies and platforms that directly support the teaching-learning process.

It is also perceived that the categories of ICT management and strategic planning should be aligned with the vision, mission and institutional policies with the firm purpose of raising the quality standards in the education of HEIs.

Likewise, it was detected in this table that the six categories considered in the study maintain at least one level of difference among them in terms of preference levels, which means that there is an adequate relationship among them.

As it has been demonstrated, the best evaluated category was quality in ICT, for which the following conclusions were found about its indicators. Based on Table 4.18 (Annexes), this study concludes that the indicators of services, products, competitiveness, impact and the quantitative indicators together represent the 83.01% of the total indicators of this category. So, the consideration of these indicators assures the quality in ICT improving the quality in education in the HEIs.

The ICT services' indicator represents 26.41% according to the authors consulted. The products' indicator (hardware, software and/or educational platforms such as Moodle, among others) and competitiveness's indicator (review of institutional missions, resources and capacities, as well as the implementation of strategies), represent 16.98% each one.

The impact indicator (social and marketing effects that promote the HEIs based on ICT) and the quantitative indicator (degree of use of computers and technological resources available to the institution, among others) represent 11.32% each one.

It is widely expected that the incorporation of this model in HEIs will significantly increase the percentage of ICT use in educational processes, hence this will directly impact on quality assurance in higher education.

4.6 Future Research

As future research, it is proposed to verify the concepts of competency and competence proposed by [25]. According to the author, the main difference is that competency is related to skills, aptitudes, abilities, knowledge and understanding while competence is visualized as an output function, that is, the performance standards of a job.

It is important to mention here that future work will consist of identifying the different KPIs of each of these concepts and incorporating them into a new model of indicators and prioritizing them through a mathematical model of decisions with the aim of improving the competitiveness of HEIs, since this model increases the educational quality in them.

Acknowledgements The authors want to thank to Universidad Politécnica de Altamira the facilities granted to develop the present work and to Laboratorio Nacional de Tecnologías de la Información the support of the TecNM project 21336.24-P and the scholarship for postgraduate studies with CVU 1260161.

Declaration of Conflicting Interests The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Annexes

See Tables [4.7](#), [4.8](#), [4.9](#), [4.10](#), [4.11](#), [4.12](#), [4.13](#), [4.14](#), [4.15](#), [4.16](#), [4.17](#), [4.18](#), [4.19](#), [4.20](#) and [4.21](#).

Table 4.7 Differences between QA and QE concepts applied to education

Quality assurance (QA)	Quality enhancement (QE)
Gives insufficient weight to the teaching/ learning processes	Gives considerable weight to the teaching/ learning processes
Tends to be associated more with assessment and accountability	Tends to be associated more with improvement and development
Meets external standards	Meets internal standards
Moves from top to lower level	Moves from lower to top level
A summative process	A formative process
A quantitative performance	A qualitative performance
Focuses on the past	Focuses on the present and the future
Less freedom (follows absolute rules)	More freedom (flexible ways)
Gives greater space to administrators	Gives a greater space to academics

Source [4]

Table 4.8 Frequency of each KPI of ICT Management category by authors

Authors	ICT management									
	Projects	Control	Maintenance	Policies for use and loan	Priorization of resources	Administrative mechanisms and Decision making	Laws and regulations	Institutional management	Organizational structure	
Área [26]				X						
OECD [27]	X		X							
Bermúdez et al. [28]				X						
Cobo [29]				X						
Tapia [30]						X				
Galindo et al. [31]				X	,					
OECD [27]	X			X	X					
Romero et al. [32]	X	X	X	X						
Soto et al. [33]	X	X	X	X						
Rodríguez et al. [34]		X			X					
Nivien [35]			X	X		X	X		X	
Cano et al. [36]						X				
Zapata [37]	X									

(continued)

Table 4.9 Frequency of each KPI of ICT Infrastructure category by authors

Authors	ICT infrastructure									
	Information system	Enterprise system	Tools	Telematics	Telecommunications	Basic components	Assets or strategic resources	Investment	User numbers	
Hernández [41]							X		X	
Lawler et al. [42]		X								
Área [26]							X			
OECD [27]					X	X				
OECD [27]					X				X	
Bermúdez et al. [28]							X			
Tello [43]					X					
Cobo [29]			X		X					
Tapia [30]							X			
Díaz [44]			X				X			
OECD [27]							X	X	X	
Romero et al. [32]							X			
Montaño [45]				X			X			
Soto et al. [33]							X			
Gámiz [46]							X			

(continued)

Table 4.10 Frequency of each KPI of ICT integration category by authors

Authors	ICT integration				
	Digital technologies	Convergence networks	Integrated services	Access	Using practices
Área [26]					X
OECD [27]		X	X	X	
Cobo [29]		X	X	X	
Galindo et al. [31]		X			
Ávila [51]				X	
Romero et al. [32]		X	X		
Montaño [45]		X		X	
OECD [27]			X		
Rodríguez et al. [34]					X
Chávez [47]		X			
Cano et al. [36]			X		
Aleksejeva [52]				X	
Zapata [37]			X		
Olufemi [48]	X	X			
Humanante et al. [38]			X	X	
Nolasco et al. [39]			X	X	
Capanegra et al. [40]			X		
Evren [49]			X		
Grande [50]				X	
% Factor	1	7	10	8	2
Average	0.0050	0.0355	0.0507	0.0406	0.0101
Category	28				
% Category	28/197 =	0.1421			

Table 4.11 Frequency of each KPI of ICT Scalability category by authors

Authors	ICT scalability			
	Measures	Speed of information	Innovation	Users number
Hernández [41]				X
Lawler, et al. [42]	X			
Área [26]	X			
Olivé [53]			X	
OECD [27]				X
Bermúdez et al. [28]			X	
Cobo [29]			X	
Pons [54]		X		
Galindo et al. [31]			X	
Ávila [51]	X			
OECD [27]	X			X
OECD [27]			X	
UNAM [55]			X	
Aleksejeva [52]			X	
Sampedro [56]		X		
Nolasco et al. [39]			X	
Grande [50]		X	X	X
% Factor	4	3	9	4
Average	0.0203	0.0152	0.0456	0.0203
Category	20			
% Category	20/197 =	0.1015		

Table 4.12 Frequency of each KPI of Quality in ICT category by authors

Authors	Impact	Services	Products	Competitiveness	Perspectives and attitudes	Technological, political and socioeconomic study	Quantitative indicators	E-learning	B-learning	PLE
OECD [27]		X								
Área [26]					X		X			
OECD [27]			X							
Olivé [53]	X	X	X							
OECD [27]		X								
Tello [43]				X						
Cobo [29]	X	X								
Zambrano, et al. [57]								X		
Galindo et al. [31]			X	X						
Díaz [44]									X	
Ávila [51]							X			
OECD [27]		X		X	X					
Ángeles [58]						X				
Romero et al. [32]		X								
OECD [27]		X								
UNAM [55]							X			

(continued)

Table 4.12 (continued)

Quality in ICT										
Authors	Impact	Services	Products	Competitiveness	Perspectives and attitudes	Technological, political and socioeconomic study	Quantitative indicators	E-learning	B-learning	PLE
López [59]	X		X	X			X			
Soto et al. [33]			X							
Gámiz [46]				X					X	
Cano et al. [36]		X								
Velázquez, et al. [60]				X						
Aleksejeva [52]		X		X			X			
Zapata [37]	X	X				X				
Olufemi [48]			X							
Humanante et al. [38]		X								X
Nolasco et al. [39]	X	X		X			X			
Capanegra et al. [40]		X	X	X						
Evren [49]		X								
Grande [50]	X		X							
Mejía, et al. [61]			X					X		

(continued)

Table 4.13 Frequency of each KPI of Strategic Planning category by authors

Strategic planning									
Indicators	Kaplan [62]	Mintzberg [63]	Mintzberg [64]	Porter [65]	Bermúdez [28]	Ángeles [58]	Mendoza [66]	% Factor	
Teaching innovation					X	X		2	0.0101
Organizational changes						X		1	0.0050
Threat of competitors				X			X	2	0.0101
Suppliers				X				1	0.0050
Customers	X			X				2	0.0101
Substitute products				X				1	0.0050
Government				X				1	0.0050
Plan			X					1	0.0050
Pattern			X					1	0.0050
Position			X					1	0.0050
Perspective			X		X			2	0.0101
Financial elements	X							1	0.0050
Internal business processes	X							1	0.0050
Market creation							X	1	0.0050

(continued)

Table 4.13 (continued)

Strategic planning									
Indicators	Kaplan [62]	Mintzberg [63]	Mintzberg [64]	Porter [65]	Bermúdez [28]	Ángeles [58]	Mendoza [66]		% Factor
Key piece of organization		X						1	0.0050
Type of decentralization		X						1	0.0050
Totals								20	0.1015

Table 4.14 Order of the categories according to the citations

ICT category	Amount	Percent (%)
Quality in ICT	53	26.90
ICT infrastructure	42	21.31
ICT management	34	17.25
ICT integration	28	14.21
Strategic planning	20	10.15
ICT scalability	20	10.15
Total	197	100.00

Table 4.15 Analysis of HEIs based on interviews with their ICT managers

ICT infrastructure elements	HEI1	HEI2	HEI3	HEI4	HEI5	Average (%)
(a) Institutional page and educational platform	✓	✓	X	✓	✓	80
(b) Servers speed	✓	✓	✓	✓	✓	100
(c) Security of databases	✓	✓	X	✓	✓	80
(d) Congruence of institutional policies – ICT	✓	X	X	✓	✓	60
(e) ICT capacity for enrolment assistance	✓	✓	✓	✓	✓	100
(f) Messaging and mail services	✓	✓	X	✓	✓	80
(g) Expectations of growth of ICT	✓	✓	✓	✓	✓	100
(h) ICT technical support services	✓	✓	✓	✓	✓	100
(i) Tele-education	✓	X	X	X	✓	40
(j) Appropriate ICT in classrooms and laboratories	✓	✓	X	✓	✓	80
(k) Enough licenses of educational software	✓	✓	X	✓	✓	80
(l) Algorithmic prioritization of the ICT distribution	✓	✓	X	X	X	40
(m) Telecommunications cover the physical HEI area	✓	✓	X	✓	✓	80
(n) The HEI has updated ICT	✓	✓	X	✓	✓	80
(o) The HEI has ICT indicators measuring systems	✓	X	X	✓	✓	60
Total	100%	80%	26.66%	86.66%	93.33%	77.33%

Table 4.16 Sorting of categories by HEIs

Measuring instrument		KPIs	
Category	(%)	Category	%
ICT infrastructure	82.66	ICT infrastructure	85.32
ICT scalability	82.00	ICT scalability	79.90
ICT integration	80.33	ICT integration	79.08
ICT management	76.00	Quality in ICT	76.53
Quality in ICT	73.45	Strategic planning	69.66
Strategic planning	68.24	ICT management	68.68
Average	77.11%		76.52%

Table 4.17 Standardization of categories by authors, measuring instrument and KPIs

Authors	Measuring instrument						Indicators							
	CT	X_i	$X_i - \mu$	$(X_i - \mu)^2$	Zi	CT	X_i	$X_i - \mu$	$(X_i - \mu)^2$	Zi	CT	X_i	$X_i - \mu$	$(X_i - \mu)^2$
CT	26.90	10.24	104.85	1.70	IF	82.66	5.55	30.80	1.08	IF	85.32	8.8	77.44	1.51
QU	21.31	4.65	21.62	0.77	SC	82.00	4.89	23.91	0.95	SC	79.90	3.38	11.42	0.58
IF	17.25	0.59	0.3481	0.098	IT	80.33	3.22	10.36	0.62	IT	79.08	2.56	6.55	0.43
MA	14.21	- 2.45	6.00	- 0.406	MA	76.00	- 1.11	1.23	- 0.21	QU	76.53	0.01	0.0001	0.0017
IT	10.15	- 6.51	42.38	- 1.08	QU	73.45	- 3.66	13.39	- 0.71	SP	69.66	- 6.86	47.05	- 1.17
SP	10.15	- 6.51	42.38	- 1.08	SP	68.24	- 8.87	78.67	- 1.72	MA	68.68	- 7.84	61.46	- 1.34
SC	16.66		217.57			77.11		158.36			76.52		203.92	

Table 4.18 KPIs of the category of Quality in ICT with their percentages

KPI	Amount	%
Services	14	7.10
Products	9	4.56
Competitiveness	9	4.56
Impact	6	3.04
Quantitative indicators	6	3.04
Perspectives and attitudes	2	1.01
Technological study	2	1.01
e-learning	2	1.01
b-learning	2	1.01
PLE	1	0.50
Total	53	26.90

Table 4.19 Calculation of the KPIs Quality in ICT according to the HEIs

KPI	Items	Summation	Average (%)
QU-02	a-f-h-k-m-n	$80 + 80 + 100 + 80 + 80 + 80 = 500$	83.33
QU-01	a-b-c-e-f-i-j-k-m-n	$80 + 100 + 80 + 100 + 80 + 40 + + 80 + 80 + 80 + 80 = 800$	80.00
QU-10	a-b-c-e-f-i-j-k-m-n	$80 + 100 + 80 + 100 + 80 + 40 + 80 + 80 + 80 + 80 = 800$	80.00
QU-09	a-b-c-e-f-i-j-k-m	$80 + 100 + 80 + 100 + 80 + 40 + + 80 + 80 + 80 = 720$	80.00
QU-08	a-b-c-e-f-i-k-m	$80 + 100 + 80 + 100 + 80 + 40 + 80 + 80 = 640$	80.00
QU-04	a-e-g-i-j-k-n	$80 + 100 + 100 + 40 + 80 + 80 + 80 = 560$	80.00
QU-03	a-j-k-n	$80 + 80 + 80 + 80 = 320$	80.00
QU-06	d-e-g-l-n-o	$60 + 100 + 100 + 40 + 80 + 60 = 440$	73.33
QU-05	d-g-l	$60 + 100 + 40 = 200$	66.66
QU-07	o	60	60.00
Totals			76.53

Table 4.20 Standardized values of the KPIs Quality in ICT

KPI	Authors				KPIs			
	X_i	$X_i - \mu$	$(X_i - \mu)^2$	Z_i	X_i	$X_i - \mu$	$(X_i - \mu)^2$	Z_i
QU-02	7.10	4.416	19.50	2.135	83.33	6.99	48.972	0.99
QU-03	4.56	1.876	3.519	0.907	80.00	3.66	13.454	0.519
QU-04	4.56	1.876	3.519	0.907	80.00	3.66	13.454	0.519
QU-01	3.04	0.356	0.126	0.172	80.00	3.66	13.454	0.519
QU-07	3.04	0.356	0.126	0.172	60.00	-16.33	266.73	-2.31
QU-05	1.01	-1.67	2.802	-0.80	66.66	-9.67	93.547	-1.36
QU-06	1.01	-1.67	2.802	-0.80	73.33	-3.002	9.012	-0.42
QU-08	1.01	-1.67	2.802	-0.80	80.00	3.668	13.454	0.519
QU-09	1.01	-1.67	2.802	-0.80	80.00	3.668	13.454	0.519
QU-10	0.50	-2.18	4.769	-1.05	80.00	3.668	13.454	0.519
Average	2.68		42.77		76.33		498.99	

Table 4.21 Results obtained by AHPy Python library after 1000 iterations with quality ICT KPI

	QU-02	QU-03	QU-04	QU-01	QU-08	QU-09	QU-10	QU-07
QU-02	1	0.333	0.5	0.333	0.111	0.166	0.142	0.142
QU-03	3	1	0.5	0.333	0.5	0.2	0.2	0.2
QU-04	2	2	1	0.5	0.5	0.2	0.25	0.166
QU-01	3	3	2	1	0.5	0.333	0.333	0.25
QU-08	9	2	2	2	1	0.5	0.333	0.5
QU-09	6	5	5	3	2	1	0.5	0.5
QU-10	7	5	4	3	3	2	1	0.5
QU-07	7	5	6	4	2	2	2	1

References

1. Olaskoaga, J., et al.: Semantic diversity and the political nature of the notions of quality in Higher Education in Mexico. High School magazine. Revista de la Educación Superior. **44**(173), 85–102 (2015)
2. Rivera de Parada, A.: Characterization of Higher Education teachers and their conceptualization of university quality. Revista Ciencia, Cultura y Sociedad. **2**(1), 27–38 (2015)
3. Prisacariu, A.: New perspectives of quality assurance in European Higher Education. Procedia Soc. Behav. Sci. **180**(5), 119–126 (2015). <https://doi.org/10.1016/j.sbspro.2015.02.094>
4. Elassy, N.: The concepts of quality, quality assurance and quality enhancement. Qual. Assur. Educ. **23**(3), 250–261 (2015). <https://doi.org/10.1108/QAE-11-2012-0046>
5. Badawy, M., et al.: A survey on exploring key performance indicators. Future Comput. Inf. J. **47**–52 (2016) <https://doi.org/10.1016/j.fcij.2016.04.001>
6. Peng, W., et al.: A semi-automatic system with an iterative learning method for discovering the leading indicators in business processes. In: Proceedings of the 2007 International Workshop on Domain Driven Data Mining (2007). <https://doi.org/10.1145/1288552.1288557>

7. Lazić, Z., et al.: Improvement of quality of Higher Education Institutions as a basis for improvement of quality of life. *Sustainability* **13**, 1–27 (2021). <https://doi.org/10.3390/su13084149>
8. Aly, F., et al.: Prioritizing faculty of engineering education performance by using AHP-TOPSIS and balanced scorecard approach. *Int J Eng Sci Innov Technol* **3**(1), 1–13 (2014)
9. Yean Ong, M., et al.: User acceptance of key performance indicators management systems in a Higher Education Institution in Malaysia: a pilot study. *Int. Online J. Educ. Scie.* 22–31 (2013)
10. Shepherd, S.: Managerialism: an ideal type. *Stud. High. Educ.* 1–11 (2017) <https://doi.org/10.1080/03075079.2017.1281239>
11. Davis, A., et al.: The impact of managerialism on the strategy work. *Stud. High. Educ.* **41**(8), 1480–1494 (2014). <https://doi.org/10.1080/03075079.2014.981518>
12. Cadez, S., et al.: Research, teaching and performance evaluation in academia: the salience of quality. *Stud. High. Educ.* **42**(8), 1455–1473 (2015). <https://doi.org/10.1080/03075079.2015.1104659>
13. Nabi, G., et al.: Does entrepreneurship education in the first year of higher education develop entrepreneurial intentions? The role of learning and inspiration. *Stud. High. Educ.* 1–16 (2016). <https://doi.org/10.1080/03075079.2016.1177716>
14. Bunce, L., et al.: The student as consumer approach in higher education and its effects on academic performance. *Stud. High. Educ.* **42**(11), 1958–1978 (2016). <https://doi.org/10.1080/03075079.2015.1127908>
15. Montgomery, D., Borror, C.: Systems for modern quality and business. *Qual. Technol. Quant. Manage.* 1–10 (2017). <https://doi.org/10.1080/16843703.2017.1304032>
16. Suryadi, K.: Framework of measuring key performance indicators for decision support in Higher Education Institution. *J. Appl. Sci. Res.* **3**(12), 1689–1695 (2007)
17. Khalid, S., et al.: Balanced scoreboard, the performance tool in higher education: Establishment of performance indicators. *Procedia Soc. Behav. Sci.* 4552 – 4558 (2014). <https://doi.org/10.1016/j.sbspro.2014.01.984>
18. Mendenhall, W., et al.: Introduction to probability and statistics. Cengage Learning (2006)
19. Li, H., Ma, L.: Ranking decision alternatives by integrated DEA, AHP and Gower Plot techniques. *Int. J. Inf. Technol. Decis. Mak.* 241–258 (2008) <https://doi.org/10.1142/S0219622008002922>
20. Griffith P.: *ahpy 2.0*. <https://pypi.org/project/ahpy/> (2021). Accessed 13 Mar 2024.
21. Núñez, Y., Rodríguez, C.: Árboles de clasificación para jerarquizar los recursos intangibles asociados a la innovación en las instituciones de educación superior latinoamericanas. *Interiencia* **39**(3), 149–155 (2014)
22. Medina, A., et al.: Identificación y selección de competencias genéricas: Caso Educación Superior Tecnológica en México. *Revista de Estudios y Experiencias en Educación* **11**(22), 99–122 (2012)
23. Pérez, L., et al.: Intuitionistic Fuzzy dimensional analysis for multi-criteria decision making. *Iran. J. Fuzzy Syst.* **15**(6), 17–40 (2018)
24. Pérez L., et al.: Hesitant fuzzy linguistic term and TOPSIS to assess lean performance. *Appl. Sci.* 1–13 (2019). <https://doi.org/10.3390/app9050873>
25. Corbett, S.: Developing contextualised literature-informed competency frameworks for middle managers in education. *Educ. Manage. Admin. Leadership* **51**(6), 1–21 (2021). <https://doi.org/10.1177/17411432211043873>
26. Área M (2005) Tecnologías de la Información y Comunicación en el sistema escolar. Una revisión de las líneas de investigación. *Revista Electrónica de Investigación y Evaluación Educativa* **11**(1), 3–25 (2005)
27. OECD: Resumen ejecutivo la optimización del uso de las TIC. <https://www.oecd.org/governance/50480724.pdf> (2012). Accessed 25 Dec 2012
28. Bermúdez, J., et al.: Las tecnologías de información en las nuevas universidades politécnicas. *Revista Electrónica de Estudios Telemáticos Telemática.* **7**(2), 118–128 (2008)
29. Cobo, J.: El concepto de tecnologías de la información. *Benchmarking sobre las definiciones de las TIC en la sociedad del conocimiento. zer.* **14**(27), 295–318 (2009)

30. Tapia, M.J.: Implementación de las nuevas tecnologías en el proceso educativo: Retos y perspectivas. *Didasc@lia: Didáctica y Educación*. 87–98 (2010)
31. Galindo, J., et al.: La universidad ante el reto de la transferencia del conocimiento 2.0: Análisis de las herramientas digitales a disposición del gestor de transferencia. *Investigaciones Europeas de Dirección y Economía de la Empresa* **17**(3), 111–126 (2011). [https://doi.org/10.1016/S1135-2523\(12\)60123-3](https://doi.org/10.1016/S1135-2523(12)60123-3)
32. Romero, M., et al.: Percepciones en torno al coordinador TIC en los centros educativos inteligentes. Un estudio de caso. *Educación* **2014**, **50**(1), 167–184 (2013)
33. Soto, J.C., et al.: Desarrollo de una metodología para integrar las TIC en las IE (Instituciones Educativas) de Montería. *Revista del Instituto de Estudios en Educación Universidad del Norte*. **21**, 34–51 (2014)
34. Rodríguez, A.L., et al.: Las tecnologías de la información y las comunicaciones y el desarrollo de la capacidad de dirección. *Ciencias Holguín* **3**, 1–11 (2014)
35. Nivien, A.: Factors Affecting the distribution of information and communication technologies in an Egyptian Public University: a case study of the faculty of education at Ain Shams University. *Int. J. Soc. Educ.* **3**(2), 167–187 (2014). <https://doi.org/10.4471/rise.2014.11>
36. Cano, J.A., Baena, J.J.: Tendencias en el uso de las tecnologías de información y comunicación para la negociación internacional. Elsevier (2015). <https://doi.org/10.1016/j.estger.2015.03.003>
37. Zapata, G.: Decisiones claves y aspectos críticos en la Gestión de las TIC. In: IV Congreso Iberoamericano de Enseñanza de la Ingeniería, Venezuela 1–9 May 2013
38. Humanante, P.R., et al.: PLEs en contextos móviles: Nuevas formas para personalizar el aprendizaje. *VAEP-RITA* **4**(1), 33–39 (2016)
39. Nolasco, P., Ojeda, M.M.: La evaluación de la integración de las TIC en la educación superior: fundamento para una metodología. *RED-Revista de Educación a Distancia* **48**(9), 1–24 (2019). <https://doi.org/10.6018/red/48/9>
40. Capanegra, H.A., et al.: El empleo de las TICS en el ámbito universitario. *DAAPGE* **26**, 159–190 (2016)
41. Hernández, P.: Formación de usuarios: modelo para diseñar programas sobre el uso de TI en IES. *Documentación de las Ciencias de la Información* **24**, 151–179 (2001)
42. Lawler, J., Kitchenham, B.: measurement modeling technology. *IEEE Comput. Soc.* 68–75 (2003)
43. Tello, C.G.: Gestionar la escuela en Latinoamérica. *Gestión educativa, realidad y política. Revista Iberoamericana de educación*. **45**(6), 1–10 (2008)
44. Díaz, A.L., Canales, A.: Aplicación de las TIC en la Educación Superior: El caso del SUAyED-UNAM. *Reencuentro* **62**, 30–36 (2011)
45. Montaña, M.J., et al.: Funcionamiento de las Redes Educativas de Centros Escolares: Desarrollo de un trabajo colaborativo. *REOP* **24**(1), 25–41 (2013)
46. Gámiz, V.M.: Lecciones aprendidas de estudios sobre blended-learning en IES. *Revista de Educación Mediática y TIC. Edmeti*, **3**(2), 52–68 (2014). <https://doi.org/10.21071/edmeti.v3i2.2889>
47. Chávez, M.M., et al.: Competencias digitales en el estudiante adulto trabajador. *Revista Interamericana de Educación de Adultos*. **37**(2), 10–24 (2015)
48. Olufemi, A.B., et al.: Effects of computer mediated power point presentation on secondary student's learning outcomes in basic science in Oyo State, Nigeria. *J. Sci. Technol. Math. Educ.* **12**(1), 229–240 (2016)
49. Evren, E., et al.: Delphi technique as a graduate course activity: elementary science teachers' TPACK competences. *SHS Web of Conf.* **26**, 1–6 (2016). <https://doi.org/10.1051/shsconf/20162601135>
50. Grande, M., et al.: Tecnologías de la Información y la Comunicación: Evolución del Concepto y características. *Int. J. Educ. Res. Innov.* **6**, 218–230 (2016)
51. Ávila, G.P., Riascos, S.C.: Propuesta para la medición del impacto de las TIC en la enseñanza universitaria. *Educación y Educadores* **14**(1), 169–188 (2011)
52. Aleksejeva, L.: Country's competitiveness and sustainability: Higher Education impact. *J. Secur. Sustain. Issues* **5**(3), 355–363 (2015). [https://doi.org/10.9770/jssi.2015.5.3\(4\)](https://doi.org/10.9770/jssi.2015.5.3(4))

53. Olivé, L.: Los desafíos de la sociedad del conocimiento: cultura científico-tecnológica, diversidad cultural y exclusión. *Revista Científica de Información y Comunicación* **3**, 29–51 (2006)
54. Pons, J.P.: Higher Education and the knowledge society. Information and digital competencies. *Revista de Universidad y Sociedad del Conocimiento*. **7**(2), 1–15 (2010)
55. UNAM.: Indicadores de desempeño para facultades y escuelas de educación superior UNAM. https://www.planeacion.unam.mx/Planeacion/Apoyo/IndDesFinal_oct31.pdf (2013). Accessed 31 Oct 2013
56. Sampedro, B.E.: Las TIC y la educación social en el siglo XXI. *Revista de Educación Mediática y TIC. Edmetec* **5**(1), 8–24 (2015)
57. Zambrano, W.R., Medina, V.H.: Creación, implementación y validación de un modelo de aprendizaje virtual para la Educación Superior en tecnologías WEB 2.0. *Signo y Pensamiento* **29**(56), 288–303 (2010)
58. Ángeles, A.: Planes estratégicos integrales para la incorporación y uso de TIC: claves para administrar el cambio. *Razón y Palabra*. **79**, 1–14 (2012)
59. López, S.: La calidad de las universidades públicas estatales en México desde la perspectiva de un multi-ranking. *Revista de la Educación Superior* **42**(166), 57–80 (2013)
60. Velázquez, E.C., et al.: La dirección estratégica en la universidad pública: una investigación en las universidades tecnológicas de México. *Universidad & Empresa, Bogotá (Colombia)* **17**(28), 87–104 (2015)
61. Mejía, J.F., López, D.: Modelo de Calidad de e-learning para IES en Colombia. *Formación Universitaria* **9**(2), 59–72 (2016). <https://doi.org/10.4067/S0718-50062016000200007>
62. Kaplan, R.S., Norton, D.P.: *The balanced scorecard: Translating strategy into action*. Harvard Business School Press, Boston (1996)
63. Mintzberg, H., Westley, F.: Sustaining the institutional environment. *Organ. Stud.* **21**(1), 71–94 (2000). <https://doi.org/10.1177/0170840600210005>
64. Mintzberg H., Ahlstrand B.: *Strategy bites back*. Pitman Publishing, London (2005)
65. Porter, M.: The five competitive forces that shape strategy. *Harvard Bus. Rev.* 78–93 (2008)
66. Mendoza, T.: La estrategia del océano azul para emprendedores. *Apunt. cienc. soc* 76–80 (2013). <https://doi.org/10.18259/acs.2013009>

Chapter 5

A Multicriteria Model for the Selection of Online Travel Agencies



Jesus Jaime Solano Noriega, Juan Bernal, and Juan Carlos Leyva Lopez

Abstract The hospitality and tourism industries operate under dynamic business models, where decision-makers continually strive to attract visitors and foster sector development. Given the intricate practical challenges inherent in these industries, decision-makers face the task of evaluating multiple alternatives based on their performance to identify the optimal strategy that maximizes diverse business benefits. Traditionally, travel agents have served as pivotal intermediaries connecting travel service suppliers with travelers. However, advancements in information technology have introduced online travel agencies (OTAs) as a viable alternative for consumers. While OTAs have emerged as a key marketing strategy for hotels to meet their booking objectives, the optimal selection problem for OTAs remains relatively unexplored. Recognizing the economic significance of the sector, this research aims to address the gap in OTAs selection studies. Specifically, it seeks to model the OTAs selection problem under a multicriteria decision-making (MCDA) framework to aid hotel managers in choosing the most suitable OTA. The objective is to optimize the booking process and enhance overall performance. This study proposes an MCDA evaluation framework for OTAs, employing a hierarchical approach with the Electre III method (*h*-ELECTRE III). This method enables the decomposition of the evaluation problem into subtasks, providing decision-makers with insights into the relationships between various evaluation perspectives. The applicability of this evaluation framework is demonstrated through a real-case scenario involving a group of sun and beach hotels, showcasing its effectiveness in such contexts. By leveraging the hierarchical MCDA approach, this research contributes to a more nuanced under-

J. J. Solano Noriega (✉)

Universidad Autónoma de Occidente, Blvd. Macario Gaxiola y Carretera internacional México 15, 81223 Los Mochis, Mexico

e-mail: jaime.solano@uadeo.mx

J. Bernal · J. C. Leyva Lopez

Universidad Autónoma de Occidente, Blvd. Lola Beltrán y Blvd. Mario López Valdez, 80020 Culiacán, Mexico

e-mail: jbbc_62@yahoo.com.mx

J. C. Leyva Lopez

e-mail: juan.leyva@uadeo.mx

standing of the complex OTAs selection problem. The insights derived can empower decision-makers in the hospitality and tourism industries to make informed choices, ultimately optimizing their strategies for attracting and serving visitors.

5.1 Introduction

The hospitality and tourism industries represent dynamic business models, where strategic endeavors are undertaken by industry leaders to allure visitors and foster the growth of this vital economic sphere. Such industries play a pivotal role in bolstering national economies through substantial investments, job creation, and tax revenues, thereby contributing significantly to their vitality and prosperity. As reported by the World Travel and Tourism Council, in 2019, the hospitality and tourism sectors contributed to over 330 million jobs globally and accounted for over 10% of the total Gross Domestic Product (GDP) worldwide [13]. Within the hospitality industry, the hotel sector aims to providing excellent customer service, anticipating guests' needs, and ensuring comfort and satisfaction. A key market strategy used by hotels' managers to attract travelers is by means of travel agents (TA) which are recognized as the pivotal link connecting hotels with travelers, thus establishing a business model wherein TAs oversee, consolidate, categorize available offerings, optimize distribution processes to reduce costs, and format information to be user-friendly and booking-friendly for both suppliers and travelers [9].

Travel agents had been traditionally operating on an agent-principal relationship where the traveler has direct contact with the TA to reach a travel product or service and the TA suggests one of them. However, innovations in business models caused by developments in the field of information technology and internet, the agent-principal model has been surpassed by online travel agents (OTAs) which offer consumers an alternative channel to booking in a more efficient way [11]. OTAs are flexible platforms that let travelers booking any kind of travel product such as flight booking, tours, and hotel rooms, among others. In the hotel sector OTAs have become a key marketing strategy to achieve their booking goals where hoteliers choose to be present in OTAs platforms with the aim to move forward new booking technologies to increase competitiveness and innovation [8].

Selecting an OTA is not a trivial problem and it is considered a supplier selection problem for hotel managers because it can affect revenue, distribution, pricing, brand image, and technology integration [15]. The choice of OTAs can significantly affect the hotel's competitiveness and profitability, making it a critical and complex decision-making problem for hotel managers [17]. However, besides selecting the right OTA is crucial for the profitability and market reach for hotels, this problem remains little studied in the literature and has been mainly focused on studying qualitative studies such as analysis of critical factors regarding customer satisfaction and/or experiences [4], performance and/or quality analysis of OTAs websites [16, 20]. From our point of view, the work of [14], is the only one that proposes an evaluating framework for OTAs selection which discusses the process of evaluating an

online travel website's adherence to its web strategy. It emphasizes the importance of an initial internal evaluation by experts, followed by external surveys. The evaluation process involves the use of Multi-Criteria Decision Analysis (MCDA) methods such as the fuzzy Delphi method to adjust selection criteria, the Decision-making Trial and Evaluation Laboratory (DEMATEL) method to identify interdependencies among perspectives, and the Analytic Network Process (ANP) [19] to rank the available OTAs.

The gaps found in the literature for OTAs selection is wide thus more research can be useful to help Decision Makers (DMs) in the hotel sector with flexible evaluation models. Despite advances in the field, there are still gaps that this study aims to address. This research aims to contribute to research in the field by proposing a MCDA model for the OTAs selection problem using an approach based on evaluating the set of OTAs under different perspective and modeled using a hierarchical structure of point of view from the DM. This approach offers a direct relationship between the DM and the analyst in such a way that the multicriteria model is built by joint agreement between both actors when there exists a hierarchical structure of criteria. The utilization of a hierarchical approach brings notable advantages to modeling multicriteria decision problems [12]. This methodology facilitates the decomposition of intricate decision challenges into more manageable subtasks, simplifying the analytical process. Organizing criteria into levels enhances computational efficiency, allowing for more focused analyses and reducing the computational load. The hierarchical framework provides transparency in decision-making, aiding stakeholders in comprehending the relationships between criteria and their contributions to the overall process. In addition, the flexibility of a hierarchical structure allows the consideration of preference relations within subsets of criteria at any hierarchical level, offering a realistic representation of decision contexts. Moreover, the adaptation of hierarchical structures of criteria into MCDA methods, such as the Elimination and Choice Expressing the Reality (ELECTRE) [7], to handle interaction effects between criteria, such as the *h*-ELECTRE III [2, 3], adds sophistication to decision modeling, enabling a nuanced evaluation of interdependencies among criteria.

The remainder of this paper is structured as follows: Sect. 5.2 presents a formal scheme for structuring decision analysis problems. Section 5.3 introduces the evaluation framework for the OTAs selection problem on a real case study. Section 5.4 presents the results and corresponding discussion. Finally, Sect. 5.5 outlines the conclusions drawn from this study.

5.2 Evaluation Framework for Decision Making

In this section, we present a problem structuring approach aimed at formally delineating the online travel agencies selection problem within the framework of MCDA. MCDA is a process of evaluating a group of alternatives regarding multiple criteria according to the opinions of experts. Since the seventies, MCDA has been advancing

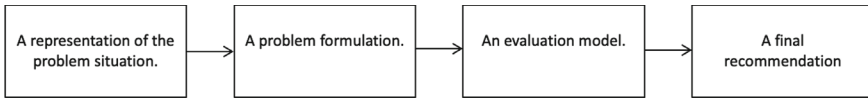


Fig. 5.1 Evaluation scheme for a multicriteria decision making model

together with the development of Information Technology and Computer Science, to support decision makers to adequately address complex decision problems [5, 18].

The selection of an MCDA approach to tackle this decision problem is based and justified on two main premises [13]: (i) due to the complex practical problems involved in hospitality and tourism industries, DMs need to compare multiple alternatives according to their performance and select the optimal scheme to maximize business benefits; and (ii) stakeholders in hospitality and tourism industries are involved in many decision-making scenarios, hence, MCDA methods can help DMs to take objective decisions.

Evaluating a group of alternatives is a complex cognitive process that involves structuring a decision problem to make it more understandable and easier to manage. In this process, it is necessary to define the elements to evaluate, determine an evaluation framework, gather the information, and obtain an assessment by means of the evaluation process aiming to obtain information about the worth of an item (product, service, material, etc.) with the goal to create a clear and coherent representation of the problem, facilitating effective analysis and decision-making. According to [1], a decision problem can be structured following four steps of the decision aiding process (refer to Fig. 5.1).

5.2.1 *A Representation of the Problem Situation*

The initial step is geared towards elucidating the intricacies of the problem, involving a collaborative dialogue between the analyst and the client to gather essential elements for defining the issue. The objective is to acquire pertinent information related to the problem scenario, through inquiries that delve into identifying the entity confronting the problem, understanding its perceived significance, determining responsibility, pinpointing individuals involved in payment decisions, and elucidating what holds utmost importance for the client. This analysis allows the client to gain insights into their position within the decision-making process for which they seek the analyst's assistance. There are mainly three elements that can be used to represent this step a set T of the actors that interact with the process, the set of objectives O or aspirations that each actor has in the process, and a set of resources S allocated by each actor to each object of their interest. This stage gains significance as it enables us to capture a snapshot of the problematic situation that needs to be

depicted when assistance is sought. These elements can form a triplet to represent the problem situation:

$$P = \{T, O, S\}$$

5.2.2 A Problem Formulation

After presenting the problem situation, the analyst might offer the client one or more problem formulations, marking a pivotal stage in the decision process. While the initial output primarily serves a descriptive or explanatory purpose, creating a problem formulation aims to formalize the situation and to incorporate the use of decision support language. The outcome is inherently simplified compared to the complexity of the actual decision process. In this stage is usual to describe the set of possible alternatives A available to the client concerning the problem scenario P , the set of perspectives or points of view V from which the potential alternatives can be viewed, examined, assessed, and compared, and the problem statement π which outlines what is expected to get from the elements in A after the evaluation. The elements described in this stage can be conceptualized as a triplet:

$$P = \{A, V, \pi\}$$

5.2.3 An Evaluation Model

Once formulating a problem, the next step is to develop an evaluation model, which involves organizing available information to generate a formal response to the problem statement. This pivotal task requires to apply a methodological framework to the data provided by the DM, resulting in a model suitable for Decision Analysis. Initially, in this task it should be defined the set of potential alternatives, denoted as A , to which the model applies. While typically established during the problem formulation stage, set A can be modified or adjusted as necessary. Set A is characterized by a set of dimensions D which encompass pertinent information about each element in A . These dimensions serve means under which elements of A are observed, described, and measured, potentially arranged hierarchically. While some dimensions may serve as constraints to establish set A , others are crucial for evaluation, enabling the assessment of each alternative's performance across specific characteristics. The final component of the evaluation model is the selection of a precise method, denoted as R , to generate a solution. This decision is significant, as different methods may lead to varying conclusions. R consists of operators that facilitate the synthesis of information from A through D into a concise evaluation, potentially

resulting in a final recommendation. Although there could be more elements than can be used to create the evaluation model, such as the presence of uncertainty, for the goals of this paper we define the evaluation model as a triplet:

$$M = \{A, D, R\}$$

5.2.4 A Final Recommendation

The evaluation model yields results in the decision support language, however, the final recommendation that serves as the ultimate output, translates these findings into language comprehensible to the DM. While the output aligns with the model, it may not necessarily reflect the concerns of the DM or the nuances of the decision process. Therefore, it's essential for the DM to exercise caution before formulating the final recommendation. Questions to consider include: How does the suggested solution respond to variations in model parameters? What range of parameter values maintains a consistent solution structure? A solution sensitive to minor parameter changes suggests a dependence on technical factors rather than preferential information. In such cases, a thorough examination of the model is warranted to ensure robustness and reliability.

5.3 An Evaluation Framework for OTAs Selection

In this section, we delve into the fundamental components for constructing the evaluation framework model for the OTAs selection problem offering a more comprehensive and detailed specification than what was previously outlined.

5.3.1 A Representation of the Problem Situation

With the aim to provide context for the issue discussed in this paper, we present a real-life scenario within a hotel group based in Mazatlán, Mexico. The primary decision-maker in this context, the sales director of the hotel group, is currently struggling with the challenge of choosing the optimal OTA to establish an annual service contract. The hotel group encompasses four beach hotels, collectively featuring 609 rooms. The core marketing strategy used by the hotel group involves a channel mix that includes wholesalers, their proprietary website and reservation system, traditional travel agencies, and the recent addition of one OTA, which has shown promising results.

While the implemented marketing strategies have achieved relative success, the intense competition in this tourist destination has significantly impacted the occu-

pancy rates in the hotels of the group, consequently affecting overall profitability. This challenging situation has necessitated the formulation of new promotional strategies and the exploration of new markets to enhance occupancy rates, income, and overall profitability.

Acknowledging the potential of the recently incorporated OTA as a significant marketing channel, the sales director is committed to conduct a comprehensive analysis of the available OTAs. The objective is to formally evaluate and choose the most effective OTA that will yield optimal outcomes for the hotel group. This strategic decision-making process is crucial in addressing the challenges posed by intense competition in the tourist destination, aiming to boost occupancy rates, revenue, and overall profitability for the group's four beach hotels in Mazatlán, Mexico.

5.3.2 A Problem Formulation

The presented problem scenario serves as a descriptive overview of the contextual challenges. To transform this narrative into a more actionable form, it becomes imperative to translate the problem situation into a formal structured framework suitable for resolution through a formal model. This formulation is designed to provide a systematic guide for the decision-making process, steering it away from potential ambiguity inherent in judgmental or intuitive decision-making approaches.

In alignment with the decision process outlined in the preceding section, the problem formulation presented in this section entails the definition of a set A , representing the set of alternatives, a set V comprising the various perspectives to evaluate alternatives within A , and π denoting the problem statement elucidating the desired outcomes expected from the elements in A . This structured approach not only facilitates a more precise representation of the problem but also lays the groundwork for generating informed recommendations as part of the decision-making process.

5.3.2.1 Points of View for the Evaluation of Alternatives

In the OTA selection problem, hoteliers face the complex task of evaluating various factors to gauge the competitiveness of each alternative with the aim to choose the one that maximizes value for their hotel's operations and marketing strategies. To systematically analyze and structure the different points of view for the evaluation in this decision-making process, this paper employs the Value-Focused Thinking (VFT) methodology [10] which is an approach designed to enhance decision-making in a multicriteria context by leveraging values.

In the initial stages of the VFT methodology, the emphasis is on establishing and understanding the values that potentially could enhance the decision making. The next step involves structuring these values and formulating objectives, providing decision-makers with a more profound and precise understanding of their priorities within the decision context. The VFT introduces a three-level hierarchical structure

of objectives conformed by: strategic, fundamental, and means-end objectives. The Strategic objective represents the overarching goal that decision-makers aspire to achieve; fundamental objectives delineate the end results that define the primary reasons for interest in the decision and illustrate the essential cause of interest in a given decision situation; finally, means-end objectives are crucial as they outline the steps required to achieve the fundamental objectives. Identifying, structuring, analyzing, and comprehensively understanding these objectives using appropriate strategies is vital to ensuring their completeness.

Each objective integrated into this study was directly derived from the information provided by the DM. Furthermore, the DM received additional support through a comprehensive literature review focused on OTA evaluations. This literature review aimed to enhance the DM's understanding of characteristics that could wield significant influence in the evaluation process. By extracting and identifying the most pertinent assessing aspects from the literature, the study ensures alignment with the preferences expressed by the DM.

The overarching objectives, meticulously defined by the DM, encompass diverse perspectives to facilitate an optimal OTA selection, ensuring alignment with the decision maker's preferences. OTAs are an interesting means of achieving reservations and therefore increasing the revenue of the hotel group. Therefore, the overarching strategic objective is to analyze and select the most optimal OTA. This strategic decision is pivotal in ensuring that the chosen OTA aligns seamlessly with the group's financial objectives, thereby maximizing the potential for meeting and surpassing established financial goals. From this pivotal objective, the DM delineated five distinct functional objectives. Firstly, the evaluation should assess the user experience during website interaction. Additionally, there is a specific emphasis on prioritizing ease of use, emphasizing simplicity and intuitiveness within this interaction. The second functional objective centers on evaluating platform operations, focusing on flexibility for users to modify bookings, ensuring responsive support for addressing booking problems, and implementing a transparent fee structure for services. The third objective delves into assessing platform credibility, vital for customer satisfaction, incorporating considerations such as the number of users and the establishment of a reputable brand image. The fourth objective centers on evaluating the variety of products offered, emphasizing their potential to add value to the hotel group. Lastly, the fifth objective directs attention towards assessing the company's commitment fulfillment with its business partners, encompassing conflict resolution and technical support availability. This comprehensive framework ensures a systematic and holistic approach to OTA evaluation, addressing a spectrum of crucial aspects for the decision-making process. Figure 5.2 shows the hierarchical structure of these objectives.

5.3.2.2 Definition of the Set of Alternatives

The primary objective in this real-life case is to carefully choose the most suitable OTA that aligns with the group's expectations and that can yield the desired outcomes,

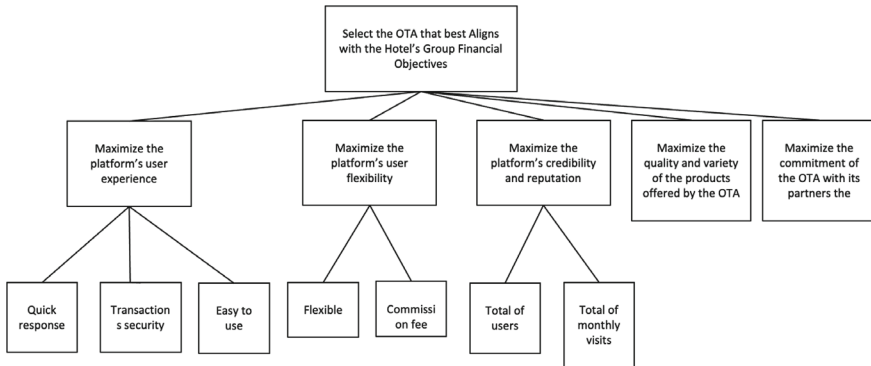


Fig. 5.2 Hierarchical objectives' structure for the OTAs selection problem

specifically an increase in the profitability of the hotel group. Given this objective, it becomes evident that the set *A* will consist of existing options, rendering the creation or formulation of additional alternatives unnecessary for the decision-maker.

To accomplish this, a preliminary analysis is underway to identify the most successful OTAs in the market. Subsequently, their operational approaches will be scrutinized to delineate the most relevant criteria for consideration in selecting the optimal option.

In the contemporary landscape, the internet and technological access serve as pivotal tools for OTAs in offering their tourist services. Each OTA employs unique strategies, combining technology and methods to enhance competitiveness and appeal to their clientele, thereby delivering value to customers.

The array of OTAs available for evaluation is vast, making the strategic decision faced by the decision-maker involve defining the perspectives for OTA selection. In the initial analysis, the decision-maker identified that some of the most successful OTAs in the market include Airbnb, Booking, Despegar, Expedia, Hostelworld, HotelTonight, PriceTravel. This step marks the commencement of a comprehensive evaluation process to determine the best-suited OTA that aligns with the hotel group's objectives and operational preferences. Table 5.1 gives a description for each OTA in the set *A*.

5.3.2.3 Problem Statement

Addressing the problem of selecting the optimal OTA involves evaluating the defining set of OTAs and subsequently ranking them to identify the best overall option. Employing a ranking system is well-suited for this task as it enables the identification of the next most crucial alternative, useful in scenarios where negotiations with the primary OTA prove unsuccessful. This comprehensive ranking approach provides

Table 5.1 The set of OTAs and their corresponding descriptions

OTA	Description
Booking.com	Booking.com is one of the largest and best-known travel websites in the world. Every day it obtains more than 1 and a half million reservations. Both its website and mobile application are available in 43 languages and gather more than 50 million reviews from verified guests of properties in more than 200 countries. Its parent company, Booking Holdings, includes brands such as Priceline.com and metasearch engines such as Kayak and HotelsCombines, Agoda, Rentalcars.com and Open Table
Expedia	Expedia is the star website of the Expedia Group. It brings together 200 travel booking websites such as Hotwire, Orbitz and Travelocity. It receives more than 600 million visits per month from 75 countries and is available in 35 languages. This website allows you to search for all types of accommodation, as well as offers on cheap flights, car rentals, cruises and vacation packages. Hotel properties that advertise on Expedia are simultaneously advertised on Hotels.com, Wotif and other channels
Airbnb	Airbnb is known for its unique accommodation offering around the world. The vast majority of properties in its catalog are located outside the main hotel areas. The company continues to expand into hospitality after acquiring HotelTonight (known for offering last-minute bookings), but its community is still looking for special getaways and experiences when booking. This platform is ideal for booking a more original vacation rental with experiences oriented towards local environments
Hostelworld	HostelWorld is a travel website where you can find the best hostel deals anywhere in the world. It has more than 12 million guest reviews, and in addition to hostels, its inventory includes hotels, B&Bs and other types of budget accommodation. Hostelworld aims to be the leader in social travel through its booking options, mobile apps and attention to the needs of guests looking to connect with others while traveling the world. Its website and mobile app are available in 20 languages
Despegar	Despegar/Decolar is the largest online travel company in Latin America. Despegar is the company's global brand and Decolar represents the brand in the Brazilian market. Its marketplace operates in 20 markets and offers a wide selection of products and services in the region, including airline tickets, hotel reservations and travel packages to a huge customer base
HotelTonight	Is a travel agency and metasearch engine owned by Airbnb and accessible via website and mobile app. It is used to book last-minute lodging in the Americas, Europe, Japan and Australia
Price Travel	Began its operations in the Mexican city of Cancun in the year 2000. Our team is made up of people from different nations such as Mexico, Argentina, Austria, Germany, and the United States, bringing together more than 60 years of experience in different areas of knowledge including internet, travel, tourism, and sales

the DM with a broad perspective on all the OTAs, offering valuable insights for informed decision-making and guiding the negotiation strategy to be implemented.

5.3.3 An Evaluation Model

Following the elucidation of the problem formulation, the subsequent crucial step involves crafting an evaluation model, aimed at organizing available information to provide a formal response to the problem statement π . In this study, the evaluation model comprises the explicitly defined set of alternatives, which encompasses the identified OTAs outlined in the problem formulation.

The model incorporates a coherent family of criteria D inferred from the means-end objectives defined in the problem formulation which offers a comprehensive framework for the assessment of the defined set of OTAs A . Table 5.2 describe the set of criteria D . The set A can be then described through the set D representing the relevant knowledge that the DM has about A .

5.3.3.1 Set of Alternatives' Performances on Each Criterion

To evaluate the performance of each alternative across criteria, the DM gathered information tailored to the nature of each criterion. Table 5.3 provides a comprehensive overview of the evaluation tool and the type of data applicable to each criterion. The assessments for criteria g_1 , g_2 , g_3 , and g_4 were conducted using Grader, a web evaluation platform available at <https://website.grader.com/>, specifically designed to analyze the performance of websites. For criteria g_6 and g_7 , performances were gauged through Google Analytics 4 <https://developers.google.com/analytics/devguides/collection/ga4>, allowing measurement of traffic and interaction between websites and applications. Lastly, for criteria g_5 , g_8 , and g_9 , the DM performed a subjective evaluation based on personal experiences, knowledge, and perceptions regarding each OTA. Table 5.3 shows the derived performance matrix.

5.3.3.2 Model

The final component to establish within the evaluation model is the specific method \mathcal{R} to be employed in crafting a solution. The choice of \mathcal{R} is contingent upon the problem statement π adopted in the problem formulation and must align seamlessly with the utilized information. The multicriteria decision analysis tool employed to execute the evaluation model and derive a definitive ranking for the OTA selection problem is the h -ELECTRE III method. This method, leveraging criteria hierarchy, outranking principles, and preference and indifference thresholds, proves to be a judicious choice for situations akin to those outlined in the study. It was defined to generate a ranking of alternatives and individual rankings for each non-elementary

Table 5.2 Description of the set of criteria and the perspective where each criteria belongs to in the hierarchical structure (Dir: Preference direction of criteria, Minimize or Maximize; UM: Unit of measure)

Perspective	Criteria	Description	Dir	UM
Use	g_1 : Quick.	Ability to complete transactions quickly.	Min	Seconds
	g_2 : Secure	Ability to complete transactions securely.	Max	Scale 1–10
	g_3 : Easy	Easy to use for the end user.	Max	Scale 1–30
System	g_4 : Flexible	Allows users to modify bookings.	Max	Scale 1–30
	g_5 : Commission-based	Commission fee is charged to the hotel.	Min	Percentage
Credibility	g_6 : User	Number of APP Users that has the platform.	Max	Millions
	g_7 : Reputable	Number of monthly visits.	Max	Millions
Product	g_8 : Product	Has variety of high quality products.	Max	Scale 1–10
Company	g_9 : Company	Acknowledging and respecting needs and goals of relationship partner.	Max	Scale 1–10

Table 5.3 Performance matrix

	Use			System		Credibility		Product	Company
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
a_1 : Airbnb	15.7	10	20	30	18%	100	106.8	7	9
a_2 : Booking	25.7	5	20	30	20%	500	678.4	10	9
a_3 : Despegar	3.7	5	30	25	16%	10	1.3	9	8
a_4 : Expedia	25	10	20	30	20%	50	92.4	10	9
a_5 : Hostelworld	16.3	10	30	25	16%	5	61.7	8	8
a_6 : HotelTonight	17	5	20	30	16%	0.1	3.1	8	10
a_7 : PriceTravel	3.1	10	30	30	14%	0.1	2.9	8	10

criterion, arranged in a descending order of preference. As with any outranking method the *h*-ELECTRE III follows a construction and an exploitation stages. The former stage creates a model of preferences of the DM by aggregating the given input, while the last stage creates a partial preorder of the alternatives from the aggregated model of preferences.

Aggregation Procedure

When addressing a non-elementary criterion g_r from the set G , in *h*-ELECTRE III, a partial outranking relation S_r is established over $A \times A$. This binary relation indicates that $a S_r b$ “ a is at least as good as b ” with regard to criterion g_r . Following this, each pair of elements (a, b) , is assessed to confirm the assertion $a S_r b$. The criteria significance are denoted by their weights W_t which are normalized such as $\sum_{g_t \in G^L} W_t = 1, \forall g_t \in G^L$ and $W_t > 0$.

Indifference q_t , preference p_t , and veto threshold v_t have to be defined $\forall g_t \in G^L$. The indifference threshold signifies the maximum performance difference between alternatives a and b on g_t that allows for their indifference on g_t . The preference threshold represents the minimum performance difference between a and b on g_t necessary for expressing a preference of one over the other on g_t . Meanwhile, the veto threshold denotes the minimum performance difference between b and a on g_t that contradicts the outranking of a and b on any criterion g_r from which g_t is derived, specifically when $g_t \in G(g_r)$. To maintain consistency, it is required that $v_t > p_t \geq q_t \geq 0$.

For any pair of alternatives (a, b) belonging to the set A , where $g_t(a) \geq g_t(b)$ holds true for all $g_t \in G(g_r)$, three preference relations can be established:

Per-Criterion Indifference Relation: When it comes to the criterion g_t , alternatives a and b are considered indifferent $a I_t b$ whenever the absolute difference between their performances, $|g_t(a) - g_t(b)|$, is less than or equal to q_t .

Per-Criterion Strict Preference Relation: In terms of the criterion g_t , alternative a is distinctly favored over b $a P_t b$ whenever $g_t(a) - g_t(b) > p_t$. Let $C(a P_t b)$ designate the subgroups of criteria for which $a P_t b$ holds true.

Per-Criterion Weak Preference Relation: For the criterion g_t , alternative a is weakly preferred over b $a Q_t b$, whenever $q_t < g_t(a)/g_t(b) \leq p_t$. Subsequently, let $C(a Q_t b)$ indicate the subgroups of criteria where $a Q_t b$ is applicable.

These three preference relations can be consolidated into a single outranking relation which encompasses the three respective scenarios $S_t = P_t \cup Q_t \cup I_t$, where $a S_t b$ (a outranks b) indicates that “ a is at least as good as b ” regarding the criterion g_t .

In the establishment of a partial outranking relation, denoted as $a S_r b$, the *h*-ELECTRE III incorporates the concordance principle. This principle mandates that, following the assessment of their relative significance, a majority of elementary subcriteria must align in favor of the statement “alternative a outranks alternative b ”. Additionally, a non-discordance principle is employed, stipulating that among the minority of elementary subcriteria that don’t endorse this claim, none should vehemently oppose it.

In *h*-ELECTRE III a concordance index is established $c_r(a, b), \forall(a, b) \in A \times A$, indicating a sufficiently strong concordant group of criteria supporting the assertion “*a* outranks *b*”. The strength of each elementary criterion is determined by its weight, w_t . Specifically, the potency of the concordant group is derived from the criteria endorsing the statement “*a* outranks *b*” along with a fraction of the influence of those criteria where “*b* is weakly preferred to *a*”. This perspective can be represented by the following partial concordance index for each non-elementary criterion $g_r \in G$:

$$c_r(a, b) = \sum_{g_t \in G_r, \hat{g}_t \in C(aI, Q, Pb)} W_t + \sum_{g_t \in G_r, \hat{g}_t \in C(bQa)} \varphi W_t \quad (1)$$

where $\varphi_t = \frac{p_t - [g_t(b) - g_t(a)]}{p_t - q_t} \in [0, 1]$, assuming $\sum_{g_t \in G^L} W_t = 1$.

Should be noted that $c_r(a, b) \in [0, W_r]$ where $W_r = \sum_{g_t \in G^L} W_t$ and $c_r(a, b) = 0$ if $g_t(a) + p_t \leq g_t(b)$, for all $g_t \in G(g_r)$ (*b* is strictly preferred to *a* on all elementary sub-criteria descending from g_r), and $c_r(a, b) = W_r$ if $g_t(a) + q_t \geq g_t(b)$, for all $g_t \in G(g_r)$ (*a* outranks *b* on all elementary sub-criteria descending from g_r). When $r = 0, c(a, b) \in [0, 1]$ because $G(g) = G^L$ and $W = 1$.

The concept of non-discordance suggests the absence of a significant minority coalition among elementary criteria that could challenge the assertion that “*a* outranks *b*”. The counteracting influence of the elementary criterion $g_t \in G(g_r)$ is determined by the veto threshold v_t . This perspective could be represented by an elementary criterion discordance index for each elementary criterion g_t structured as follows:

$$d_t(a, b) = \begin{cases} 1 & \text{if } g_t(b) - g_t(a) \geq v_t \\ \frac{[g_t(b) - g_t(a)] - p_t}{v_t - p_t} & \text{if } p_t < g_t(b) - g_t(a) < v_t \\ 0 & \text{if } g_t(b) - g_t(a) \leq p_t \end{cases} \quad (5.1)$$

The last stage in the aggregation procedure involves merging these two indices to generate a partial credibility index $\sigma_r(a, b)$, where $0 \leq \sigma_r(a, b) \leq 1$, for every non-elementary criterion g_r , quantifying the extent of outranking between *a* and *b*. The resulting $\sigma_r(a, b)$ is calculated as:

$$\sigma_r(a, b) = c_r(a, b) \prod_{g_t \in G(g_r)} T_t(a, b) \quad (5.2)$$

where

$$T_t(a, b) = \begin{cases} \frac{1 - d_t(a, b)}{1 - c_r(a, b)} & \text{if } d_t(a, b) > c_r(a, b) \\ 1 & \text{otherwise} \end{cases} \quad (5.3)$$

Equation (5.3) operates under the assumption that if the power of the partial concordance surpasses that of elementary discordance, the partial concordance value remains unaltered. However, if this condition is not met, we need to question the assertion $aS_r b$ and adjust $c_r(a, b)$ according to the provided equation. Consequently, we have established a partial fuzzy outranking relation $S_A^{\sigma_r}$ defined over $A \times A$. This means that each ordered pair $(a, b) \in A \times A$ is associated with a real number

$\sigma_r(a, b)$, where $0 \leq \sigma_r(a, b) \leq W_r$, indicating the degree of support for the crisp outranking relation $aS_r b$. With this, the construction of the aggregated model of preferences is finalized.

Exploitation Procedure

The subsequent stage in the multicriteria outranking methodology involves the exploitation of the model to attain a conclusive partial preorder of alternatives through the fuzzy outranking relation S_A^σ for each non-elementary criterion g_r . This process is facilitated by the distillation procedure [2], a method grounded in multiple cuts that yields two complete preorders, known as descending and ascending distillations. In the descending distillation, alternatives are arranged from the best to the worst, while the ascending distillation reverses the order, ranking alternatives from the worst to the best. The intersection of these preorders yields a partial order, suggesting the final ranking of the alternatives. We find this distillation procedure particularly apt for this problem, as it has demonstrated efficacy, especially in scenarios involving a modest set of alternatives within the ELECTRE III framework.

To proceed in creating both distillations for the h -ELECTRE III we can follow the next steps:

With a non-elementary criterion and a collection of lambda cut levels, both distillations address the fuzzy outranking relation to construct a crisp outranking relation as follows :

$$aS_r^{\lambda_k} b \iff \begin{cases} \sigma_r(a, b) > \lambda_k \\ \sigma_r(a, b) > \sigma_r(b, a) + s(\sigma_r(a, b)) \end{cases} \quad (5.4)$$

Where $s(\lambda) = \alpha\lambda + \beta$, $\alpha = -0.15$, and $\beta = 0.3$ (Note that these parameters and their values are suggested in [2]).

Then, from the crisp outranking relation $aS_r^{\lambda_k} b$, some calculations are made for each alternative on every non-elementary criterion in G_r

- The λ_k -power of a , $p_{r,A}^{\lambda_k}(a) = |\{a \in A : aS_r^{\lambda_k} b\}|$: counts the alternatives that are outranked by a on G_r
- The λ_k -weakness of a , $f_{r,A}^{\lambda_k}(a) = |\{b \in A : aS_r^{\lambda_k} a\}|$: counts the alternatives that outrank a on G_r

After calculated λ_k -power and λ_k -weakness of a its relative position is calculated as:

$$q_{r,A}^{\lambda_k}(a) = p_{r,A}^{\lambda_k}(a) - f_{r,A}^{\lambda_k}(a) \quad (5.5)$$

Algorithm 1 delines the steps to perform the distillation procedure.

From the ascending and descending distillations, we obtain two comprehensive pre-orders. Within each, alternatives are organized into groups based on their ranking equivalence. To get the final ranking a partial pre-order is derived from the intersection of both pre-orders (ascending and descending). Such a partial pre-order facilitates

Algorithm 1 Distillation Procedure

Require: $\sigma_r(a, b)$ on $G_r, \forall a, b \in A$

- 1: $n \leftarrow 0$
 - 2: $\bar{A}_0 \leftarrow A$ or $\underline{A}_0 \leftarrow A$
 - 3: $\lambda_0 \leftarrow \max_{\substack{a, b \in \bar{A}_n \\ a \neq b}} \sigma_r(a, b)$ or $\lambda_0 \leftarrow \max_{\substack{a, b \in \underline{A}_n \\ a \neq b}} \sigma_r(a, b)$
 - 4: $k \leftarrow 0$
 - 5: $D_0 \leftarrow \bar{A}_n$ or $D_0 \leftarrow \underline{A}_n$
 - 6: $\lambda_{k+1} \leftarrow \max_{\substack{\sigma_r(a, b) > \lambda_k - s(\lambda_k) \\ a, b \in D_k}} \sigma_r(a, b)$ ▷ Should be noted that: $\forall a, b \in D_k$, if $\sigma_r(a, b) > \lambda_k - s(\lambda_{k+1})$ then $\lambda_{k+1} \leftarrow 0$
 - 7: $\forall a \in D_k$ on G_r determine λ_k -quantifications
 - 8: Find the maximum \bar{q}_{D_k} or minimum \underline{q}_{D_k} λ_k -quantifications
 - 9: Calculate $\bar{D}_{k+1} \leftarrow \{a \in D_k : q_{r,A}^{\lambda_{k+1}}(a) = \bar{q}_r, D_k\}$ or $\underline{D}_{k+1} \leftarrow \{a \in D_k : q_{r,A}^{\lambda_{k+1}}(a) = \underline{q}_r, D_k\}$
 - 10: **if** $|\bar{D}_{k+1}| = 1$ or $|\underline{D}_{k+1}| = 1$ or $\lambda_{k+1} = 0$ **then**
 - 11: Go to Step 16
 - 12: **else**
 - 13: $k \leftarrow k + 1, D_k \leftarrow \bar{D}_k$ or $D_k \leftarrow \underline{D}_k$
 - 14: Go to Step 6
 - 15: **end if**
 - 16: $\bar{C}_{n+1} \leftarrow \bar{D}_{n+1}$ or $\underline{C}_{n+1} \leftarrow \underline{D}_{n+1}$
 - 17: $\bar{A}_{n+1} \leftarrow \bar{A}_n \setminus \bar{C}_{n+1}$ or $\underline{A}_{n+1} \leftarrow \underline{A}_n \setminus \underline{C}_{n+1}$
 - 18: **if** $\bar{A}_{n+1} = \emptyset$ or $\underline{A}_{n+1} = \emptyset$ **then**
 - 19: $n \leftarrow n + 1$
 - 20: Go to Step 3
 - 21: **else**
 - 22: The distillation procedure is completed.
 - 23: **end if**
-

the comparisons between alternatives and highlights potential incomparabilities. The intersection of both pre-order can be calculated following three rules:

- An alternative a will be deemed superior to b if, in one of the distillations, a outranks b , and in the other distillation, a is ranked equally to or higher than b .
- An alternative a will be deemed equivalent to b if both alternatives are part of the same equivalence class in both preorders.
- Alternatives a and b are considered incomparable if a is ranked higher than b in the ascending distillation while b is ranked higher than a in the descending distillation, or vice versa.

Modeling the OTAs Selection Problem

As mentioned before, a critical characteristic of h -ELECTRE III is the set of parameters that should be given to get an output. These parameters represent preferential information of the DM. These are basically four mandatory parameters in order to get a recommendation from h -ELECTRE III: set weights, preference, indifference, and veto thresholds. For this case study, preference and indifferent thresholds were given from the DM after a conscious analysis on the data of the performance matrix.

Table 5.4 Weights (W), Indifference (q), and Preference (p) thresholds

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
W	0.05	0.10	0.05	0.20	0.30	0.10	0.05	0.10	0.05
q	1	1	5	1	0.1	1	5	0.1	0.1
p	3	2	10	5	0.2	10	10	1	1

Veto thresholds were not considered by the DM thus a value of zero was set for veto in all the criteria. To determine the weights, the DM employed the 'Pack of Cards' Technique from [6]. Table 5.4 shows the final weights, and the indifference and preference thresholds set by the DM. Finally data from the performance matrix (Table 5.3) and the corresponding weights and thresholds (Table 5.4), were modeled using the h -ELECTRE III and the distillation procedure to construct a partial order of classes of alternatives.

5.4 Results and Discussion

We present in this section the results obtained after aggregating and exploiting the data in this case study using the h -ELECTRE III method and the distillation procedure. We discuss the results at the level of the overall criterion and on each fundamental objective. Starting with the overall ranking we can see in Fig. 5.3a the final order of the set of alternatives A on the global perspective. Such a Figure suggests a ranking with five classes of alternatives, some classes contain one alternative while others contain two alternatives. The first class in the ranking contains a_7 (Price Travel) which is the most preferred alternative according to the preferences of the DM. In the second ranking, there is an equivalence class with alternatives a_2 (Booking) and a_3 (Despegar) that are considered to be indifferent to each other. Next, we have in the third ranking, a class with alternative a_1 (Airbnb), followed by a class on rank four with alternative a_6 (Hotel Tonight). Finally at the bottom of the ranking, we find a class with alternatives a_4 (Expedia) and a_5 (Hostel World) which seems to be the less preferred alternatives with the lowest performances.

From the perspective of each fundamental objective, we can see how each alternative performs in each stem of the structure. Results from exploiting elementary criteria on the fundamental objective "User", the ranking suggests that the best alternative is a_7 (Price Travel) followed by a_5 (Hostel World); in the third position we find an equivalence class with a_1 (Airbnb) and a_3 (Despegar); in the fourth, fifth and sixth position in the ranking are a_4 (Expedia), a_6 (Hotel Tonight), and a_2 (Booking), respectively. Respecting the fundamental objective "System", it shows that a_2 (Booking) is the most preferred alternative followed by a_1 (Airbnb). In the third ranking there is a_4 (Expedia) followed by rank fourth with an equivalence class that includes alternatives a_3 (Despegar) and a_5 (Hostel World). At the bottom of the ranking there

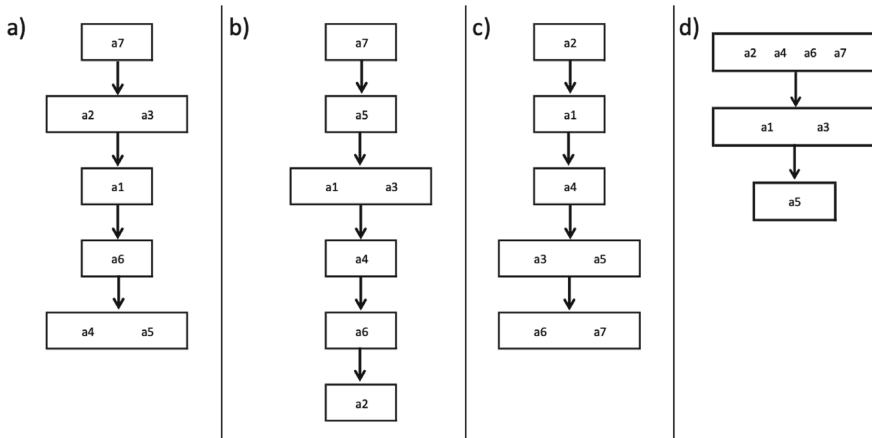


Fig. 5.3 a Depicts the Global ranking; b–d depict the partial rankings for the fundamental objectives “Use”, “System”, and “Credibility” respectively

is an equivalence class with a_6 (Hotel Tonight) and a_7 (Price Travel) that are considered the alternatives with less performance under this fundamental criterion. In the fundamental objective “Credibility” we can observe that in the first position there is an equivalence class that includes a_2 (Booking), a_4 (Expedia), a_6 (Hotel Tonight), and a_7 (Price Travel); in the second position we find a_1 (Airbnb) and a_3 (Despegar) that are considered indifference each other; in the last ranking there is a_5 (Hostel World) as the less proffered alternative. This decomposition of the evaluation on the fundamental criteria, let us have a different perspective on the global ranking, and let us justify the final solution that the DM could take from the different points of view that were established for the evaluation framework. From these perspectives, that is, different Rankings, we can observe that the global ranking reflects an evaluation on all fundamental criteria and is a_7 (Price Travel) the one that seems to be the more proffered alternative as it appears in the first position in all rankings, but on the ranking on fundamental objective “System”, where it is the less proffered alternative along a_6 (Hotel Tonight). The rest of the alternatives hesitates in different position in all rankings. It also should be noted that alternative a_2 (Booking) persists near the top in two fundamental objectives, “System” and “Credibility”, and in the global ranking, hence could be considered a good candidate as a second selection option.

Sensitivity Analysis of the Global Ranking

Analyzing how the global ranking responds to variations in parameters such as weights (w), indifference (q), and preference thresholds (p) offers valuable insights into the robustness of the decision-making process. To understand the impact of these parameters on the global ranking of the OTAs, we conducted a sensitivity analysis. The results of this analysis, presented in Tables 5.5 and 5.6, reveal the intricacies of the presented evaluation methodology. These two tables show the criterion that was modified, the weight or threshold value that was set for the sensitivity analy-

Table 5.5 Resulting global ranking after variations in criteria weights

Criterion	Weight	Global ranking
g_1	0.1	$a_7 > a_2, a_3 > a_1 > a_6 > a_4, a_5$
g_2	0.2	$a_7 > a_2, a_3 > a_1 > a_4, a_5, a_6$
g_3	0.1	$a_7 > a_1 > a_2, a_3, a_5 > a_4 > a_6$
g_4	0.4	$a_7 > a_1 > a_6 > a_4 > a_2, a_3 > a_5$
g_5	0.6	$a_7 > a_1 > a_3 > a_4 > a_2 > a_5, a_6$
g_6	0.2	$a_7 > a_1 > a_6 > a_2 > a_4 > a_3, a_5$
g_7	0.1	$a_7 > a_6 > a_5 > a_3 > a_1 > a_2, a_4$
g_8	0.2	$a_7 > a_2 > a_1 > a_3 > a_5, a_6 > a_4$
g_9	0.1	$a_7 > a_3 > a_6 > a_4 > a_1 > a_2 > a_5$

Table 5.6 Resulting global ranking after variations in indifference and preference thresholds

Criterion	Thresholds	Global ranking
g_1	$q=2, p=5$	$a_7 > a_2 > a_3 > a_1 > a_6 > a_4, a_5$
g_5	$q=1, p=2$	$a_7 > a_6 > a_2 > a_3, a_5 > a_1 > a_4$
g_6	$q=50, p=100$	$a_7 > a_6 > a_2 > a_3, a_5 > a_1 > a_4$
g_7	$q=50, p=100$	$a_7 > a_2, a_3 > a_1 > a_6 > a_4, a_5$
g_8	$q=1, p=2$	$a_7 > a_2 > a_1 > a_3, a_4, a_5, a_6$
g_9	$q=1, p=2$	$a_7 > a_2, a_3 > a_1 > a_4, a_5, a_6$

sis and the resulting global ranking with such a modification. Despite introducing a total of 15 adjustments, from the original parameter values detailed in Table 5.3, the global ranking outlined in Fig. 5.3 remained largely unchanged. This suggests that, within the spectrum of parameter modifications explored, the influence on our results (ranking) was deemed minimal. Alternative a_7 (Price Travel) remains stable in all rankings, while the rest of alternatives show slight deviations from the ranking created from the original parameters).

5.5 Conclusions and Future Research

The hospitality and tourism industries are crucial components of the global economy, providing millions of jobs and contributing significantly to the GDP. Within this landscape, the hotel sector relies heavily on TAs and, more recently, OTAs to attract travelers and drive bookings. The shift from traditional travel agent models to OTAs has revolutionized the way travelers book accommodations, offering more flexibility and efficiency. For hotel managers, selecting the right OTA is a critical decision that can impact revenue, distribution, brand image, and technology integration. Despite the importance of OTA selection, this area remains under explored in the literature.

Existing studies have mainly focused on qualitative analyses of customer satisfaction, OTA website performance, and critical factors influencing selection. There is a lack of comprehensive, quantitative models to guide hotel managers in OTA selection.

To address this gap, this research proposed a MCDA model, based on the *h*-ELECTRE III, for the OTAs selection that offered a systematic framework for evaluating OTAs based on various perspectives and criteria, allowing hotel managers to make informed decisions.

The hierarchical structure of the MCDA model proposed offers several advantages over the traditional approaches to OTA selection: (i) it simplifies complex decision challenges by breaking them down into manageable subtasks, making the analytical process more straightforward; (ii) it provides transparency in decision-making, helping stakeholders understand the relationships between criteria and their contributions to the overall process; and (iii) the hierarchical structure allows for the consideration of preference relations within subsets of criteria, providing a realistic representation of different points of view or decision contexts. By utilizing the MCDA model, hotel managers can evaluate OTAs based on their specific needs and preferences, leading to more informed and strategic decisions. This research contributes to the field by providing a structured approach to OTA selection, filling a crucial gap in the existing literature. However, there is still ample room for improvement when analyzing the OTAs selection problem. Therefore, future work could focus on integrating additional qualitative criteria and exploring different MCDA approaches to more accurately model qualitative factors. Moreover, further research could delve into modeling this problem as a group MCDA, particularly in sectors where decision-making involves a board of managers who collectively choose providers.

References

1. Bouyssou, D., Marchant, T., Pirlot, M., Tsoukias, A., Vincke, P.: Evaluation and Decision Models with Multiple Criteria: Stepping Stones for the Analyst, vol. 86. Springer Science & Business Media (2006)
2. Corrente, S., Figueira, J.R., Greco, S., Słowiński, R.: A robust ranking method extending electre iii to hierarchy of interacting criteria, imprecise weights and stochastic analysis. *Omega* **73**, 1–17 (2017) <https://doi.org/10.1016/j.omega.2016.11.008>. <https://www.sciencedirect.com/science/article/pii/S0305048316301621>
3. Corrente, S., Greco, S., Słowiński, R.: Multiple criteria hierarchy process with electre and promethee. *Omega* **41**, 820–846 (2013). <https://doi.org/10.1016/j.omega.2012.10.009>. <https://www.sciencedirect.com/science/article/pii/S0305048312002046>
4. Dutta, S., Chauhan, R.K., Chauhan, K.: Factors affecting customer satisfaction of online travel agencies in India. *Tour. Hosp. Manag.* **23**, 267–277 (2017)
5. Figueira, J., Greco, S., Ehrgott, M.: Multiple Criteria Decision Analysis: State of the Art Surveys. Springer Science & Business Media (2005)
6. Figueira, J., Roy, B.: Determining the weights of criteria in the ELECTRE type methods with a revised Simos' procedure. *European J. Operat. Res.* **139**, 317–326 (2002)
7. Figueira, J.R., Greco, S., Roy, B., Słowiński, R.: An overview of ELECTRE methods and their recent extensions. *J. Multi-Crit. Dec. Anal.* **20**, 61–85 (2013)

8. Inversini, A., Masiero, L.: Selling Rooms Online: The Use of Social Media and Online Travel Agents [Summary]. Virginia Tech (2014)
9. Jager, K.D.: Choosing Between Travel Agencies and the Internet. University of Johannesburg, South Africa (2015)
10. Keeney, R.L.: A Path to Creative Decisionmaking. Harvard University Press (1992). <https://doi.org/10.2307/j.ctv322v4g7>. <http://www.jstor.org/stable/j.ctv322v4g7>
11. Lee, H.A., Guillet, B.D., Law, R.: An examination of the relationship between online travel agents and hotels: A case study of choice hotels international and expedia.com. *Cornell Hosp. Quart.* **54**, 95–107 (2013)
12. Leyva, J.C., Flores, S., Solares, E., León, M., Díaz, R., Flores, A.: Multicriteria decision model to support the evaluation of common jurisdiction violence in the capital cities of the states of Mexico. *IEEE Acc.* (2023)
13. Liao, H., Yang, S., Zavadskas, E.K., Škare, M.: An overview of fuzzy multi-criteria decision-making methods in hospitality and tourism industries: bibliometrics, methodologies, applications and future directions. *Economic Research-Ekonomska Istraživanja* **36**, 2150871 (2023)
14. Liao, P., Ye, F., Wu, X.: A comparison of the merchant and agency models in the hotel industry. *Int. Trans. Operat. Res.* **26**, 1052–1073 (2019). <https://doi.org/10.1111/itor.12365>
15. Onder, E., Kabadayi, N.: Supplier selection in hospitality industry using ANP. *Int. J. Acad. Res. Bus. Soc. Sci.* **5** (2015)
16. Park, S., Huang, Y.: Motivators and inhibitors in booking a hotel via smartphones. *Int. J. Contemp. Hosp. Manag.* **29**, 161–178 (2017)
17. Raab, C., Berezan, O., Christodoulidou, N., Jiang, L., Shoemaker, S.: Creating strategic relationships with online travel agents to drive hotel room revenue: an OTA perspective. *J. Hosp. Tourism Tech.* **9**, 125–140 (2018)
18. Rivera, G., Florencia, R., Guerrero, M., Porras, R., Sánchez-Solís, J.P.: Online multi-criteria portfolio analysis through compromise programming models built on the underlying principles of fuzzy outranking. *Inform. Sci.* **580**, 734–755 (2021)
19. Saaty, T.L., Vargas, L.G.: *The Analytic Network Process*, 1–40. Springer, USA (2013). https://doi.org/10.1007/978-1-4614-7279-7_1
20. Yee, B.Y., Faziharudean, T.M.: Factors affecting customer loyalty of using internet banking in Malaysia. *J. Elect. Banking Syst.* **21** (2010)

Chapter 6

A New Methodology Based on Multicriteria Ordinal Classification for the Management of Financial Resources with Application to Real Data from the Stock Market



Efrain Solares , Eduardo Fernández , Eyrán Díaz-Gurrola , Reimundo Moreno-Cepeda, Emmanuel Contreras-Medina , and Edy López Cervantes

Abstract Stock selection is highly complex due to the high heterogeneity of its factors. The determination of the value of a stock depends to a large extent on the investor's preferences towards such factors. This paper describes and evaluates a new methodology that uses investor decision policy to assign stocks to preferentially ordered classes. These classes can be of the “Don't Buy”, “Doubt” or “Buy” type. The classes are identified by limiting profiles at the boundary of each pair of consecutive classes and are given a priori by the investor. A back-testing strategy is used to evaluate the proposal and its results are compared with those of some benchmark approaches. The primary findings highlight that the stocks classified within the best class not only yielded better average returns compared to the broader market but

E. Solares (✉) · E. Fernández · E. Díaz-Gurrola · R. Moreno-Cepeda · E. Contreras-Medina
Research Center for Sustainable Development and Business Innovation, Universidad Autónoma de Coahuila, Torreón, Mexico
e-mail: efrain.solares@uadec.edu.mx

Universidad Autónoma de Coahuila, Saltillo, Mexico

E. Fernández
e-mail: eduardo.fernandez@uadec.edu.mx

E. Díaz-Gurrola
e-mail: eyran_diaz@uadec.edu.mx

R. Moreno-Cepeda
e-mail: reimundo.moreno@uadec.edu.mx

E. Contreras-Medina
e-mail: emmanuelmedina@uadec.edu.mx

E. López Cervantes
Facultad de Informática, Universidad Autónoma de Sinaloa, Sinaloa, Mexico
e-mail: edylopezc@gmail.com

also exhibited significantly lower volatility, suggesting a more favorable risk-reward balance and outperforming conventional methods and market benchmarks in terms of both returns and risk management.

Keywords Decision aiding · Outranking approach · Stock selection

6.1 Introduction

Many methodologies for stock selection have been presented in the related literature in recent decades. The interest of the scientific community in this field is due to the important repercussions that it entails, both in the academic ground and for society in general. Its practical and theoretical limitations have been addressed from various perspectives and theories, producing a considerably high number of factors involved in the selection process. However, determining the best way to aggregate the information provided by these factors is an open question.

The so-called multicriteria decision aiding (MCDA) [1] is an important branch of decision theory that can deal with the aggregation of information from multiple factors. In MCDA, a set of actions (decision alternatives) must be assessed based on a set of features called criteria to address choosing, ranking and sorting problems. Sorting problems, also called ordinal classification problems, consist of deciding how to assign actions to a collection of classes or categories that have been ordered (e.g., from worst to best) through the preferences of a decision maker (DM). In multicriteria ordinal classification, or multicriteria sorting, each action is assessed on multiple criteria and, aided by a decision model or decision policy, the actions are compared to reference profiles (reference actions) that characterize or define each class. Depending on the preferences within the decision model, the comparison allows the DM to assign the action to one or more classes.

Within the MCDA theory and in the context of multicriteria sorting, the most widely used method is ELECTRE TRI [2, 3], later called ELECTRE TRI-B by [3]. A recent generalization of ELECTRE TRI-B was introduced in [4]. This method, named INTERCLASS-nB, fulfills all the consistency properties imposed on multicriteria sorting methods and, like ELECTRE TRI-B, uses reference profiles at the limiting boundary of each pair of consecutive classes to perform the classification. Unlike its predecessor, INTERCLASS-nB can exploit more than one reference profile at each boundary. This is a major improvement on the original method, as it can give the sorting process more discerning power. In addition, the new method is versatile with respect to the elicitation of parameter values, since they can be represented by real or interval numbers, which reduces the cognitive effort of the DM to define these values and the time required to define the “best” parameter values. Here, we explore the abilities of INTERCLASS-nB to aggregate the information provided by all common factors and assign stocks to preferentially ordered classes that can lead to the final selection of the most convenient stocks.

In this paper, INTERCLASS-nB is used for first time to address the stock selection problem. Following [5] and, as described in Sect. 6.3, the methodology used to assess the proposal is defined as follows:

1. The set of decision alternatives is composed of the stocks in the S&P500 index [6].
2. The set of criteria is composed of the so-called fundamental factors [7] as well as price forecasting [8].
3. The collection and preparation of the input data is performed through capitaliq.com.
4. Expert investors were simulated to create a unified (although imprecise) decision policy that will provide the parameter values to exploit the multicriteria sorting method.
5. Assign the stocks (actions) to preferentially ordered classes.
6. Only the stocks that have been assigned to the best overall class are bought in a buy-and-hold strategy. A uniform distribution of resources is used to buy the selected stocks.
7. The created portfolio is assessed with a back-testing strategy.

This paper is structured as follows. Section 6.2 provides a review of the related literature and describes the INTERCLASS-nB method. Section 6.3 presents how this method is exploited to address the stock selection problem. The results are shown in Sect. 6.4. Section 6.5 concludes the paper.

6.2 Background

This section begins by mentioning and briefly describing the most outstanding works related to this paper; later, the details of the multicriteria sorting method used are provided.

6.2.1 *Review of the Related Literature*

Determining the most convenient stocks from a large universe of options is defined as stock selection. There are many factors involved in defining this convenience. Common factors in the related literature come from fundamental analysis [7].

Fundamental analysis uses data that is regularly published by the organizations underlying the stocks. This data is used to calculate indicators that investors often use to evaluate stocks. Indicators are both qualitative and quantitative, and typically include information that allows investors to compare the indicators (which represent the actual value of organizations) with current stock prices. Therefore, if one of these indicators shows evidence that the stock is undervalued, then it supports the decision that the investor should support the stock. If a sufficiently large number of indicators

provide such evidence, then the stock should be selected for investment. However, in practice, the decision is not straightforward since indicators do not usually provide evidence to reach the same conclusions.

We found in the literature review that the most used fundamental indicators used during stock selection are the following [9, 10].

- Profitability.
- Leverage.
- Liquidity.
- Efficiency.
- Growth.
- Solvency.
- Operational efficiency.

Several fundamental indicators were used in [7] for stock selection purposes. The indicators used in that work relied mainly on the price of the stocks and their relationship with the financial information published by the organizations underlying the stocks. The indicators used in that work are Debt to Equity, Price to Earnings, and Profit to Earn. Similarly, [11] used the Price to Earnings Ratio and New Loan to Market Capitalization ratios. As many as seventeen indicators were used in [10] for similar purposes.

The literature shows that fundamental indicators are commonly used in combination with other types of information during stock selection. For example, technical analysis is employed to complement fundamental information. In [12, 13], eight technical indicators were used together with eight fundamental indicators [14] combined fundamental indicators with technical indices. On the other hand, price forecasting is also a very common approach used to complement fundamental analysis [9] used an artificial neural network approach to forecast stock prices and combined them with twelve fundamental indicators.

Various stock selection methods have been employed, with notable ones including artificial neural networks, data envelopment analysis, evolutionary algorithms, sentiment analysis, and support vector machines [15]. In [16], data envelopment analysis is combined with multicriteria decision aiding theory for fund selection. Another study [17] introduces a novel three-stage network model in multiplier data envelopment analysis. In a different approach [18], a hybrid model between a feed-forward neural network and an adaptive neural fuzzy inference system is proposed. Additionally, a study [9] suggests using differential evolution to optimize an objective function based on historical prices, which in turn weights a set of indicators derived from fundamental analysis. Lastly, support vector machines are utilized in two studies [19, 20].

6.2.2 The INTERCLASS-nB Method

INTERCLASS-nB is a multicriteria sorting method that generalizes the most used sorting method within the MCDA theory, the so-called ELECTRE TRI-B. The main characteristics of INTERCLASS-nB are i) its ability to assign actions to ordered classes through multiple profiles in the limiting boundary of each pair of consecutive classes (as opposed to ELECTRE TRI-B, that can use only one), and ii) its capacity to cope with imprecision in the definition of its parameter values (through the so-called interval numbers).

Imprecise, vague, or ill-defined values can be assigned to the parameters of INTERCLASS-nB to reflect situations where the decision maker does not want/ is not willing to engage in an arduous process to define these parameters, and/or when the information to define the impacts on the criteria (factors) is not precisely known. This is accomplished in INTERCLASS-nB through the concept of interval number. An interval number can be defined as a variety of values (in the form of interval) that an unknown quantity can achieve. Formally, let's consider an amount r whose value is uncertain. The possible range of this value can be defined by its highest attainable value r^+ and its lowest attainable value r^- . This allows us to capture the variability and potential volatility of r within a bounded range, providing a clear framework to analyze its fluctuations between r^+ and r^- , respectively, $r = [r^-, r^+]$ is called interval number of r . We use bold font to identify interval numbers.

$A = \{a_1, a_2, a_3, \dots\}$ denotes the set of decision actions, while the collection of classes is denoted by C_k , $k = 1, \dots, M$, with C_{k+1} being preferred to C_k . To differentiate the boundary between each pair of classes, a set $B_k = \{b_{k,j}; j = 1, \dots, \text{card}(B_k)\}$ of limiting actions (profiles) $b_{k,j}$ is used, where $\{B_0, B_1, \dots, B_M, B_{M+1}\}$ is the set of all the limiting profiles (B_0 , and B_{M+1} are composed of the anti-ideal and ideal actions, respectively).

Consider δ and β established following the preferences of the decision maker. Given the assertion " $x \in A$ dominates $y \in A$ ", we denote its credibility as $xD(\alpha)y$. The assertion " x dominates y " is accepted when $xD(\alpha)y \geq \delta$. Similarly, the assertion " x is at least as good as y " is denoted by $\eta(x, y)$. The assertion " x is at least as good as y " is accepted when $\eta(x, y) \geq \beta$, and is denoted as $xS(\beta)y$. Furthermore, if $xS(\beta)y$ is hold but $yS(\beta)x$ does not, then it is said that " x is preferred to y ", which is denoted by $xPr(\delta, \lambda)y$. The specific steps to calculate $xD(\alpha)y$ and $\eta(x, y)$ are described in [4]. INTERCLASS-nB also exploits the following conditions to sort the actions in A into classes.

The following conditions are imposed to INTERCLASS-nB for the method to fulfill consistency properties [4]:

1. C_k is defined through a set of reference upper limiting profiles, B_k , and through a set of reference lower limiting profiles, B_{k-1} . It is assumed that all $b_{k,j}$ of B_k are in C_{k+1} .
2. B_0 (respectively, B_M) is composed of the anti-ideal (resp. the ideal) action.
3. For all k , there is no pair $(b_{k,j}, b_{k,i})$ such that $b_{k,j} Pr(\delta, \lambda)b_{k,i}$.
4. For all (k, h) ($h > k$), there is no pair $(b_{k,j}, b_{h,i})$ such that $b_{k,j} Pr(\delta, \lambda)b_{h,i}$.

5. For all k and for each action w in B_k , there is at least one action z in B_{k+1} such that $zD(\alpha)w$ (with $\alpha \geq \delta$).
6. For all k and for each action w in B_{k+1} , there is at least one action z in B_k such that $wD(\alpha)z$ (with $\alpha \geq \delta$).

Preferential relations can also be established between an individual action x and a set of profiles B_k as follows:

$xS(\delta, \lambda)B_k \Leftrightarrow xS(\delta, \lambda)w$ for any $w \in B_k$ and there is no $z \in B_k$ with $zPr(\delta, \lambda)x$.

$B_kS(\delta, \lambda)x \Leftrightarrow wS(\delta, \lambda)x$ for any $w \in B_k$ and there is no $z \in B_k$ with $xPr(\delta, \lambda)z$.

The assignment of alternatives to classes in INTERCLASS-nB is carried out through two interconnected methods: the pseudo-conjunctive and pseudo-disjunctive procedures. These procedures rely on the sets of profiles meeting specific conditions, as outlined previously. Here's a closer look at how each procedure functions:

- **Pseudo-Conjunctive Procedure:** This approach aims to ensure that an alternative satisfies a certain set of criteria, similar to a conjunctive evaluation. Only if the alternative meets the necessary conditions from the profiles will it be assigned to a corresponding class.
- **Pseudo-Disjunctive Procedure:** In contrast, this method allows for more flexibility by assigning an alternative to a class if it meets any one of the specified conditions from the profiles, resembling a disjunctive evaluation.

Both procedures work together to provide a comprehensive classification framework, enabling accurate and balanced assignments to the appropriate classes.

Pseudo-conjunctive procedure

1. Compare x to B_k for $k = M - 1, \dots, 0$, until the first value, k , such that $xS(\delta, \lambda)B_k$;
2. Assign x to class C_{k+1} .

Pseudo-disjunctive procedure

1. Compare x to B_k for $k = 1, \dots, M$, until the first value, k , such that $B_kPr(\delta, \lambda)x$;
2. Assign x to class C_k .

For more details of this method, such as general steps, the step-by-step algorithm, an optimization-based inference method for the INTERCLASS-nB, and other considerations, the reader is referred to [4].

6.3 Proposed Methodology

The proposed methodology exploits the INTERCLASS-nB method to assign each element in a set of actions to preferentially ordered classes; then, creating a portfolio with the stocks that were assigned to the best class and supporting each of these with an equal amount of resources; finally, using a back-testing strategy to compare the results of the proposal with those of the benchmark approaches.

This methodology is shown in Fig. 6.1 and described in the rest of this section.

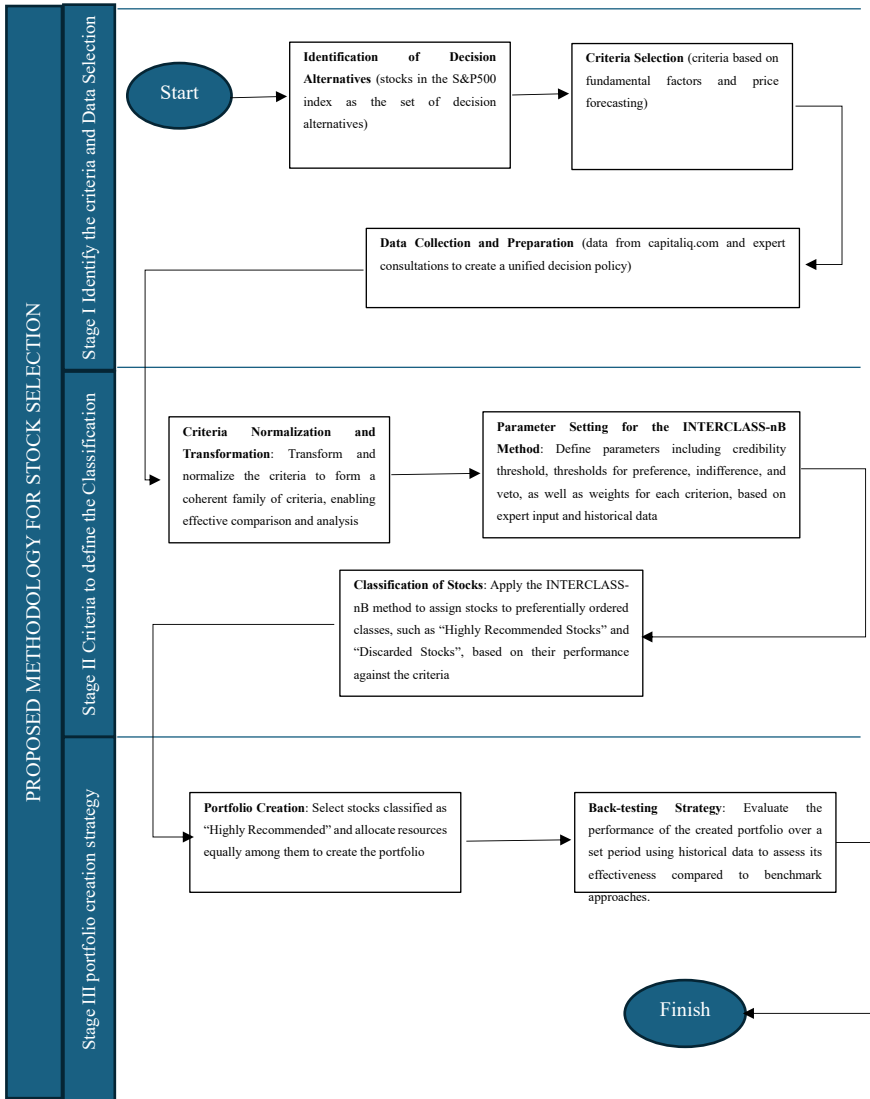


Fig. 6.1 Proposed methodology

6.3.1 *Input Data*

The Standard and Poor's 500 (S&P500) index is perhaps the most important stock benchmark worldwide, as it contains five hundred of the most representative companies of the United States of America. To perform the back-testing strategy, we used daily data for the last ninety business days. We use historical data provided by capitaliq.com about the stocks in the S&P500 index consisting of daily prices from March 5 to July 14, 2021 (ninety periods). Such a platform allows us to access financial statements and ratios preparing the input data particularly by discarding missing information (as opposed to fulfilling it with artificial information). Sixty periods are used to "train" the model. Then, the multicriteria sorting method assigns the stocks in the period immediately after the sixty training periods to finally select the stocks of the best class. This is done in the form of a sliding window, so that the selection of stocks is performed in thirty different periods that represent different contexts and trends of the market (assessing robustness of the proposal). Capitaliq.com is particularly used for downloading and preparing the data, while the rest of the procedure is performed using basic tabular tools such as Excel.

Table 6.1 provides a sample of the organizations in the S&P500.

Among the different options of fundamental factors used in the literature, given the complementarity of the information they provide, here we use the following factors based on the recommendations provided by [5, 9, 10, 13, 18]:

1. Return on asset (RoA): indicates the profitability of the organization regarding its total assets. Provides an insight about the efficiency of the organization exploiting its assets to generate profits. It is calculated by dividing its net income by its total assets.
2. Price to Earnings (P/E): measures the stock price of the organization regarding its earnings per share (that is, the profit of the organization divided by the outstanding shares of its common stock). Provides an insight about the relative value of the shares of the organization.
3. Price to Book (P/B): indicates the price per share in the market regarding its book value. Provides an insight considering if the value of the organization is priced properly by the market.
4. Price to Sales (P/S), indicates the price per share in the market regarding the revenue per share. Provides an insight about the stock price of an organization with respect to its revenues; that is, denotes the value that the market has puts on each dollar the organization has sold.
5. Return on equity (RoE), indicates the efficiency of the organization in generating profits. It can be calculated as the net income of the organization divided by its average shareholder's equity.

Following the trend in the literature to combine multiple types of factors, in addition to these fundamental indicators that provide a wide range of perspectives regarding the financial situation of the organizations underlying the stocks, we use stock price forecasts based on more traditional estimations. We utilize an estimated

Table 6.1 Organizations in the S&P500 index

Constituents name	Constituents	Constituents name	Constituents
3M Company	NYSE:MMM
A. O. Smith Corporation	NYSE:AOS	Valero Energy Corporation	NYSE:VLO
Abbott Laboratories	NYSE:ABT	Ventas Inc.	NYSE:VTR
AbbVie Inc.	NYSE:ABBV	VeriSign Inc.	NasdaqGS:VRSN
Abiomed Inc.	NasdaqGS:ABMD	Verisk Analytics Inc.	NasdaqGS:VRSK
Accenture plc	NYSE:ACN	Verizon Communications Inc.	NYSE:VZ
Activision Blizzard Inc.	NasdaqGS:ATVI	Vertex Pharmaceuticals Incorporated	NasdaqGS:VRTx
Adobe Inc.	NasdaqGS:ADBE	ViacomCBS Inc.	NasdaqGS:VIAC
Advance Auto Parts Inc.	NYSE:AAP	Viatis Inc.	NasdaqGS:VTRS
Advanced Micro Devices Inc.	NasdaqGS:AMD	Visa Inc.	NYSE:V
Aflac Incorporated	NYSE:AFL	Vornado Realty Trust	NYSE:VNO
Agilent Technologies Inc.	NYSE:A	Vulcan Materials Company	NYSE:VMC
Air Products and Chemicals Inc.	NYSE:APD	W. R. Berkley Corporation	NYSE:WRB
Akamai Technologies Inc.	NasdaqGS:AKAM	W.W. Grainger Inc.	NYSE:GWW
Alaska Air Group Inc.	NYSE:ALK	Walgreens Boots Alliance Inc	NasdaqGS:WBA
Albemarle Corporation	NYSE:ALB	Walmart Inc.	NYSE:WMT
Alexandria Real Estate Equities Inc.	NYSE:ARE	Waste Management Inc.	NYSE:WM
Alexion Pharmaceuticals Inc.	NasdaqGS:ALON	Waters Corporation	NYSE:WAT
Align Technology Inc.	NasdaqGS:ALGN	WEC Energy Group Inc.	NYSE:WEC
Allegion plc	NYSE:ALLE	Wells Fargo & Company	NYSE:WFC
Alliant Energy Corporation	NasdaqGS:LNT	Welltower Inc	NYSE:WELL
Alphabet Inc.	NasdaqGS:GOOG	West Pharmaceutical Services Inc.	NYSE:WST
Alphabet Inc.	NasdaqGS:GOOG.L	Western Digital Corporation	NasdaqGS:WDC
Altria Group Inc.	NYSE:MO	Westinghouse Air Brake Technologies Corporation	NYSE:WAB
Amazon.com Inc.	NasdaqGS:AMZN	WestRock Company	NYSE:WRK
Amcor plc	NYSE:AMCR	Weyerhaeuser Company	NYSE:WY

(continued)

Table 6.1 (continued)

Constituents name	Constituents	Constituents name	Constituents
Ameren Corporation	NYSE:AEE	Whirlpool Corporation	NYSE:WHR
American Airlines Group Inc.	NasdaqGS:AAL	Willis Towers Watson Public Limited Company	NasdaqGS:WLTW
American Electric Power Company Inc.	NasdaqGS:AEP	Wynn Resorts Limited	NasdaqGS:WYNN
American Express Company	NYSE:AxP	Xcel Energy Inc.	NasdaqGS:XEL
...	...	Xilinx Inc	NasdaqGS:XLNX

future return for each stock, calculated as the mean return over the most recent sixty periods, plus or minus three times the standard deviation of those returns. This approach captures the expected return along with its variability, providing a range that reflects potential fluctuations based on historical data. Here, we exploit the ability of INTERCLASS-nB to deal with parameters defined as interval numbers. This is particularly useful in this situation given the high complexity usually involved in forecasting the stock returns.

Some of the factors mentioned above reflect information in various contexts. Following [1], we carry out an assessment of the factors to ensure that they can form a so-called coherent family of criteria. This way, the factors form a set of criteria that fulfill the properties of non-redundancy, completeness, and consistency. First, we confirmed that no criterion duplicates the function of another within the decision-making process. Each criterion should measure a unique aspect of the stocks, preventing any undue influence from overlapping measures. We achieved it by using factors that measure the quality of the stocks in different perspectives: forecasted returns, profitability, and company performance. Second, we ensured that the set of criteria covers all relevant aspects of the decision problem; in the context of this work, completeness refer to two basic aspects, i.e., fundamental analysis and forecasting factors. Finally, criteria are independent, meaning the evaluation of a stock under one criterion does not affect its evaluation under another. These criteria are normalized to $[0,1]$ (except for the forecasted return) considering the last sixty historical periods of stock prices. Table 6.2 shows a sample for the most recent period considered in the experiments, July 14, 2021, which is part of the input used to assess the methodology.

Table 6.2 Sample of the performance matrix for July 14, 2021

Constituents name	Constituents	RoA	PtE	PtB	PtS	RoE
3M Company	NYSE:MMM	0.015	0.514	0.470	0.796	0.155
A. O. Smith Corporation	NYSE:AOS	0.668	0.393	0.851	0.746	0.730
Abbott Laboratories	NYSE:ABT	0.787	0.212	0.534	0.401	0.808
AbbVie Inc.	NYSE:ABBV	0.000	0.736	0.684	0.943	0.000
Abiomed Inc.	NasdaqGS:ABMD	0.013	0.374	0.463	0.482	0.000
Accenture plc	NYSE:ACN	0.170	1.000	1.000	1.000	0.194
Activision Blizzard Inc.	NasdaqGS:ATVI	0.102	0.259	0.338	0.244	0.000
Adobe Inc.	NasdaqGS:ADBE	0.973	1.000	1.000	1.000	0.017
Advance Auto Parts Inc.	NYSE:AAP	0.864	0.235	0.919	0.625	0.928
Advanced Micro Devices Inc.	NasdaqGS:AMD	0.763	0.626	0.739	0.615	0.264
Aflac Incorporated	NYSE:AFL	0.730	0.115	0.549	0.271	0.784
Agilent Technologies Inc.	NYSE:A	0.535	0.619	0.958	0.949	0.894
Air Products and Chemicals Inc.	NYSE:APD	0.024	0.579	0.538	0.425	0.000
Akamai Technologies Inc.	NasdaqGS:AKAM	0.036	0.823	0.857	0.850	0.000
Alaska Air Group Inc.	NYSE:ALK	0.197	0.000	0.000	0.352	0.000
...
Waste Management Inc.	NYSE:WM	0.037	0.998	0.998	0.998	0.051
Waters Corporation	NYSE:WAT	0.073	1.000	1.000	1.000	0.000
WEC Energy Group Inc.	NYSE:WEC	0.090	0.282	0.331	0.231	0.000
Wells Fargo & Company	NYSE:WFC	1.000	0.000	1.000	0.000	1.000
Welltower Inc.	NYSE:WELL	0.068	1.000	1.000	1.000	0.000
West Pharmaceutical Services Inc.	NYSE:WST	0.799	0.696	0.961	0.951	0.904
Western Digital Corporation	NasdaqGS:WDC	0.786	0.065	0.440	0.489	0.650
Westinghouse Air Brake Technologies Corporation	NYSE:WAB	0.071	0.719	0.741	0.823	0.000
WestRock Company	NYSE:WRK	0.060	0.000	0.222	0.234	0.000
Weyerhaeuser Company	NYSE:WY	0.855	0.039	0.143	0.124	0.879
Whirlpool Corporation	NYSE:WHR	0.712	0.123	0.201	0.277	0.926
Willis Towers Watson Public Limited Company	NasdaqGS:WLTW	0.092	0.000	0.000	0.003	0.756
Wynn Resorts Limited	NasdaqGS:WYNN	0.874	0.000	0.816	0.120	0.000
Xcel Energy Inc.	NasdaqGS:XEL	0.095	0.380	0.564	0.262	0.000
Xilinx Inc.	NasdaqGS:XLNX	0.013	0.604	0.536	0.604	0.418

6.3.2 Parameter Values

Here, we describe the parameter values used to operationalize the proposed methodology. Please note that expert knowledge from the investors is simulated by determining parameter values from (1) the historical performances of the stocks, (2) preliminary experiments, and (3) the literature.

Given the outstanding feature of INTERCLASS-nB to handle imprecise information in the form of interval numbers, the cognitive effort imposed to the decision maker during the elicitation of parameter values is considerably reduced. Therefore, we assume that the decision maker is able and willing to provide such parameter values and that he/she can do it in a straightforward manner.

6.3.2.1 Representing the Preference Model of the Decision Maker

According to the discussion in Sect. 2.1 and the specific steps described in [4], INTERCLASS-nB requires the following parameters to be defined in order to reproduce the preference model of the decision maker:

- A credibility threshold, δ , to establish clear preference relations.
- A threshold, λ , to define a strong majority in the outranking relation.

n criteria weights, w_i ($i = 1, 2, \dots, n$), to denote the relative importance of the criteria.

n indifference thresholds, q_i , ($i = 1, 2, \dots, n$) to denote the maximum differences between criteria impacts such that the decision maker feels that the impacts are indifferent.

n preference thresholds, p_i , ($i = 1, 2, \dots, n$) to denote the minimum differences between criteria impacts such that the decision maker can discern which of the impact is preferred.

n veto thresholds, v_i , ($i = 1, 2, \dots, n$) to denote the minimum differences between the impacts of actions x and y on the i th criterion, say $g_i(x)$ and $g_i(y)$, such that the decision maker can veto the assertion “ x is at least as good as y according to the i th criterion” when $g_i(y) - g_i(x) \geq v_i$.

To evaluate the proposal, the following parameter values are utilized. The credibility threshold for establishing clear preference relations, δ , is set at 0.51. The threshold λ for defining a strong majority in the outranking relation is set between 0.51 and 0.66. All criteria are assumed to have equal weight, with the threshold values reflecting that the criteria impacts are normalized to the range [0, 1]. The indifference thresholds are defined as [0, 0.1], the preference thresholds as [0.1, 0.2], and the veto thresholds as [0.5, 0.7]. As stated above, these values were set given the results from the historical performances of the stocks, some preliminary experiments, and the literature. Particularly, we have noticed from previous unstructured experiments on historical data that low values of δ and λ helped to obtain greater returns in the long term, while the (indifference, preference, and veto) thresholds are supported by the

Table 6.3 Profiles used to represent the limiting boundaries between classes

Boundary	Return on asset	Price to earnings	Price to book	Price to sales	Return on equity	Forecasted return
B_0	0	0	0	0	0	- 0.08
B_1	0.6	0.6	0.6	0.6	0.6	[0.2, 0.4]
B_2	1	1	1	1	1	0.10

works [13, 21, 22]. Evidently, more structured experiments should be carried out to improve the performance of the proposal. However, this is out of the scope of this work, so the authors will address the topic as a future research line.

6.3.2.2 Representing Limiting Boundaries Between Classes Through Reference Profiles

Two classes were established for stock selection: the Discarded class ($C1$) and the Highly Recommended Stocks class ($C2$). The defining profiles at the boundaries of these classes are outlined in Table 6.3. The profile values were intuitively assigned given the ranges of criteria scores ([0,1] for the fundamental indicators and around 0 for the average forecasted returns); particularly, such values denote convenience points characterizing the limiting boundaries between the classes.

6.4 Results and Discussion

As mentioned in Sect. 6.3, we consider daily data (for business days) of stocks in the S&P500 index for the time frame from March 5 to July 14, 2021 (ninety periods). In each of these periods, the fundamental indicators were derived from data for the immediately preceding quarter of the year, while the last sixty daily historical prices were used to forecast the stock's return; later, this information was used to evaluate the stock and assign it to one of the classes. To carry out the buy and hold strategy (with a portfolio with equally distributed weights), only the stocks allocated to the highest class were selected. Below we discuss first the allocations made by the proposal and then the returns the portfolios achieved.

6.4.1 Results of the Multicriteria Sorting Method

The work of the multicriteria sorting method consists of processing the input data to evaluate each stock on the multiple criteria and assign it to a preferential class. Since we are only interested in the most outstanding stocks, we only have two classes, the

Highly Recommended Stocks and the Discarded Stocks classes. Table 6.4 provides a sample of the assignments made by the method. For the period March 5, 2021. The complete set of results can be found in this *link*.

6.4.2 Returns of the Recommended Portfolios

Table 6.5 presents a summary of the returns produced by the selected stocks (as classified by the INTERCLASS-nB method) for each of the thirty historical periods under consideration. The periods correspond to specific dates, and for each period, the performance of the recommended stocks is compared against a benchmark index. The table is structured to provide insights into the effectiveness of the stock selection methodology by showing the financial outcomes of following the proposed strategy versus market performance as represented by the benchmark.

The entries under both “Recommended Stocks” and “Benchmark Index” columns are percentage changes, reflecting the variation in investment value from the beginning to the end of each period. Positive values indicate a gain, while negative values signify a loss. The comparison between these two columns is crucial for assessing the proposed stock selection methodology’s relative performance. It demonstrates whether the method consistently outperforms, matches, or underperforms against the market benchmark, providing a measure of its effectiveness and potential value to investors.

Given the large number of stocks in the S&P500 index, the third column of Table 6.5 tends to show larger numbers than the second one. However, it can still be clearly appreciated that the selected stocks provide better overall returns. The mean return for the stocks in the index after the thirty periods is -6.21% , while that of the selected stocks is -0.28% .

Another interesting result is that the summatory of the returns produced by the recommended stocks not necessarily follows the tendency (direction) of the market, which may be helpful in periods of general losses for the market. It is important to remark that the market presents high volatility. The concept of volatility (measured, for example, by the standard deviation) is a crucial indicator of the risk that the investor would be taking if he/she participates in a given portfolio of investment. The standard deviation of the summatory of returns of the stocks in the S&P500 index is 373.80% , while that of the recommended stocks is 1.33% .

6.5 Conclusions

A novel way of selecting stocks through a multicriteria sorting method has been described and assessed. The main component of the proposed methodology is the so-called INTERCLASS-nB method. This method exploits the available information about the preferences of the decision maker to assign actions (the stocks) to predefined

Table 6.4 Sample of the assignments of stocks to ordered classes made by INTERCLASS-nB

Constituent	Class	Constituent	Class	Constituent	Class	Constituent	Class	Constituent	Class	Constituent	Class
NYSE:MMM	C1	NasdaqGS:AEP	C1	NYSE:AVY	C1	NYSE:CARR	C2	NasdaqGS:CMCS.A	C1		
NYSE:AOS	C1	NYSE:AxP	C1	NYSE:BKR	C1	NYSE:CTLT	C1	NYSE:CMA	C2		
NYSE:ABT	C1	NYSE:AIG	C1	NYSE:BLL	C1	NYSE:CAT	C1	NYSE:CAG	C1		
NYSE:ABV	C1	NYSE:AMT	C2	NYSE:BAC	C1	BATS:CBOE	C1	NYSE:COP	C1		
NasdaqGS:ABMD	C1	NYSE:AWK	C1	NYSE:BAx	C1	NYSE:CBRE	C1	NYSE:ED	C1		
NYSE:ACN	C1	NYSE:AMP	C1	NYSE:BDx	C1	NasdaqGS:CDW	C1	NYSE:STZ	C1		
NasdaqGS:ATVI	C1	NYSE:ABC	C1	NYSE:BRK.B	C1	NYSE:CE	C1	NasdaqGS:CPRT	C1		
NasdaqGS:ADBE	C1	NYSE:AME	C1	NYSE:BBY	C2	NYSE:CNC	C1	NYSE:GLW	C2		
NYSE:AAP	C1	NasdaqGS:AMGN	C1	NYSE:BIO	C1	NYSE:CNP	C1	NYSE:CTVA	C1		
NasdaqGS:AMD	C1	NYSE:APH	C1	NasdaqGS:BIIB	C1	NasdaqGS:CERN	C1	NasdaqGS:COST	C2		
NYSE:AFL	C1	NasdaqGS:ADI	C2	NYSE:BLK	C1	NYSE:CF	C1	NYSE:CCI	C1		
NYSE:A	C1	NasdaqGS:ANSS	C1	NasdaqGS:BKNG	C1	NYSE:CRL	C1	NasdaqGS:CSx	C1		
NYSE:APD	C1	NYSE:ANTM	C1	NYSE:BWA	C1	NasdaqGS:CHTR	C1	NYSE:CMI	C1		
NasdaqGS:AKAM	C1	NYSE:AON	C1	NYSE:BxP	C1	NYSE:CVx	C1	NYSE:CVS	C1		
NYSE:ALK	C1	NasdaqGS:APA	C1	NYSE:BSx	C1	NYSE:CMG	C1	NYSE:DHI	C1		
NYSE:ALB	C1	NasdaqGS:AAPL	C1	NYSE:BMY	C1	NYSE:CB	C2	NYSE:DHR	C1		
NYSE:ARE	C1	NasdaqGS:AMAT	C1	NasdaqGS:AVGO	C1	NYSE:CHD	C1	NYSE:DRI	C1		
NasdaqGS:ALxN	C1	NYSE:APTV	C1	NYSE:BR	C1	NYSE:CI	C1	NYSE:DVA	C1		
NasdaqGS:ALGN	C1	NYSE:ADM	C1	NYSE:BF.B	C1	NasdaqGS:CINF	C2	NYSE:DE	C1		
NYSE:ALLE	C2	NYSE:ANET	C1	NasdaqGS:CHRW	C1	NasdaqGS:CTAS	C1	NYSE:DAL	C1		
NasdaqGS:LNT	C1	NYSE:AJG	C1	NYSE:COG	C1	NasdaqGS:CSCO	C1	NasdaqGS:xRAY	C1		
NasdaqGS:GOOG	C1	NYSE:AIZ	C1	NasdaqGS:CDNS	C1	NYSE:C	C1	NYSE:DVN	C1		

(continued)

Table 6.5 Summary of the returns produced by the selected stocks in each of the thirty historical periods

Period	Recommended stocks (%)	Benchmark index (%)
5/3/2021	- 1.60	52.83
8/3/2021	0.07	- 98.19
9/3/2021	0.47	251.15
10/3/2021	- 1.13	- 93.48
11/3/2021	0.18	86.98
12/3/2021	- 2.40	- 229.60
15/3/2021	- 0.95	48.69
16/3/2021	1.33	178.53
17/3/2021	- 1.26	- 208.36
18/3/2021	0.48	16.75
19/3/2021	- 0.97	- 340.13
22/3/2021	- 3.58	- 507.96
23/3/2021	- 3.06	- 833.29
24/3/2021	1.96	971.61
25/3/2021	0.59	69.01
26/3/2021	0.34	- 69.39
29/3/2021	0.75	322.24
30/3/2021	1.35	364.07
31/3/2021	- 1.09	- 219.71
1/4/2021	0.15	- 55.93
5/4/2021	0.74	125.37
6/4/2021	0.94	329.89
7/4/2021	0.63	114.34
8/4/2021	- 0.63	- 436.99
9/4/2021	0.23	103.36
12/4/2021	- 1.71	- 529.83
13/4/2021	1.23	811.70
14/4/2021	0.21	134.79
15/4/2021	- 1.27	- 467.06
16/4/2021	- 0.26	- 77.65

and preferentially ordered classes. The main characteristics of this method are that (1) it uses reference profiles at the limiting boundary between each pair of consecutive classes, (2) it fulfills all the consistency properties imposed on multicriteria sorting methods, (3) it is capable of incorporating imprecise data, vague or poorly defined preference information, so that obtaining the values of its parameters is relatively

easy for the decision maker, (4) it can handle the effects of non-compensation and veto during the decision process.

This work proposes to use the INTERCLASS-nB method to determine outstanding stocks based on a set of factors taken from the fundamental analysis (an approach that is widely used by practitioners) and the forecast of stock prices. According to the literature review, combining different types of information that describe the quality of the stocks is a common practice. Furthermore, we notice a strong tendency to incorporate information from these types of analyses since they tend to make it easier to find undervalued stocks (and, therefore, with high probability of increasing their price). However, considering multiple factors usually increases the complexity of the problem, so addressing this problem is not straightforward. The characteristics of the INTERCLASS-nB method allowed us to cope with the complexity of the problem.

We assessed the proposal by simulating long-position investments with actual historical data (that is, a back-testing strategy) in stocks within the Standard and Poor's 500 (S&P500) index. To perform the back-testing strategy, we used daily data for the last ninety business days. First, sixty periods are used for "training", then the INTERCLASS-nB method assigns the stocks to the classes, so we can select the best classified stocks. This is performed in a sliding-window manner such that the selection of stocks is performed in thirty different periods representing different contexts and trends of the market, thus assessing robustness of the proposal. The results of the experiments showed that around ten percent of the stocks were assigned to the best category of stocks in each of the test periods. This is an important result because the stock selection phase (one of the stages of the overall management of stock portfolios) requires identifying a limited number of the best stocks. The actual return of the selected stocks is then compared to that of the stocks in the S&P500 index. Table 6.5 shows that the proposal outperformed the market in most scenarios in the context of summary of returns and with respect to volatility and cumulative return. Therefore, we conclude that the proposed approach is adequate to be considered by practitioners.

As can be seen in Table 6.4, INTERCLASS-nB only assigned around ten percent of the stocks to the best category (C_2), which is mainly due to the high demand imposed through the profiles in the boundaries between pairs of classes. In general terms, a portfolio with fewer shares is more convenient to exercise better control and achieve lower levels of commissions.

In future works, more experiments will be carried out to investigate the impact that such a requirement produces on the number of supported stocks. In any case, the decision maker can always provide the requirements and number of profiles with which he/she feels most comfortable.

The proposal should be further assessed using other preference models to represent the different behaviors of decision makers. Particularly, different number of profiles per limiting boundary, different sets of fundamental factors, other ways of forecasting stock prices, and diverse factors coming from other types of sources (such as technical and sentimental analyses).

References

1. Roy, B.: *Multicriteria Methodology for Decision Aiding*. Kluwer Academic Publishers, The Netherlands (1996)
2. Roy, B., Bouyssou, D.: *Aide multicritère à la décision: méthodes et cas*. Economica Paris (1993)
3. Almeida-Dias, J., Figueira, J.R., Roy, B.: Electre Tri-C: a multiple criteria sorting method based on characteristic reference actions. *Eur. J. Oper. Res.* **204**, 565–580 (2010). <https://doi.org/10.1016/j.ejor.2009.10.018>
4. Fernández, E., Figueira, J.R., Navarro, J.: Interval-based extensions of two outranking methods for multi-criteria ordinal classification. *Omega (Westport)*. **95**, 102065 (2020). <https://doi.org/10.1016/j.omega.2019.05.001>
5. Solares, E., de-León-Gómez, V., Fernández, E., Contreras-Medina, E., Lopez, O.: Multicriteria ordinal classification to improve strategic planning in the financial sector of the company. *Int. J. Combinatorial Optim. Probl. Inf.* **13** (2022)
6. Frino, A., Gallagher, D.R.: Tracking S&P 500 index funds. *J. Portfolio Manage.* **28**, 44–55 (2001)
7. Xidonas, P., Mavrotas, G., Psarras, J.: A multicriteria methodology for equity selection using financial analysis. *Comput. Oper. Res.* **36**, 3187–3203 (2009). <https://doi.org/10.1016/j.cor.2009.02.009>
8. Solares, E., Salas, F.G., De-Leon-Gomez, V., Díaz, R.: A comprehensive soft computing-based approach to portfolio management by discarding undesirable stocks. *IEEE Access*. **10**, 40467–40481 (2022). <https://doi.org/10.1109/ACCESS.2022.3167153>
9. Yang, F., Chen, Z., Li, J., Tang, L.: A novel hybrid stock selection method with stock prediction. *Appl. Soft Comput.* **80**, 820–831 (2019). <https://doi.org/10.1016/j.asoc.2019.03.028>
10. Shen, K.-Y., Tzeng, G.-H.: Combined soft computing model for value stock selection based on fundamental analysis. *Appl. Soft Comput.* **37**, 142–155 (2015). <https://doi.org/10.1016/j.asoc.2015.07.030>
11. Chai, J., Du, J., Lai, K.K., Lee, Y.P.: A hybrid least square support vector machine model with parameters optimization for stock forecasting. *Math. Probl. Eng.* **2015**, (2015). <https://doi.org/10.1155/2015/231394>
12. Fernandez, E., Navarro, J., Solares, E., Coello, C.C.: A novel approach to select the best portfolio considering the preferences of the decision maker. *Swarm Evol. Comput.* **46**, 140–153 (2019). <https://doi.org/10.1016/j.swevo.2019.02.002>
13. Fernandez, E., Navarro, J., Solares, E., Coello, C.C.: Using evolutionary computation to infer the decision maker's preference model in presence of imperfect knowledge: a case study in portfolio optimization. *Swarm Evol. Comput.* **54**, 100648 (2020). <https://doi.org/10.1016/j.swevo.2020.100648>
14. Zarandi, M.H.F., Rezaee, B., Turksen, I.B., Neshat, E.: A type-2 fuzzy rule-based expert system model for stock price analysis. *Expert Syst. Appl.* **36**, 139–154 (2009). <https://doi.org/10.1016/j.eswa.2007.09.034>
15. Andriopoulos, D., Doumpos, M., Pardalos, P.M., Zopounidis, C.: Computational approaches and data analytics in financial services: a literature review. *J. Oper. Res. Soc.* **70**, 1581–1599 (2019). <https://doi.org/10.1080/01605682.2019.1595193>
16. do Castelo Gouveia, M., Duarte Neves, E., Cândido Dias, L., Henggeler Antunes, C.: Performance evaluation of Portuguese mutual fund portfolios using the value-based DEA method. *J. Oper. Res. Soc.* **69**, 1628–1639 (2018). <https://doi.org/10.1057/s41274-017-0259-7>
17. Galagedera, D.U.A., Roshdi, I., Fukuyama, H., Zhu, J.: A new network DEA model for mutual fund performance appraisal: an application to US equity mutual funds. *Omega (Westport)* **77**, 168–179 (2018). <https://doi.org/10.1016/j.omega.2017.06.006>
18. Huang, Y., Capretz, L.F., Ho, D.: Neural network models for stock selection based on fundamental analysis. In: 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), pp. 1–4 (2019). <https://doi.org/10.1109/CCECE.2019.8861550>

19. Huang, H., Wei, X., Zhou, Y.: A sparse method for least squares twin support vector regression. *Neurocomputing* **211**, 150–158 (2016). <https://doi.org/10.1016/j.neucom.2015.12.133>
20. Huang, H., Wei, X., Zhou, Y.: Twin support vector machines: a survey. *Neurocomputing* **300**, 34–43 (2018). <https://doi.org/10.1016/j.neucom.2018.01.093>
21. Xidonas, P., Askounis, D., Psarras, J.: Common stock portfolio selection: a multiple criteria decision making methodology and an application to the Athens Stock Exchange. *Oper. Res. Int. J.* **9**, 55–79 (2009). <https://doi.org/10.1007/s12351-008-0027-1>
22. Xidonas, P., Mavrotas, G., Krintas, T., Psarras, J., Zopounidis, C., Xidonas, P., Mavrotas, G., Krintas, T., Psarras, J., Zopounidis, C.: Stock selection. *Multicrit. Portfolio Manag.* 23–55 (2012) https://doi.org/10.1007/978-1-4614-3670-6_2

Part II
Intelligent Optimization

Chapter 7

Warm Starting Integer Programming for the Internet SHopping Optimization Problem with Multiple Item Units (ISHOP-U)



Fernando Ornelas, Alejandro Santiago , José Antonio Castan Rocha , Salvador Ibarra Martínez , and Alejandro H. García 

Abstract In this discrete optimization work, we deal with Internet purchases that have become very popular and increased yearly. In this chapter, we study the Internet SHopping Optimization Problem with multiple item Units (ISHOP-U), a combinatorial NP-hard variant of the original ISHOP that considers purchasing one or more units of a product in a set of products. We performed a warm start in an Integer programming model in CPLEX using as a starting point the best-found solutions from two evolutionary algorithms: A Cellular Genetic Algorithm (CGA), A Genetic Algorithm (GA), and a nature-inspired Water Cycle Algorithm (WCA) present in the state-of-the-art. The test setup is on 15 synthetic instances in the literature, where unit costs and delivery prices adhere to a random uniform distribution. The results were as follows: for instances with 10 products and 25 stores, the average difference in CPU ticks between a cold start and a warm start is 0.222 CPU ticks when using solutions constructed by a GA and a CGA. Additionally, the difference for solutions built by a WCA is 0.022 CPU ticks.

Keywords Integer programming; Linear programming · ISHOP-U · Warm start · Cold start · CPLEX

F. Ornelas · A. Santiago · J. A. Castan Rocha (✉) · S. Ibarra Martínez · A. H. García
Faculty of Engineering Tampico, Autonomous University of Tamaulipas, Centro Universitario
Sur, 8109 Tampico, Tamaulipas, Mexico
e-mail: jacastan@docentes.uat.edu.mx

F. Ornelas
e-mail: a2193338003@alumnos.uat.edu.mx

A. Santiago
e-mail: aurelio.santiago@uat.edu.mx

S. Ibarra Martínez
e-mail: sibarram@docentes.uat.edu.mx

A. H. García
e-mail: ahgarcia@docentes.uat.edu.mx

7.1 Introduction

A wide variety of everyday problems require obtaining the best possible solution, considering a set of criteria met, whether to obtain the maximum benefit or the lowest cost. The topic of this chapter is online shopping. We observe the activities involved and examine which have received special attention from the scientific community. For example, order picking represents around 60% of warehouse operating costs [20]. Given that speed is a crucial factor for customer satisfaction, researchers aim to reduce operating costs by optimizing the locations of certain products and then determining the fastest routes to complete an order. Assuming robots are used for order picking, they will follow the shortest routes, calculating the expense of situating a product as the sum of the most direct trajectories for all orders enumerated (e.g., [23, 35]).

In mathematical optimization, when we focus on solving optimization problems where the variables can only take discrete values, we deal with integer programming (IP) problems, essentially planning problems with discrete values. IP is utilized in the deployment of shipping containers, with the goal of optimizing the quantity of goods loaded into a single transport unit. This involves considering rotation, dimensions, and stackability, where most products are similar. Delorme and Wagenaar [12] presents an integer programming model that accurately addresses this issue. The process involves the creation of item columns and proceeds with the resolution of a two-dimensional knapsack problem. The use of exact algorithms expedites the search process by capitalizing on the structure of the problem. The ultimate goal is to mitigate container loading issues.

A subset of Integer Programming (IP) problems are the Integer Linear Programming (ILP) problems, where the objective function and constraints incorporate only terms of the first degree, i.e., products of a variable by a real number. There are highly efficient algorithms for solving ILP problems, such as the simplex method, which relies on linear geometry and properties of systems of linear equations. In the context of developing new projects, budget and time problems are common, often caused by the complexity of project management. Most organizations tailor their project plans, resulting in suboptimal scheduling. This challenge is known as Resource-Constrained Project Scheduling (RCPS), which aims to reduce project time and costs. In [26], an ILP develops that incorporates person-based metrics, allows manual schedule adjustments to address unforeseen needs, and employs a sequential approach by dividing a group of projects into smaller subgroups. The aim is to mitigate problem complexity and provide quality solutions for each subgroup.

Mixed-integer linear Programming (MILP) is a commonly used subset of Integer Programming (IP). It deals with linear problems where the decision variables might be fractional. Both the objective function and constraints in MILP can contain linear terms with coefficients that are either integers or fractions. Some examples of MILP problems in the literature are last-mile delivery problems, which refer to the logistical challenge of delivering products from a warehouse or store to a customer's address. These problems are characterized by high delivery costs, a lack of transparency

throughout the process, interrupted order tracking, multiple delivery attempts, and inefficient routing. Munoz-Villamizar et al. [30] suggests a metaheuristic method that utilizes the Tabu Search algorithm. This approach tackles large-scale issues in tandem with a Mixed Integer Linear Programming (MILP) model. The model takes into account future demand and a cost framework to determine which deliveries should be immediate and which should be delayed. The delivery using unmanned aerial vehicles (UAVs), commonly known as drones, has generated special interest due to their environmental friendliness, shorter delivery times, and ability to reach remote areas. In [29], where drone usage is restricted solely to package delivery, prioritizing cost over route length, a Mixed Integer Linear Program (MILP) model accompanied by a collection of valid inequalities is presented. In all the cases examined, it is demonstrated that the use of drones is a promising option for last-mile delivery.

Nowadays, online purchasing has become a titanic task due to the vast number of suppliers for the same products and the multitude of purchasing combinations. Computational systems for automating online purchases can assist with this laborious task and reduce costs for buyers/companies by exploring store/product combinations. This chapter carries out the first exact implementation of a cold-start ISHOP-U integer programming model. Additionally, tests are conducted on the same integer programming model using the first warm start proposal to address the ISHOP-U problem in CPLEX. Currently, there is no analysis of when it is advisable to use a warm start versus a cold start within the state of the art. Warm start techniques offer a potential solution to obtaining optimal solutions within limited computational time budgets.

Few articles have been written about the problem of online shopping using exact methods. A description of these methods can be found in the Sect. 7.2. In the case of ISHOP-U in the current field, no work exists regarding exact methods for the problem of Internet Shopping with multiple units (ISHOP-U). Therefore, this work would be the first of its kind in the state of the art. This study compares initiating the exact solution from a warm start (given initial solutions) versus a cold start (without any initial solution). The initial solutions are obtained from the most recent solution algorithms in the latest advancements, which can be found in [36]. The initial solution-generating algorithms include a cellular genetic algorithm, a genetic algorithm, and a water cycle algorithm.

ISHOP-U is an NP-Hard problem, which means that finding the optimal solution to this problem in polynomial time is impossible. Consequently, computing solutions become infeasible as the problem size significantly increases, leading to exponential computation times. The instances used in this study are described in the Sect. 7.5.1. Problems belonging to the NP-Hard class interest the scientific community because they are considered the most challenging problems. An efficient solution for this problem has not been found, which justifies further research and developing new solution methods. In the literature, various integer programming problems fall into the NP-Hard class. Some examples include the set cover problem [27] and the knapsack problem [15].

Our main contribution in this chapter is conducting an empirical study comparing the performance of cold start versus warm start approaches for the Internet

Shopping Optimization Problem with Multiple Item Units (ISHOP-U). This study provides insights into specific scenarios where a warm start is beneficial and when it is not. Our experimental results and conclusions highlight these cases. Here is how the chapter is structured: Sect. 7.2 provides information on related studies that are relevant. Section 7.3 describes the Integer Linear Programming model. Section 7.4 presents a solution representation. Section 7.5 details the test setup. Section 7.6 displays our quantitative results. Section 7.7 presents our graphic results. Finally, Sect. 7.8 presents our conclusions.

7.2 Related Studies

The internet has entirely transformed the way we shop. Many people enjoy buying products on the internet these days. More online stores lead to greater competition, which lowers the prices of items. This makes products more affordable for customers and aligns with the business-to-consumer model. The above can result in feeling overwhelmed by the purchase process due to the large number of stores and delivery costs. One of the first efforts to solve this problem was intelligent software agents by Tolle and Chen [43]. These agents collect deals from various online stores and arrange them according to customers' needs. Later, Kwon and Sadeh suggested a method of online shopping that is aware of the context and allows for comparisons [21]. However, it had a downside: their method aimed to buy only a single item. This type of research led to the development of price comparison sites. Since platforms are commercial projects, they often prioritize maximizing revenue by guiding customers towards specific online retailers playing the role of recommender systems [39]. An unintended consequence of the previously mentioned issue is a decrease in customer trust.

Internet Shopping Optimization Problem (ISHOP) was proposed as an alternative to price comparison sites by Blazewicz [5] and then expanded in [3, 4]. The main idea is that the user creates a list of required items, and considering unit prices and delivery costs in a given set of stores, the algorithm chooses the cheapest solution, moving away from buying a single item as happens with the price comparators. This problem is addressed through two approaches, exact methods and heuristics, to develop satisfactory solutions. Exact methods guarantee that we find the optimum solution and are very useful when instances are not on a large scale. Conversely, heuristics allow us to find approximations to the optimum. In the state-of-art, we can see the implementation of both methods, as seen in [3, 4, 32, 33] where a Branch and Bound exact method developed according to literature [22] and then compared to heuristics. In the same way, López-Lóces [24] uses exact methods but employs an Integer Linear Programming Model (ILP) using IBM CPLEX optimization software [42].

ILP has been extensively utilized in the literature to solve a large number of problems, such as the Financial Budget of Universities in China [11] to analyze and optimize factors such as resource allocation, cost estimation, risk assessment, and scheduling for the execution of construction projects. The constructed model uses a Neural Network with ILP-based Combinatorial Optimization.

In like manner, in an airport for tactical flight rescheduling in Paris [41] if travelers cannot board their plane due to disruptions like subway closures or other disruptive events, understanding passenger delays helps in deciding to delay the flight departure, thus minimizing the number of passengers left stranded. They proposed a linear formulation of the problem, which allows for an exact solution approach using commercial software for solving Integer Linear Programming (ILP) models.

In the same way, the strategic placement of store branches [37] in Ahvaz, Iran, reduces costs and distance between stores and customers, as well as commute distances for employees to reach their workplace. A model based on integer linear programming proposes to divide a specific area within the city of Ahvaz into various scenarios. The aim is to determine the ideal number of stores while also managing the distance between operational stores.

In the search for the optimal solution, researchers implement an integer linear programming model, as shown above, in a solver. It is possible to provide hints to assist in finding an initial solution. These hints can include known variables and values, referred to as warm or advanced start. A warm start may be a feasible, infeasible, or incomplete solution. In the literature, we can find examples that use warm start, such as [7] presented a branch-and-cut algorithm to solve last-mile delivery logistics, which is the movement of orders from a distribution center to their final destinations, as they connect shippers, customers, and independent carriers to fulfill delivery requests, with a warm start heuristic to speed up the process. In a similar vein, [34] introduced an Automated Logistic Regression Solution Framework (ALRSF). This framework tackles the problem of best subset selection in logistic regression. It solves a mixed integer programming (MIP) formulation and employs a warm start for efficiency, allowing the solution in one iteration to be the entry point in the next iteration.

Now, we will present the work carried out in recent years. In 2020, an evolutionary approach was introduced using a genetic algorithm to address the ISHOP in [44], proposing shopping portals along with the products in stock at the shopping portal at a minimum price. A dataset with ten products in twenty shopping portals was used. In this work, the user sends a shopping list from their mobile device to a remote server that performs the calculations and returns a response.

In 2021, a robust optimization model applied to ISHOP in [8] that allows us to make decisions in fluctuating environments. The discussion focuses on obtaining an upper delivery time limit through an optimization model. When the delivery time is inconsistent, they depict the uncertainty as a problem of maximum flow that includes cyclical demand for the study. In addition, they take into account polyhedral uncertainty when setting up the adjustable robust counterpart optimization model.

In 2022, a bibliometric analysis of the literature was carried out along with a PRISMA meta-analysis [25], which allows for transparent reporting on why the review was conducted, what methods were used, and what the researchers found about the contributions to ISHOP. It highlights an annual research growth rate of 11.61% for the ISHOP.

In 2023, another study was conducted; this one focused on the proposed models and their solution methods. The following models are detailed in [28], which analyzes the following works. The proposed solution is an Internet shopping optimization problem with shipping costs. The proposed MinMin heuristic algorithm [24] traverses each shop, assigns a product, and checks if it reduces the overall expenditure. A cellular processing metaheuristic that works in parallel is also proposed, as well as a genetic algorithm [44], a water cycle algorithm [40], and a memetic algorithm-based metaheuristic [17]. Another model is the Internet shopping optimization problem with shipping costs and discounts, assuming all products are available in all stores. This model is limited as only one product of the same type can be purchased. The model first appeared in [32]. Simple heuristics were also proposed to solve it in [6], and it was modified by associating it with the facility location problem. In addition, a Tabu search and simulated annealing are proposed in [19]. Taking into account specific factors in the process of making choices, such as quantity, product weight, and availability. Other mentioned variants are the Internet Shopping Optimization Problem with price-sensitive discounts [3], where metaheuristics based on the cellular optimization process and a new greedy algorithm are considered, and the Trustworthy online shopping with price impact [31] in which a trust factor uses for cases where market reputation is important.

Finally, the Internet Shopping Optimization Problem with delivery constraints [10] is described, which is a bi-objective problem that considers time constraints and purchase cost, two heuristics proposed to obtain the optimal Pareto set. Also, in that same year, 2023, the Benders decomposition method [18] is proposed on a robust optimization model [1] that is treated in [9]. The case studies were applied to five products purchased from six stores by partitioning them into small subproblems since it is easier to solve computational calculations. Finally, in [14], a new solution method is presented that seeks to minimize each product's total cost and delivery time by applying two heuristics and achieving an Approximate Pareto Front (APF). They propose a decomposition-based algorithm as a solution to tackle the bi-objective optimization problem in online shopping.

In this chapter, we delve into the Internet SHopping Optimization Problem with multiple item Units (ISHOP-U), a particularly challenging variant of the ISHOP. This version of the problem arises in scenarios where acquiring multiple units of a particular product is necessary. To tackle this complex problem, we applied both exact and heuristic approaches. Specifically, we compared the Integer Programming Model for the ISHOP-U, implemented in IBM CPLEX optimization software, using a warm strategy vs a cold start, the warm start employing the best solutions produced by the algorithms in [36] as initial bounds. That is, starting from the best results obtained by a pair of evolutionary algorithms, a Genetic Algorithm (GA) [16], a Cellular Genetic Algorithm (CGA) [2], and a nature-inspired Water Cycle Algorithm (WCA) [13] in the state-of-the-art.

7.3 Integer Programming Model of the ISHOP-U

This section describes the Internet Shopping Optimization Problem with multiple item Units (ISHOP-U) that we implement in the CPLEX IBM optimization software as an Integer Programming problem.

7.3.1 ISHOP-U Mathematical Model

The fundamental components of the mathematical model include a collection of necessary quantities $L = \{l_1, l_2, \dots, l_n\}$ for purchasing n products where l_k signifies the quantity of product k to be purchased; an availability matrix $A = \{a_{1,1}, \dots, a_{i,j}, \dots, a_{m,n}\}$ for n products across m stores where each element $a_{i,j}$ specifies that store i has j accessible product units; a viable candidate solution S where each element $s_{i,j}$ denotes the quantity of product j to be purchased from store i , a unit cost matrix C where each element $c_{i,j}$ signifies the unit cost of product j in store i ; a collection of delivery fees $D = \{d_1, d_2, \dots, d_m\}$ for m stores. The following represents the minimization problem model for the ISHOP-U objective function:

$$F(S) = \sum_{i=1}^m \left(\sum_{j=1}^n s_{i,j} c_{i,j} + y_i d_i \right) \quad (7.1)$$

7.3.2 Constraints

y_i is a binary variable from the ISHOP-U Model, y_i returns one when a product or more products are bought, regardless of the amount, from the store i and *zero* otherwise as Eq. (7.2) indicates.

Subject to

$$y_i = \begin{cases} 1, & \sum_{j=1}^n s_{i,j} \neq 0 \\ 0, & \text{otherwise} \end{cases} \quad (7.2)$$

$$s_{i,j} \leq a_{i,j} \quad (7.3)$$

$$\sum_{i=1}^m s_{i,j} = l_j \quad (7.4)$$

In Eq. (7.3), the $s_{i,j}$ values of the candidate solution S cannot take a value greater than $a_{i,j}$ value of the availability matrix. For this constraint, the product must be available in a given set of stores. Finally, the last constraint in Eq. (7.4) guarantees to buy the required number of units per product.

7.4 ISHOP-U Numerical Example

This part provides a numerical illustration of the ISHOP-U problem. In this example, five items are labeled from V to Z. Table 7.1 presents the units available at each store, and Table 7.2 displays the unit costs and delivery fees.

It's necessary to understand the number of item units needed per product. Table 7.3 displays the units we need to purchase.

Being aware of availability, requirements, and their unit price of units, the solution design formulation can be like Table 7.4.

Table 7.1 Available items offered in the five internet stores

Stores	V	W	X	Y	Z
1	2	3	7	9	6
2	6	4	8	3	2
3	4	2	3	4	5

Table 7.2 Unit cost and delivery offered by five internet stores

Stores	V	W	X	Y	Z	Delivery
1	24	40	29	48	57	15
2	18	45	20	57	53	10
3	22	42	17	55	44	15

Table 7.3 Unit cost and delivery fees provided by five online retailers

	V	W	X	Y	Z
Units	7	2	4	8	6

Table 7.4 Feasible candidate solution

Stores	V	W	X	Y	Z
1	0	2	0	8	0
2	6	0	1	0	1
3	1	0	3	0	5

Finally, We compute the multiplication of the unit price and the units allocated to the store, and then add the aggregate of the delivery costs for each store, considering that The delivery fee is assigned a single time if multiple products come from the same shop. The objective value of this numerical example is 978.

7.5 Test Setup

This section outlines the collection of instances (Sect. 7.5.1), nature-inspired and evolutionary algorithms used (Sect. 7.5.2), and the computational experiments (Sect. 7.5.3).

7.5.1 *Synthetic Instances*

To evaluate the warm start, we use a set of 3 different sizes of uniform synthetic instances of the original ISHOP-U., a small subset of 5 instances (10 products with 25 stores), a medium subset of 5 instances (25 products with 50 stores), and a large subset of 5 instances (50 products with 100 stores). The unit price varies in the interval [5,50], and delivery prices vary in the range of [0,10]. Instances available at <https://github.com/Fernando-Ornelas/IShopUWarmStart/tree/main/ISHOP-U>.

7.5.2 *Nature-Inspired and Evolutionary Algorithms*

To analyze the warm starting, we use evolutionary algorithms (GA and CGA), choosing the optimal configuration as documented in the latest research [36] ($p_r = 1.0$, $p_m = 0.05$). The population equals 100 individuals with a stop condition equivalent to 250 generations. Moreover, We use the Water Cycle Algorithm [13], the Ali Sadollah (2022), Unconstrained Discrete version 2 serves as the foundation for ISHOP-U implementation [38]. The parameter configurations are adjusted following the author's suggestions.

7.5.3 *Computational Experiments*

Computational experiments use the best-found solution from 30 independent runs for all the considered algorithms. The experimental computer is an Alienware M14 \times 2.4Ghz Intel Core i7 with 16GB of RAM. The CPU time in CPU ticks is the average of the 30 independent executions. The computer was not executing other algorithms or applications, except for the boot ones in the Windows 10 pro

operating system. The algorithm's implementation was under Java, and the Integer programming model was on the OPL (Optimization programming language) from IBM ILOG CPLEX Studio 22.1.0. The warm start experiment over the interactive optimizer 22.1.0.0. Code available in <https://github.com/Fernando-Ornelas/IShopUWarmStart>.

7.6 Quantitative Results

This section presents the comparative results of the IP model described in 7.3.1 and the evolutionary algorithms: Cellular Genetic Algorithm (CGA), Genetic Algorithm (GA), and a nature-inspired Water Cycle Algorithm (WCA) for the 15 uniform instances present in the literature as is described in 7.5.2. Table 7.5 shows the average time of 30 independent executions in IBM CPLEX Interactive optimizer 22.1.0.0 measured in CPU ticks, highlighting the best average times in light gray. In the first column, we can observe the cold execution of the IP model in CPLEX. In the remaining columns, the warm start execution of the model employs the best solution derived from the methods mentioned above. The GA and CGA algorithms achieve the best warm start performance for instances of 10 products and 25 stores. For the rest of the instances, the cold start outperforms. The equal CPU ticks between GA and CGA are because both evolutionary algorithms possess identical computational complexity $O(nm)$ for both the crossover and mutation operators.

Table 7.5 Average times for CPLEX warm started by GA, CGA, WCA and CPLEX cold start

Instance	CPLEX	GA	CGA	WCA
UniformS1	4.55	4.19	4.19	4.49
UniformS2	3.94	3.85	3.85	3.93
UniformS3	4.18	4.04	4.04	4.16
UniformS4	4.71	4.45	4.45	4.69
UniformS5	4.78	4.52	4.52	4.78
UniformM1	13.56	15.70	15.70	15.70
UniformM2	10.83	12.96	12.96	12.96
UniformM3	11.43	13.55	13.55	13.55
UniformM4	11.16	13.30	13.30	13.30
UniformM5	10.80	12.94	12.94	12.94
UniformL1	65.25	71.97	72.95	72.81
UniformL2	64.45	71.75	72.61	73.89
UniformL3	77.07	88.37	88.41	94.98
UniformL4	55.37	66.63	66.63	80.94
UniformL5	50.14	57.88	57.88	58.02

Table 7.6 Objective values

Instance	CPLEX	GA	CGA	WCA
UniformS1	447.20	447.20	447.20	494.08
UniformS2	447.22	447.22	477.22	476.57
UniformS3	524.25	524.25	524.25	594.05
UniformS4	462.92	462.92	462.92	507.81
UniformS5	489.62	489.62	489.62	529.50
UniformM1	2164.58	3289.45	3442.83	3634.32
UniformM2	3069.90	4526.51	4730.03	5039.40
UniformM3	3248.58	5083.37	5345.31	5463.01
UniformM4	3259.24	4929.07	5157.62	5140.60
UniformM5	3536.99	5234.01	5506.20	5835.63
UniformL1	9110.89	25174.45	25238.80	23871.20
UniformL2	8072.24	21472.10	21473.02	20630.38
UniformL3	7554.83	20817.67	21195.73	19892.48
UniformL4	8846.62	23001.14	23964.49	22336.05
UniformL5	8351.03	22916.85	23722.15	22917.44

In Table 7.6, the optimal time obtained by CPLEX is displayed in the first column. The time the algorithms obtain is in the remaining columns (CPU ticks units).

The CPLEX interactive optimizer uses the representation of matrix S in an mst file, equivalent to the initial solution in the warm start. The best result is emphasized in dark gray, while the second-best result is light gray for each row.

Subsequently, Table 7.7 presents the approximation factor for the analyzed warm starts by the evolutionary algorithms and a nature-inspired algorithm. It refers to the correlation between the solution’s value, which the approximation algorithm achieves, and the optimal solution’s value. The approximation factor is $\rho = F(x)/F(X)^*$, where $F(x)$ represents the solution found by the approximation and $F(X)^*$ denotes the optimal solution, providing an insight into how much the obtained solution deviates from the global optimum. A small approximation factor indicates that the algorithm produces solutions close to the optimum, while a large one indicates that the algorithm may produce significantly worse solutions than the optimum. The table highlights the minor approximation factors obtained by the warm starts.

Furthermore, a Wilcoxon Signed-Rank Test is presented, which evaluates whether a significant difference exists between the values of two related samples. It compares whether the differences between paired data follow a symmetric distribution, meaning there are equal values to the right and left of the median. Consequently, there are the same number of positive and negative signed deviations around a central value. A significant result in the Wilcoxon Signed-Rank Test indicates a low probability that the difference between the two samples is due to chance. Conversely, if no significant difference exists, the samples are equivalent. Table 7.8 shows the statistical

Table 7.7 Approximation factor

Instance	CPLEX + GA	CPLEX + CGA	CPLEX + WCA
UniformS1	1.00	1.00	1.10
UniformS2	1.00	1.07	1.07
UniformS3	1.00	1.00	1.13
UniformS4	1.00	1.00	1.10
UniformS5	1.00	1.00	1.08
UniformM1	1.52	1.59	1.68
UniformM2	1.47	1.54	1.64
UniformM3	1.56	1.65	1.68
UniformM4	1.51	1.58	1.58
UniformM5	1.48	1.56	1.65
UniformL1	2.76	2.77	2.62
UniformL2	2.66	2.66	2.56
UniformL3	2.76	2.81	2.63
UniformL4	2.60	2.71	2.52
UniformL5	2.74	2.84	2.74

Table 7.8 Wilcoxon test for statistical significance over CPU time (p-value < 0.05)

Instance	ColdStart/WarmGA	ColdStart/WarmCGA	ColdStart/WarmWCA
UniformS1	Yes	Yes	Yes
UniformS2	Yes	Yes	Yes
UniformS3	Yes	Yes	Yes
UniformS4	Yes	Yes	Yes
UniformS5	Yes	Yes	No
UniformM1	Yes	Yes	Yes
UniformM2	Yes	Yes	Yes
UniformM3	Yes	Yes	Yes
UniformM4	Yes	Yes	Yes
UniformM5	Yes	Yes	Yes
UniformL1	Yes	Yes	Yes
UniformL2	Yes	Yes	Yes
UniformL3	Yes	Yes	Yes
UniformL4	Yes	Yes	Yes
UniformL5	Yes	Yes	Yes

significance test related to CPU time. No significant differences were found for the Uniform instance S5 when comparing cold start versus warm start initiated with the WCA algorithm. However, significant differences were observed in all other cases regarding the generated times.

Table 7.9 Wilcoxon Test for statistical significance over objective values (p-value < 0.05)

Instance	ColdStart/WarmGA	ColdStart/WarmCGA	ColdStart/WarmWCA
UniformS1	No	No	Yes
UniformS2	No	Yes	Yes
UniformS3	Yes	Yes	Yes
UniformS4	No	Yes	Yes
UniformS5	No	Yes	Yes
UniformM1	Yes	Yes	Yes
UniformM2	Yes	Yes	Yes
UniformM3	Yes	Yes	Yes
UniformM4	Yes	Yes	Yes
UniformM5	Yes	Yes	Yes
UniformL1	Yes	Yes	Yes
UniformL2	Yes	Yes	Yes
UniformL3	Yes	Yes	Yes
UniformL4	Yes	Yes	Yes
UniformL5	Yes	Yes	Yes

Finally, in Table 7.9, No statistical significance was found in the differences of objective values for the Uniform instances UniformS1, UniformS2, UniformS4, and UniformS5 regarding cold start versus the warm start of both the GA and the CGA, specifically in the case of instance S1. However, significant differences were observed in all other cases.

Three conditions exist to determine whether it is advisable to apply a warm start: (1) there is a significant difference in CPU time. (2) There appears to be no substantial variation in the objective value. (3) That the CPU time of the warm start is less than the time it takes for the cold start. The result was that a warm start GA for instances UniformS1, UniformS2, UniformS4, and UniformS5 is advisable. In addition, a warm start CGA, such as UniformS1, is advisable.

7.7 Graphic Results

In Fig. 7.1, section (a) depicts the average time CPLEX takes to optimize small instances after 30 executions. Notably, for the UniformS1 instance, both GA and CGA warm starts exhibit a difference of 0.36 CPU ticks compared to cold starts. Similarly, for UniformS2, the warm starts of GA and CGA show a difference of 0.09 CPU ticks. In the case of UniformS3, a difference of 0.14 CPU ticks is observed. For UniformS4 and UniformS5, the warm starts of GA and CGA present a difference of 0.26 compared to cold starts. Moving to section (b), we notice that for UniformM1, UniformM4, and UniformM5, cold starts have a difference of 2.14 CPU ticks com-

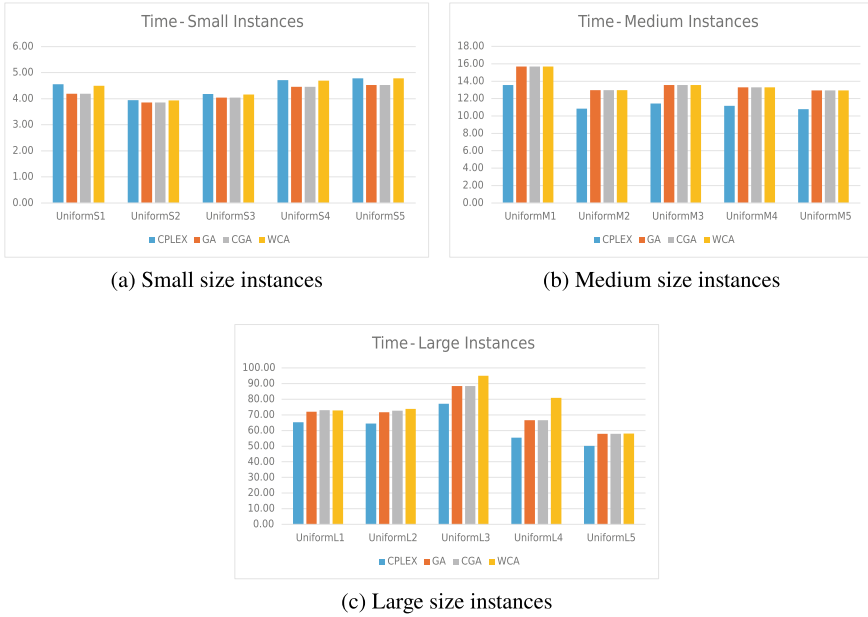


Fig. 7.1 Graphical results in CPU ticks time, **a** five small instances, **b** five medium instances, **c** five large instances

pared to GA and CGA. However, for UniformM2 and UniformM3, the difference between cold starts reduces to 2.13 and 2.12 CPU ticks relative to GA and CGA. Finally, in section (c), the best start strategy is the cold start, which exhibits a difference of 6.72, 7.3, and 11.3 CPU ticks for UniformL1, UniformL2, and UniformL3 instances, respectively, compared to GA. Additionally, for GA and CGA, the difference is 11.26 and 7.74 CPU ticks, respectively. Another point to highlight is that as the instance size increases, its CPU time also increases. Additionally, we can observe that cold start tends to improve with larger instance sizes.

Subsequently, Fig. 7.2, section (a) presents the optimal value obtained by CPLEX, GA, and CGA for the small instances with a difference of 46.88, 29.35, 69.80, 44.89, and 39.88 for instances UniformS1, UniformS2, UniformS3, UniformS4, and UniformS5 respectively in purchase cost obtained by WCA versus the optimal value of the instance. For section (b), the optimal value obtained by CPLEX and the highest objective function value obtained by WCA for the medium instances shows a difference of 1, 469.74, 1, 969.50, 2, 214.43, 1, 881.36, and 2, 298.64 for instances UniformM1, UniformM2, UniformM3, UniformM4, and UniformM5 in the given order in total purchase cost. Finally, for section (c), the first column exposes the optimal value obtained by CPLEX. It compares it with the highest objective function value obtained by CGA for the large instances with a disparity of 16, 127.91, 13, 400.78, 13, 640.90, 15, 117.87, and 15, 371.12 for instances UniformL1, UniformL2, UniformL3, UniformL4, UniformL5 for the purchase cost.

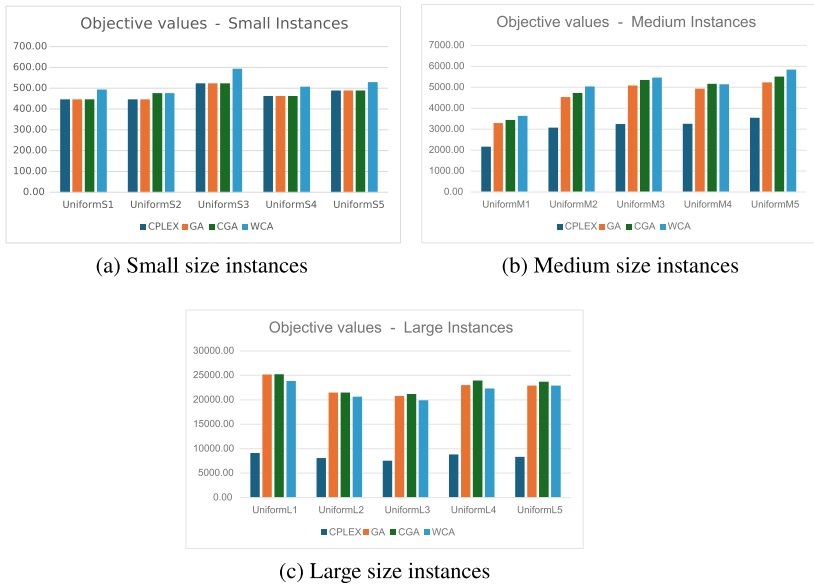


Fig. 7.2 Graphical results in purchase cost, **a** five small instances, **b** five medium instances, **c** five large instances

7.8 Discussion and Conclusion

In this work, we developed an Integer Programming model in CPLEX for ISHOP-U. Subsequently, we conducted 30 independent executions of two evolutionary algorithms, GA, CGA, and a nature-inspired WCA, for five small, medium, and large instances. We obtained their best-found solution, which became a warm start to the CPLEX mst file. The mst file constitutes a warm start for each instance and each algorithm. We performed 30 independent executions of the warm starts for each instance/algorithm best-found solution and 30 cold starts, obtaining the average execution time in CPU ticks. The objective values from which the warm start begins are presented, along with an approximation factor indicating how far the solution deviates from the global optimum. The initial solutions in the warm start with an approximation factor higher than $\rho = 1.07$ result in a higher CPU time than a cold start. We can highlight that a warm start can help reduce the calculation time as it starts from an established solution. However, without having a solution close to the optimum, the predefined solution could increase search time. On the other hand, a cold start does not use any previous solution, which can cause the search process to be slower as it is necessary to explore the entire solution space. This study is limited to the algorithms present in the state of the art for ISHOP-U. Future research directions are described below. It is interesting to consider new solution methods for the warm start using heuristics and metaheuristics. In addition, a hybridized heuristic approach with exact methods.

Acknowledgements Alejandro Santiago would like to thank CONAHCYT Mexico for the SNI salary award.

References

1. Agustini, R.A., Chaerani, D., Hertini, E.: Adjustable robust counterpart optimization model for maximum flow problems with box uncertainty. *World Scient. News* **141**, 91–102 (2020)
2. Alba, E., Dorronsoro, B.: *Introduction to Cellular Genetic Algorithms*, pp. 3–20. Springer US, Boston, MA (2008). https://doi.org/10.1007/978-0-387-77610-1_1
3. Blazewicz, J., Bouvry, P., Kovalyov, M.Y., Musial, J.: Internet shopping with price sensitive discounts. *4OR* **12**, 35–48 (2014). <https://doi.org/10.1007/s10288-013-0230-7>
4. Blazewicz, J., Cherière, N., Dutot, P.F., Musial, J., Trystram, D.: Novel dual discounting functions for the internet shopping optimization problem: new algorithms. *J. Schedul.* **19**, 245–255 (2016). <https://doi.org/10.1007/s10951-014-0390-0>
5. Blazewicz, J., Kovalyov, M., Musial, J., Urbanski, A., Wojciechowski, A.: Internet shopping optimization problem. *Int. J. Appl. Math. Comp. Sci.*, 385–390 (2010). <https://doi.org/10.2478/v10006-010-0028-0>
6. Błażewicz, J., Musiał, J.: E-commerce evaluation—multi-item internet shopping. Optimization and heuristic algorithms. In: *Operations Research Proceedings 2010*, pp. 149–154. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20009-0_24
7. Cerulli, M., Archetti, C., Fernandez, E., Ljubic, I.: A bilevel approach for compensation and routing decisions in last-mile delivery (2023). arXiv preprint . <https://doi.org/10.48550/arXiv.2304.09170>
8. Chaerani, D., Rusyaman, E., Mahrudinda, Marcia, A., Fridayana, A.: Adjustable robust counterpart optimization model for internet shopping online problem. *J. Phys.: Conf. Series* **1722**(1), 012074 (2021). <https://doi.org/10.1088/1742-6596/1722/1/012074>
9. Chaerani, D., Saksmitena, S., Irmansyah, A.Z., Hertini, E., Rusyaman, E., Paulus, E.: Benders decomposition method on adjustable robust counterpart optimization model for internet shopping online problem. *Comput.* **11**(2), 37 (2023)
10. Chung, J.B.: Internet shopping optimization problem with delivery constraints. *J. Distrib. Sci.* **15**, 15–20 (2017). <https://doi.org/10.15722/jds.15.2.201702.15>
11. Dai, L.D.L., et al.: Neural network algorithm optimization for financial budget of universities. *J. Elect. Syst.* **20**(1), 158–175 (2024). <https://doi.org/10.52783/jes.674>
12. Delorme, M., Wagenaar, J.: Exact decomposition approaches for a single container loading problem with stacking constraints and medium-sized weakly heterogeneous items. *Omega* **125**, 103039 (2024). <https://doi.org/10.1016/j.omega.2024.103039>
13. Eskandar, H., Sadollah, A., Bahreininejad, A., Hamdi, M.: Water cycle algorithm—a novel metaheuristic optimization method for solving constrained engineering optimization problems. *Comp. Struct.* **110–111**, 151–166 (2012). <https://doi.org/10.1016/j.compstruc.2012.07.010>
14. García-Morales, M.A., Brambila-Hernández, J.A., Fraire-Huacuja, H.J., Frausto-Solis, J., Cruz-Reyes, L., Gómez-Santillán, C.G., Valadez, J.M.C., Aguirre-Lam, M.A.: Multi-objective evolutionary algorithm based on decomposition to solve the bi-objective internet shopping optimization problem (moea/d-bishop). In: *Mexican International Conference on Artificial Intelligence*, pp. 326–336. Springer (2023). https://doi.org/10.1007/978-3-031-51940-6_24
15. He, Q., Xu, Z.: Simple and faster algorithms for knapsack. In: *2024 Symposium on Simplicity in Algorithms (SOSA)*, pp. 56–62. SIAM (2024). <https://doi.org/10.1137/1.9781611977936.6>
16. Holland, J.H.: *Adaptation in Natural and Artificial Systems*. The University of Michigan (1975). <https://doi.org/10.1137/1018105>
17. Huacuja, H.J.F., Morales, M.Á.G., Locés, M.C.L., Santillán, C.G.G., Reyes, L.C., Rodríguez, M.L.M.: Optimization of the Internet Shopping Problem with Shipping Costs, 249–255.

- Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-68776-2_14
18. Irmansyah, A.Z., Chaerani, D., Rusyaman, E.: A systematic review on integer multi-objective adjustable robust counterpart optimization model using benders decomposition. *JTAM (Jurnal Teori dan Aplikasi Matematika)* **6**(3), 678–698 (2022). <https://doi.org/10.31764/jtam.v6i3.8578>
 19. Józefczyk, J., Ławrynowicz, M.: Heuristic algorithms for the internet shopping optimization problem with price sensitivity discounts. *Kybernetes* **47**(4), 831–852 (2018). <https://doi.org/10.1108/K-07-2017-0264>
 20. Kordos, M., Boryczko, J., Blachnik, M., Golak, S.: Optimization of warehouse operations with genetic algorithms. *App. Sci.* **10**(14) (2020). <https://doi.org/10.3390/app10144817>. <https://www.mdpi.com/2076-3417/10/14/4817>
 21. Kwon, O.B., Sadeh, N.: Applying case-based reasoning and multi-agent intelligent system to context-aware comparative shopping. *Decis. Supp. Syst.* **37**(2), 199–213 (2004). [https://doi.org/10.1016/S0167-9236\(03\)00007-1](https://doi.org/10.1016/S0167-9236(03)00007-1)
 22. Land, A.H., Doig, A.G.: *An Automatic Method for Solving Discrete Programming Problems*, 105–132. Springer Berlin Heidelberg, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-540-68279-0_5
 23. López, I.B., Saldaña, R., Rivera, G.: Modular framework for crowd simulation “menge” from a production warehouse simulation perspective. In: *Technological and Industrial Applications Associated with Intelligent Logistics*, 301–311 (2021). https://doi.org/10.1007/978-3-030-68655-0_16
 24. Lopez-Loces, M.C., Musial, J., Pecero, J.E., Fraire-Huacuja, H.J., Blazewicz, J., Bouvry, P.: Exact and heuristic approaches to solve the internet shopping optimization problem with delivery costs. *Int. J. Appl. Math. Comp. Sci.* **26**(2), 391–406 (2016). <https://doi.org/10.1515/amcs-2016-0028>
 25. Mahrudinda, M., Chaerani, D., Rusyaman, E.: Systematic literature review on adjustable robust counterpart for internet shopping optimization problem. *Int. J. Data Netw. Sci.* **6**(2), 581–594 (2022). <https://doi.org/10.52677/j.ijdns.2021.11.006>
 26. Manousakis, K., Savva, G., Papadouri, N., Mavrovouniotis, M., Christofides, A., Kolokotroni, N., Ellinas, G.: A practical approach for resource-constrained project scheduling. *IEEE Acc.* **12**, 12976–12991 (2024). <https://doi.org/10.1109/ACCESS.2024.3352438>
 27. Meck, M., Pelz, P.F.: A set-covering approach to minimize the variety of standard chemical process pumps in equipment pools. *Comp. Chem. Eng.* **185**, 108673 (2024). <https://doi.org/10.1016/j.compchemeng.2024.108673>
 28. Morales, M.Á.G., Huacuja, H.J.F., Solís, J.F., Reyes, L.C., Santillán, C.G.G.: *A Survey of Models and Solution Methods for the Internet Shopping Optimization Problem*, 105–122. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-22042-5_6
 29. Mulumba, T., Diabat, A.: Optimization of the drone-assisted pickup and delivery problem. *Transp. Res. Part E: Logis. Transp. Rev.* **181**, 103377 (2024). <https://doi.org/10.1016/j.tre.2023.103377>
 30. Muñoz-Villamizar, A., Velazquez-Martínez, J.C., Caballero-Caballero, S.: A large-scale last-mile consolidation model for e-commerce home delivery. *Expert Syst. Appl.* **235**, 121200 (2024). <https://doi.org/10.1016/j.eswa.2023.121200>
 31. Musial, J., Lopez-Loces, M.C.: Trustworthy online shopping with price impact. *Found. Comp. Decision Sci.* **42**(2), 121–136 (2017). <https://doi.org/10.1515/fcds-2017-0005>
 32. Musial, J., Pecero, J., Lopez-Loces, M., Fraire-Huacuja, H., Bouvry, P., Blazewicz, J.: Algorithms solving the internet shopping optimization problem with price discounts. *Bull. Polish Acad. Sci. Tech. Sci.* **64**(3), 505–516 (2016). <https://doi.org/10.1515/bpasts-2016-0056>
 33. Musial, J., Pecero, J.E., Lopez, M.C., Fraire, H.J., Bouvry, P., Blazewicz, J.: How to efficiently solve internet shopping optimization problem with price sensitive discounts? In: *2014 11th International Conference on e-Business (ICE-B)*, pp. 209–215. IEEE (2014). <https://doi.org/10.5220/0005112602090215>

34. van Niekerk, T.K., Venter, J.V., Terblanche, S.E.: An automated exact solution framework towards solving the logistic regression best subset selection problem. *South African Statist. J.* **57**(2), 89–129 (2023). <https://doi.org/10.37920/sasj.2023.57.2.2>
35. Olmos, J., Florencia, R., García, V., González, M.V., Rivera, G., Sánchez-Solís, P.: Metaheuristics for order picking optimisation: a comparison among three swarm-intelligence algorithms. In: *Technological and industrial applications associated with industry 4.0*, 177–194 (2022). https://doi.org/10.1007/978-3-030-68663-5_13
36. Ornelas, F., Santiago, A., Martínez, S.I., Ponce-Flores, M.P., Terán-Villanueva, J.D., Balderas, F., Rocha, J.A.C., García, A.H., Laria-Menchaca, J., Treviño-Berrones, M.G.: The internet shopping optimization problem with multiple item units (ISHOP-U): formulation, instances, NP-completeness, and evolutionary optimization. *Math.* **10**(14) (2022). <https://doi.org/10.3390/math10142513>
37. Paeizi, A., Makui, A.: An integer linear programming approach for a location-allocation problem in online stores industry: A real world case study. *J. Future Sustain.* **4**(2), 77–84 (2024). <https://doi.org/10.5267/j.jfs.2024.5.002>
38. Sadollah, A., Eskandar, H., Lee, H.M., Yoo, D.G., Kim, J.H.: Water cycle algorithm: a detailed standard code. *SoftwareX* **5**, 37–43 (2016). <https://doi.org/10.1016/j.softx.2016.03.001>
39. Satzger, B., Endres, M., Kießling, W.: A preference-based recommender system. In: *International Conference on Electronic Commerce and Web Technologies*, pp. 31–40. Springer (2006). https://doi.org/10.1007/11823865_4
40. Sayyaadi, H., Sadollah, A., Yadav, A., Yadav, N.: Stability and iterative convergence of water cycle algorithm for computationally expensive and combinatorial internet shopping optimisation problems. *J. Exp. Theoret. Artif. Intell.* **31**(5), 701–721 (2019). <https://doi.org/10.1080/0952813X.2018.1549109>
41. Scozzaro, G., Mancel, C., Delahaye, D., Feron, E.: An ILP approach for tactical flight rescheduling during airport access mode disruptions. *Int. Trans. Operat. Res.* **31**(3), 1426–1457 (2024). <https://doi.org/10.1111/itor.13396>
42. Studio, I.I.C.O.: V20. 1: User’s manual for cplex. IBM Corp (2020)
43. Tolle, K.M., Chen, H.: Intelligent software agents for electronic commerce. In: *Handbook on Electronic Commerce*, 365–382 (2000). https://doi.org/10.1007/978-3-642-58327-8_17
44. Verma, S., Sinha, A., Kumar, P., Maitin, A.: Optimizing online shopping using genetic algorithm. In: *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, pp. 271–275 (2020). <https://doi.org/10.1109/ICICT50521.2020.00048>

Chapter 8

Hybrid Genetic Algorithm Based on Machine Learning and Fitness Function Estimation Proposal for Ground Vehicle and Drone Cooperative Delivery Problem



Muhammed Mirac Özer 

Abstract In this study, the development of a hybrid genetic algorithm, integrating machine learning and function estimation, presents a novel approach to address the simultaneous intervention challenge involving unmanned ground vehicles (UGVs) and unmanned aerial vehicles (UAVs). The adaptability of this hybrid genetic algorithm confers a notable advantage in managing drone scenarios. Notably, this work constitutes the inaugural attempt in the literature to devise an exact solution for the concurrent intervention of a UGV and a UAV, with the added innovation of minimizing intervention time. This pioneering methodology holds promise for extending the problem domain to encompass more realistic scenarios, thereby bridging a significant gap in the literature and furnishing a foundational framework for future research endeavors.

Keywords Machine learning · Fitness function estimation · Hybrid genetic algorithm · Vehicle routing problem with drone · Traveling salesman person

8.1 Introduction

Drones, officially called unmanned aerial vehicles (UAVs), represent one of the most remarkable and rapidly developing areas of the aerospace industry today [1, 2]. This technology, which was initiated in earlier years for military applications, is now used for a variety of purposes in many different sectors [3–8]. The structure of drones is basically based on a design that requires them to be lightweight and compact [9]. These unmanned aerial vehicles are usually equipped with four

M. M. Özer (✉)

Faculty of Aeronautics and Astronautics, Unmanned/Intelligent Systems Lab, Tarsus University, Mersin 33400, Turkey

e-mail: muhammed_mirac@tarsus.edu.tr

propellers, which allow them to perform different flight maneuvers [10]. In addition, the sensors, cameras and GPS systems in the structure of drones allow these devices to perceive their environment, collect data and perform precise missions [11]. The areas of use of drones are quite wide and constantly expanding [12–20]. From construction [21] to agricultural applications [22], from security inspections [23] to environmental monitoring [24], drones are recognized as a cost-effective and efficient tool in many sectors [25–28]. Moreover, their ability to be used quickly and effectively in emergency response [29] helps in the successful execution of rescue operations [30, 31]. One of the main benefits of drone technology to the scientific and industrial world is its ability to collect and analyze data. These devices can monitor changes in different parts of the earth, collect environmental data and present this data to scientists, engineers and decision makers [32–34]. It also has great potential in energy efficiency and resource management. Because it can contribute to the development of environmentally friendly and sustainable applications. As a result, drone technology has revolutionized the field of aviation and is having a major impact on industrial and scientific fields [35–39].

In addition, the transportation [40–43] and logistics sector [44, 45] constitutes one of the cornerstones of the modern economy and remains an area that needs to be continuously improved in terms of efficiency, resource utilization and profitability [46, 47]. In this context, a large number of mathematical models and algorithms have been developed to optimize and make transportation operations with conventional vehicles more efficient [48–52]. However, today's rapidly evolving technology requires us to re-evaluate the traditional approaches in this area and search for more sustainable, fast and cost-effective solutions. "Vehicle Routing Problems with Drones" (VRP-D) is an important research area in this context, offering a new paradigm in transportation and logistics [53, 54]. VRP-D mainly involves the use of drones, in addition to traditional means of transportation and aims to make transportation processes more effective and efficient [55, 56]. Using complex mathematical optimization models and algorithms, VRP-D aims to reduce delivery costs, shorten delivery times and minimize environmental impacts by making the best use of existing transport networks and resources [57–61]. This new paradigm offers potential in many application areas [62–64]. The use of VRP-D in different sectors such as distribution logistics, emergency response, healthcare and agriculture has great potential to enable faster and more precise deliveries [65, 66]. It also brings advantages such as traffic congestion reduction, efficient route planning and energy savings in urban transportation [67, 68].

Technological advances in distribution processes have increased the demand for new methods in transportation and logistics. In this context, the integration of UAVs in distribution processes constitutes a critical research area that requires an understanding of the fundamental differences between the traditional Vehicle Routing Problem (VRP) [69] and the UAV Routing Problem (VRP-D) [70]. The traditional VRP [71] aims to optimize the distribution of goods and services by transport vehicles to customers in a given region. In this problem, the determination of vehicle routes and delivery sequences is essential. VRP-D, on the other hand, involves the integration of UAVs into distribution processes, which introduces new dynamics and

challenges. The ability of UAVs to provide airborne transportation provides different advantages than road transportation. The complexity of VRP-D stems from the fact that it involves the simultaneous deployment of trucks and UAVs. This implies the need for time and location coordination and requires a different solution strategy than the traditional VRP. At this point, unlike the existing literature, this study aims to contribute to the solution of VRP-D by developing a hybrid genetic algorithm-based function estimation approach.

In the ever-evolving landscape of logistics and transportation, efficiently allocating resources and optimizing delivery routes remain key challenges [72, 73]. The arrival of drones and the advancement of Genetic Algorithms (GA) [74–78] have transformed our approach to these optimization problems [79, 80]. However, recognizing the synergistic potential of combining these two technologies- Hybrid Genetic Algorithms (HGA) and Vehicle Routing Problems with Drones (VRP-D)- opens a new dimension in optimization and decision making [81, 82]. VRP-D addresses complex real-world scenarios that take into account air and ground logistics in delivery and route planning, while HGAs embrace the ability to use the principles of natural selection and evolution to find optimal solutions. The integration of HGA and VRP-D offers a new paradigm in logistics optimization. This approaches to address the limitations and complexities of traditional VRPs by aiming to simultaneously optimize both route optimization for conventional vehicles and drone deliveries. This is achieved by taking into account constraints such as vehicle capacity, time windows and drone flight limitations. The potential applications of this hybrid approach are quite broad. From urban parcel deliveries to last-mile logistics, from disaster relief operations to precision agriculture, HGA-VRP-D integration holds the promise of improving delivery speed, cost-effectiveness and environmental sustainability.

Mobile robots or unmanned ground vehicles (UGVs), on the other hand, represent a burgeoning field of study within academic research, offering a diverse range of applications and challenges [83]. These autonomous or semi-autonomous systems traverse various environments, from controlled laboratory settings to complex real-world scenarios, making them invaluable tools for investigating navigation, perception, control, and human–robot interaction. Current literature explore a myriad of topics, including localization and mapping techniques, path planning algorithms, sensor integration, and machine learning approaches to enhance robot autonomy and adaptability [84]. Furthermore, mobile robots play a pivotal role in interdisciplinary research, facilitating advancements in fields such as delivery services and surveillance.

The use of UAVs in delivery services also has gained increasing popularity in recent years, despite technical, legal and social challenges. This new and still emerging field is gaining prominence with the adoption of a model where trucks and UAVs can deliver simultaneously. However, this new delivery model, which requires time and location coordination of both vehicles, makes the routing problem quite complex. It is inevitable to develop new algorithms to overcome this challenge and provide efficient solutions. This paper aims to present a new approach to the deployment model by improving existing heuristic solutions in the literature.

This delivery model, in which the UAV is used simultaneously with the truck, involves multiple UAVs delivering at the same time, which reflects a possible and realistic scenario. Therefore, it is of great importance to study this scenario in detail from an operational perspective. The focus of the study is the development of an effective solution approach that can be easily adapted to multiple UAV scenarios.

In the light of the existing knowledge in this field, it is possible to further elaborate on the advantages of UAV integration and truck synchronization in the distribution model. With the increasing use of UAVs, it is clear that the integration of this technology into distribution processes can provide significant contributions in many aspects such as work efficiency, cost reduction and environmental impact. In this context, the new solution approach to be developed will allow distribution operations to be carried out in a more optimized and efficient manner.

In conclusion, this study presents a new perspective for the effective utilization of UAVs in distribution processes. The solution approach to be developed aims to overcome the existing challenges and manage the distribution operations in a more effective and sustainable manner. It is expected to provide an important basis for future research and applications in this field.

The concept of using UAVs in distribution processes has attracted the attention of not only academia but also the business community. This widespread interest in this innovative delivery model can be explained by the fact that the business community has realized its potential advantages. This new problem, involving the simultaneous deployment of trucks and UAVs, requires a detailed study of various scenarios for its real-world implementation and the identification of the contribution in this field. Therefore, the development of effective solution approaches is an important need.

A hybrid genetic algorithm based on a machine learning based function estimation method is proposed as a solution to the simultaneous deployment problem of trucks and UAVs. The results compared with the previously used simulated annealing method show that the hybrid genetic algorithm outperforms the developed hybrid genetic algorithm, especially for small and medium-sized problems. The ability of this method to be easily adapted to multiple UAV scenarios makes it a step ahead of other studies. Moreover, the hybrid genetic algorithm is the first work in the literature to be adapted to the problem of simultaneous deployment of trucks and multiple UAVs and to take into account the assumptions adopted.

This study fills the knowledge gaps in the existing literature and provides a new perspective on solving the simultaneous deployment problem of trucks and UAVs. The proposed hybrid genetic algorithm is of great importance in both academic and industrial circles for its potential and effective usability in practical applications. Therefore, the study is expected to make a valuable contribution to research and applications in this field.

First, a related work section was created to assess the existing body of knowledge in the literature. An overview of related work is presented, analyzing the gaps in the literature and existing solutions for the simultaneous deployment of UAVs and trucks. Then, in the methodology section, a detailed description of the hybrid genetic algorithm for solving the simultaneous distribution problem of trucks and UAVs is presented. The basic principles of the algorithm, the data sets used, parameter settings

and methodological details are discussed in detail. The differences and advantages of this solution from the existing literature are emphasized. Afterwards, the results section evaluates the performance of the developed algorithm. The degree of achievement of the objectives, the solution capability of the algorithm and the limitations encountered are analyzed in detail. The results obtained are evaluated by comparing them with other works in the literature. Finally, the conclusion section provides an overall evaluation of the study. A summary of the results obtained, the methodological and analytical achievements of the study, contributions to the literature and suggestions for future research are presented in this section.

8.2 Related Work

The use of UAVs in distribution services has been a rapidly expanding topic despite technical, legal and social challenges [80–85]. Advances in this field have led to a distribution model that goes beyond the traditional use of trucks. The simultaneous use of trucks and UAVs has emerged as a current research topic and complicates the route planning problem as it requires time and location coordination of the two vehicles [86]. In this regard, this thesis aims to overcome the limitations of existing solutions in the literature and develop more efficient methods.

In this new model where UAVs are integrated into distribution activities, the simultaneous distribution of a truck with multiple UAVs reflects a practical and possible scenario. Therefore, this scenario needs to be studied in detail from an operational perspective. The focus of this study is the development of a solution approach suitable for multiple UAV scenarios. As a solution-oriented approach, a hybrid genetic algorithm based on function estimation with machine learning is proposed to tackle the simultaneous deployment problem of trucks and UAVs. The results comparing this algorithm with the simulated annealing method used in previous studies show that the hybrid genetic algorithm achieves more effective results, especially for small and medium-sized problems. Moreover, the flexibility of this algorithm to be easily adapted to multiple UAV scenarios makes it a step ahead of other methods.

In this context, the development of drone-based delivery systems has gained great potential in recent years due to the high mobility and low cost of drones. As an example, Khosravi et al. [87] presents a survey focusing on important issues related to drone routing in drone-based delivery systems. By addressing three main drone routing aspects (route planning, charging, safety), this review highlights practical design considerations to ensure efficient, flexible and reliable package delivery. First, it discusses potential issues that may arise during the design of these systems. It then presents a new classification based on these three aspects. Using this classification, it provides a detailed review of each drone guidance algorithm in terms of key features and operational characteristics. It also compares these algorithms in terms of idea, advantages, limitations and performance. Finally, it presents open research challenges in order to stimulate further research in this area. As a different application area, Dinelli et al. [88] focuses on how autonomous robots and exploration systems

can be used in harsh environments such as underground mines. The study highlights the characteristics and challenges of environments that are difficult or impossible for humans to reach, such as underground and outer space, and draws attention to hybrid robotic systems that can be used in such environments. These systems, combining multiple agents such as Unmanned Ground Vehicles (UGVs) and UAVs, offer potential for underground exploration and mine emergency response. The paper discusses in detail the configurations, construction practices and hardware equipment of these hybrid systems.

Ribeiro et al. [89] discusses the importance of using UAVs in search and rescue missions in emergency and post-disaster scenarios. The study addresses the new challenges in the development of self-charging technologies and how to integrate these technologies into UAVs, which complicate the use of UAVs in such missions. The main focus of the study is on the coordinated use of UAVs and mobile charging stations. For this purpose, VRP with synchronized networks (VRPSN), a variant of a new vehicle routing problem (VRP), is defined. This problem requires routing UAVs to charging stations and synchronizing mobile charging platforms according to the movements of the UAVs. The study developed a mixed integer linear program model to solve this problem. It also presents a build-and-tune heuristic integrated with a genetic algorithm to overcome the computational limits of this model. An example application of the work was carried out at the Córrego do Feijão Mine in Brazil, showing that this method can be used as an effective planning method in search and rescue missions. This study makes an important contribution to support the effective use of UAVs in emergency and post-disaster scenarios. The research sheds light on new developments in this field by addressing the synchronized use of UAVs and mobile charging stations.

After discussing the importance of the collaboration of truck and drone technologies to overcome logistical challenges in rural areas, the cost-saving impact of the truck-drone system should be evaluated in detail to highlight the potential of the collaboration. In this context, Jiang et al. [90] investigates a multi-visit and flexible docking vehicle routing problem that combines a fleet of trucks and drones to meet pick-up and delivery demands in rural areas. Specific to this collaborative truck-drone system, each drone can serve multiple customers during the same trip, dock with different trucks, and perform pickup and delivery operations simultaneously. To address this complex scenario, the paper formulates the problem using a mixed integer linear programming model and solves it with an adaptive large neighborhood search metaheuristic. Numerical experiments show that the proposed truck-drone system achieves 34% cost savings compared to truck-only methods. Furthermore, the study evaluates in detail the effects of multiple visit services, flexible docking and simultaneous pick-up and delivery on the performance of the truck-drone system.

On the other hand, studies involving a novel algorithm for optimizing package deliveries through the collaboration of EVs and drones contribute to last-mile logistics and vehicle routing problems. For example, Mara et al. [91] sheds light on a new research direction in the context of last mile logistics, focusing on the collaboration between electric vehicles (EVs) and drones. The E-VRPD, referred to as the electric vehicle routing problem, aims to determine the optimal vehicle tour for the fastest

delivery of packages to customers in a scenario where a number of EVs are each equipped with a single drone. This paper addresses the limitations of existing techniques in solving such problems, emphasizing the importance of E-VRPD. In this context, a sequential decomposition algorithm for solving E-VRPD is proposed. This algorithm involves the development of a mathematical formulation for integrating drone sorties into EV tours. The study evaluates the proposed method on instances with 40 customers and 7 charging nodes, and the experimental results show that E-VRPD can be effectively implemented in practice.

The use of drones in delivery services has been a growing topic, but technical, legal and societal challenges need to be overcome. In particular, the development of a new model where trucks and UAVs make simultaneous deliveries makes the routing problem quite complex as it requires time and location coordination of both vehicles. Therefore, new algorithms need to be developed to solve this problem. In this paper, a hybrid genetic algorithm based on machine learning and function estimation is developed as a solution to the simultaneous distribution problem of trucks and UAVs. The results of this new approach are compared with the previously used annealing simulation method and show that the developed method provides better results for small and medium sized problems. This study aims to contribute to solving the truck and UAV simultaneous deployment problem more effectively.

This paper presents an in-depth review of the basic components, structure, uses, and scientific and industrial benefits of drone technology. In this context, we delve deep into the fundamental concepts and mathematical foundations of Hybrid Genetic Algorithms and Vehicle Routing Problems with Drones, showing how this integration has the potential to transform real-world scenarios. In exploring the complex balance between evolutionary algorithms and aerial logistics, we attempt to pave a path to shape the future of optimization in the transportation sector. To this end, the study aims to provide a new approach by improving solutions from the existing literature. With the increasing role and impact of UAVs in vehicle routing problems, the development of new algorithms to manage single drone scenarios more effectively is the foundation of research and development efforts in this field, and the importance of studies and innovations is increasing.

8.3 Methodology

In this section, a hybrid genetic algorithm based on machine learning and function estimation is developed to solve medium and large-scale VRP-D. In the first phase of the study, the process of determining the ground vehicle and UAV routes was carried out in a two-stage method. In the first stage, the ground vehicle route is determined by the genetic algorithm, while in the second stage, the UAV route is optimized. This two-stage approach aims to effectively evaluate the potential of the genetic algorithm on the ground vehicle route to optimize the UAV route. The methodology is designed in such a way that the genetic algorithm in the ground vehicle routing phase only generates routes that perform well. This comprehensive

and selective approach allows for a more efficient determination of ground vehicle routes through function estimation supported by machine learning. In the second step, the determination of the UAV route based on the optimized ground vehicle route forms the basis of the hybrid genetic algorithm. This step includes the scenario in which the ground vehicle makes simultaneous deliveries with multiple UAVs and successfully adapts the developed hybrid solution approach to multi-UAV scenarios.

In the proposed hybrid algorithm, in the first stage, ground vehicle routes are generated and in the second stage, the best UAV rounds are assigned to minimize the waiting time of the ground vehicle. In the first stage, the target delivery time of the ground vehicle is calculated, and in the second stage, the waiting time of the ground vehicle is obtained, and the sum of these two times constitutes the return time to the ground control station (GCS). In this approach, a genetic algorithm is used to generate the ground vehicle routes in the first stage. In the context of evolutionary algorithms, the population size needs to be chosen sufficiently large to avoid getting stuck in local solutions. In some scenarios, the computation of the fitness function can be quite costly. In such cases, the use of hybrid algorithms based on the estimated fitness function may be required.

Traditional package distribution processes are usually carried out by trucks. However, the limited speed of trucks and the fact that they are easily affected by terrain conditions limit the efficiency of this method. UAVs offer several advantages over trucks in package delivery. They are faster, do not require an operator, are not affected by traffic congestion and have lower transportation costs, all of which increase their potential to be preferred. However, due to technical barriers, the requirement to carry a single package per shipment and return to the warehouse limits the potential advantages of UAVs. To overcome these challenges and maximize the advantages of UAVs, a new deployment model is proposed. In this model, the vehicle fleet consists of two different vehicle types: trucks and UAVs. The UAV can be transported together with the truck or move separately and has to meet the truck again after each customer visit. This solution combines both the speed and cost advantages of the UAV and the truck's advantage of transporting various packages over long distances, providing an efficient distribution model.

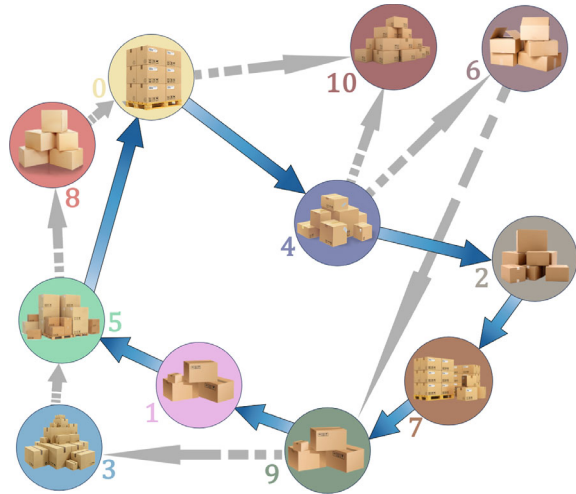
Derived from the vast literature on the vehicle routing problem, this new dispatching model presents an innovative approach that consists of a fleet of two different types of vehicles: trucks and UAVs. The UAV can either be transported together with the truck (i.e., the UAV is carried on the truck) or it can move independently of the truck. However, due to the UAV's battery life limitations, once it leaves the truck, it has to visit a single customer and meet the truck again at a different customer location. After the UAV leaves the truck at a customer location, it visits one or more customers and then heads to the customer location where it meets the UAV. The UAV can only obtain permission to leave and meet the truck at existing customer locations; this cannot be done at any location other than the customer. The truck is equipped with a system designed to place the customer's package each time the UAV leaves and to replace the UAV's battery each time it returns. In this context, the truck can be considered as a mobile warehouse for the UAV.

Some customer packages may only need to be delivered by truck, taking into account possible terrain conditions during the take-off and landing of the UAV, and situations where the UAV may not be able to deliver due to the payload capacity of the UAV. The main objective of this approach is to minimize the delivery time, with the last vehicle returning to the depot to complete the delivery. This unique distribution model can be considered as an evolution-ary extension of the vehicle routing problem and aims to increase the efficiency of distribution processes by enabling the interactive use of UAVs and trucks. The highlights of this model provide an alternative perspective to existing methods in the logistics and transportation sectors, enabling more effective and optimized management of distribution operations.

The proposed hybrid algorithm is based on generating truck routes in the first stage and assigning the optimal UAV rounds to minimize the waiting time of the truck in the second stage. In this two-stage process, the delivery time of the truck is determined in the first stage, while the second stage focuses on the waiting time of the truck and the sum of these processes reveals the overall delivery time of the vehicle returning to the depot. In the developed methodology, a genetic algorithm is used to generate truck routes in the first stage. This algorithm is one of the widely preferred meta-heuristics such as simulated an-nealing for routing problems. In this way, the optimization potential provided by the genetic algorithm is used to obtain an efficient solution in logistics distribution processes. The determination of truck routes with the genetic algorithm, which forms the basis of the approach, allows for more efficient and optimized distribution processes in logistics operations. At this point, the success of the genetic algorithm in routing problems provides an observable advantage over simulated annealing and other similar methodologies. This methodology aims to reduce operational costs and improve logistics performance by targeting the effective management of distribution processes in the logistics and transportation sectors. This innovative work has the potential to provide solutions to current challenges in the logistics industry and contribute to making future logistics operations more sustainable and effective.

Figure 8.1 shows an example solution to the VRP-D problem. In the figure, there are three different nodes and 2 different edges. The nodes are defined in different ways to distinguish the vehicles that visit them and the edges are defined in different ways to distinguish the vehicles that travel between the nodes. The node with a triangle is visited only by the UAV, the node with a circle is visited only by the ground vehicle, and the node with a square is visited by the ground vehicle traveling with the UAV. Solid dots indicate the ground vehicle route; dashed dots indicate the UAV flight. In the example, the vehicles leave the warehouse separately. The UAV takes off from the warehouse and after the delivery at Node 10, it travels to Node 4 to meet the ground vehicle. After leaving the warehouse, the ground vehicle goes directly to Node 4 to meet the UAV. Deliveries at the departure and rendezvous nodes are carried out by the ground vehicle. After leaving Node 4, the UAV picking up the package of customer 6 visits Node 6 and then travels to Node 9. The ground vehicle traveling from Node 4 to Node 9 also visits Nodes 2 and 7. After meeting at Node 9, the vehicles leave again to meet at Node 5 after the UAV has replaced its battery and loaded the package for customer 3. After Node 9, the ground vehicle visits Node

Fig. 8.1 Process for VRP-D solution



1 and arrives at Node 5. The ground vehicle returns to the warehouse after meeting the UAV at Node 5, while the UAV returns to the warehouse after visiting Node 8.

The assumptions in the study include the main constraints set for the VRP-D problem. First, it is not possible for the UAV to stay in the air and wait for the ground vehicle in order to reduce the charge consumption. This decision was taken to keep the UAV in the air for safety against external interference and to preserve battery life. Furthermore, this assumption limits the UAV's travel between the departure and rendezvous points, affecting the battery life not only during the tour but also during the ground vehicle's movements. Second, only one customer visit can take place in each UAV flight. This restriction ensures that the UAV focuses on a single service target in each of its tours. Third, the ground vehicle and UAV cannot meet at the point where the tour starts, and the departure and rendezvous points must be different. This constraint ensures efficient coordination of cooperating vehicles. Fourth, vehicles can only leave and meet at customer nodes. This prevents vehicles from interacting with the customers they serve and from meeting or leaving at points outside the depot. Fifth, a second visit to the same customer cannot take place. This restriction is important to increase customer satisfaction and the efficiency of deliveries. Finally, if a UAV tour ends at the warehouse, the vehicle is taken out of service and the UAV cannot be ventilated again. This assumption aims to avoid the impractical return of the UAV for missions outside the depot.

After the ground vehicle routes are generated with the genetic algorithm, the waiting times of the ground vehicle must be obtained in order to calculate the fitness value. This is because the objective function consists of the duration of the ground vehicle route and the waiting time of the ground vehicle. Since the calculation of the fitness function is quite costly in this approach, the approximate waiting times of the ground vehicle for the ground vehicle routes generated in each iteration are determined by using machine learning and estimating the fitness function, and the

UAV tour assignments are optimized only for a certain number of routes with the best fitness values. In each iteration, the exact solutions obtained are added to the training data, so that the difference between the exact fitness value and the approximate fitness value is reduced as the iterations progress. In this study, genetic algorithm is used to generate ground vehicle routes and extreme learning machine is used to estimate the fitness function. The hybrid genetic algorithm is summarized in Algorithm 1 and its details are explained in the subsections.

Algorithm 1. A hybrid algorithm based on eligibility estimation with machine learning

1	Training data, generate random ground vehicle route as big (N_t) of initial training data
2	Calculate the distribution time of each individual in the training data
3	Solve mixed-integer nonlinear programming (MNL) for each individual in the training data, get the best dwell time
4	Train the training data
5	Create the initial population by copying the best population size (N_p) route from the training data to the population with distribution and dwell times
6	Select random parents from the population and perform crossovers for the number of crosses (N_c) to be applied in each iteration, generating N_c children
7	Calculate the distribution time of each child individual produced
8	Using machine learning, estimate the fitness of each child produced and obtain approximate waiting times
9	Perform mutation with mutation probability (M_p)
10	if the sum of the distribution and approximate cooldowns improves as a result of the mutation, apply the mutation
11	Select the number of best children (N_b) individuals to be included in the population at the end of iteration, taking into account the approximate waiting time and the exact distribution time from the children
12	Solve MNL for selected N_b individuals, calculate exact wait times
13	Add these child individuals to the population and training data
14	Remove the worst N_b individuals from the population
15	Generate random ground vehicle routes for the number of randomly generated individuals to be included in the population at each iteration (N_{rhn})
16	Generate ground vehicle routes according to the nearest neighborhood as many as the number of individuals (N_{nhn}) created according to the nearest neighborhood in the population

(continued)

(continued)

Algorithm 1. A hybrid algorithm based on eligibility estimation with machine learning	
17	Calculate the distribution time for the N_{rhn} and N_{nhn} individuals produced
18	Solve MNLP for N_{rhn} and N_{nhn} individuals generated
19	Remove the worst $N_{nhn} + N_{nhn+1}$ individuals in the population
20	Add the generated N_{rhn} and N_{nhn} individuals to the population and training data
21	Apply local search to the best individual in the population
22	Add the individual obtained by local search to the population
23	Train the training data
24	if the number of iterations has reached the number of iterations in the Genetic algorithm (N_{gn})
25	Stop
26	if not
27	Go to step 6

The training data is generated in a randomized fashion. Each entry in the dataset represents a randomly generated ground vehicle route. When generating these ground vehicle routes, targets that will not be assigned to the drone are assigned to the ground vehicle routes. Then, for each ground vehicle route, the mixed-integer nonlinear programming (MNLP) model is analyzed and the best waiting times for each ground vehicle route are determined. In the mathematical model that minimizes the waiting time of the ground vehicle for the UAV, the first stage involves the process of determining the route of the ground vehicle and also includes the determination of the customers that the UAV will visit. This means that customers outside the ground vehicle route will be visited by the UAV. This routing process in the first phase covers the movements of the ground vehicle and the UAV's assigned customer visits. In the second stage, a mixed integer linear programming model is used to determine the customer visits of the UAV outside the ground vehicle route. This model is used to match the meeting and departure points of the UAV with the ground vehicle with the assigned customer visits. The second phase focuses on the interaction between the ground vehicle and the UAV to optimize customer visits.

At this stage, the ground vehicle route (\bar{r}_s) and the cluster of customers that the UAV will visit (D_s cluster) are known. The cluster \bar{r}_s , representing the ground vehicle route, and the cluster D_s , representing the nodes assigned to the UAV, were determined in the first iteration stage. Based on these two pieces of information, a cluster of tours that the UAV can perform can be created. The first of the three nodes that make up a UAV tour is "s", which is part of the ground vehicle route, the second is D_s , which represents the customer assigned to the UAV, and the third is the meeting point with the ground vehicle. The second node of the UAV tour should be the customer that

the UAV delivers to, so this node should be chosen from the cluster D_s . Also, the first node of each UAV tour must be visited before the third node, according to the order in the ground vehicle route. Another factor to consider during tour generation is the battery life. In order for the generated tours to be suitable tours that can be used in the solution, the tour duration should be shorter than the battery life. Also, the travel time of the ground vehicle between the start and end nodes of the tour should not exceed the battery life. Such UAV tours are critical to optimize the interaction between the ground vehicle and the UAV in logistics distribution processes. Factors such as battery life and tour durations play a decisive role in the effectiveness of the solution, emphasizing the important contribution of the study towards improving efficiency in logistics operations.

In the first stage, the duration of the ground vehicle route is calculated by determining the ground vehicle route. However, waiting situations that may be caused by UAV tours may increase the ground vehicle’s deployment time beyond the time determined in the first stage. For example, if the ground vehicle arrives at any of the designated meeting points first, the ground vehicle will not be able to continue its route and will have to wait for the UAV to replenish the UAV’s battery and load the next customer’s package. In this case, the ground vehicle route will take longer to complete, exceeding the time set in the first phase. The mathematical model was developed to assign the best UAV rounds by minimizing this waiting time of the ground vehicle and without increasing the completion time of the ground vehicle route as much as possible. In this context, since waiting situations can only occur at meeting points, the waiting times at the warehouse with the customers on the ground vehicle route are taken into account, as shown in Eq. 8.1.

$$Z = \min \sum_{i \in C \cup \{N+1\} / D_s} w_i \tag{8.1}$$

Equation 8.2 calculates the time the ground vehicle waits for the UAV at the meeting point. By this stage, the ground vehicle’s route has been determined, so it is known when the ground vehicle will arrive at which customer. The waiting time of the ground vehicle at node i is determined by calculating the difference between the arrival time at node i from node j , where they leave with the UAV, and the duration of the UAV’s tour starting at j and ending at i . The inner parenthesis in the equation indicates how long it takes the ground vehicle to cover the distance between node j and node i , where the tour p starts and ends at node i . If the UAV tour between i and j is longer than the time it takes the ground vehicle to cover this distance, the ground vehicle is put on hold; otherwise, the waiting time of the ground vehicle is set to 0.

$$\sum_{p \in P_s} \left\{ d_p l_{ip} x_p - \left(t_i l_{ip} x_p - \sum_{j \in C \cup \{0\} / D_s} t_j f_{jp} x_p \right) \right\} \leq w_i, \quad i \in C \cup \{N+1\} / D_s \tag{8.2}$$

In order for the tour assignments in the problem to be valid, certain conditions must be fulfilled. The first of these conditions is clearly stated in Eq. 8.3. As stated

in this equation, for each customer assigned to the UAV in the first stage, only one UAV tour must be selected in which these customers are intermediate nodes.

$$\sum_{p \in P_s} a_{ip} x_p = 1, \quad i \in D_s \tag{8.3}$$

Furthermore, as expressed in Eq. 8.4, each customer on the ground vehicle route, including the warehouse, can be selected as the start node of only one UAV tour (the node where the UAV leaves the ground vehicle).

$$\sum_{p \in P_s} f_{ip} x_p \leq 1, \quad i \in C \cup \{0\}/D_s \tag{8.4}$$

Similarly, each customer on the ground vehicle route, including the warehouse, can be selected as the end node (the node where the vehicles meet) in only one UAV round, as specified in Eq. 8.5.

$$\sum_{p \in P_s} l_{ip} x_p \leq 1, \quad i \in C \cup \{N + 1\}/D_s \tag{8.5}$$

As shown in Fig. 8.2, it defines two types of negative round assignments. The last constraint that prevents these two types of negative assignments is expressed in Eq. 8.6.

$$\sum_{p \in P_s} f_{mkp} x_p + \sum_{p \in P_s} l_{mkp} x_p \leq 2 \times \left(1 - \sum_{p \in P_s} f_{mip} l_{mjp} x_p \right),$$

$$i = 0, 1, \dots, n_s - 1, \quad j = i + 2, \dots, n_s + 1, \quad k \in H, \quad H = \{h | i < h < j\}, \quad i \neq j \tag{8.6}$$

This constraint guarantees that the UAV does not start another tour before completing a tour. For example, at iteration s , n_s customers are assigned to the ground vehicle route. If a UAV tour is selected that starts at a customer at position

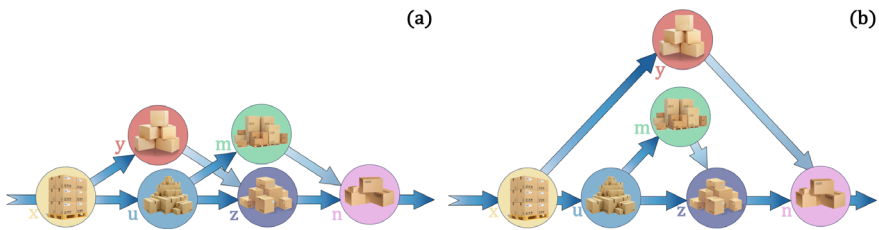


Fig. 8.2 **a** Example of a type 1 no-go UAV tour assignment, **b** example of a type 2 no-go UAV tour assignment

i and ends at a customer at position j on the ground vehicle route, no UAV tour can be selected that starts or ends at a customer at any position k between these two positions. Definitions of variables are given in Eqs. 8.7 and 8.8.

$$x_p \in \{0, 1\}, p \in P_s \quad (8.7)$$

$$wt_i \geq 0, x \in C \cup \{N + 1\}/D_s \quad (8.8)$$

Ground vehicle routes and the best waiting times of these routes constitute the input and target values for machine learning. The initial population is formed by selecting the best routes from the training data up to the population size. Using the overlearning machine, the weights to be used for function estimation are calculated for the training data. Then, a certain number of children are generated by crossover. Mutation is applied with a certain probability and the approximate fitness value of each generated child is calculated by the fitness function estimation. While the fitness estimation is done for the waiting time of the generated child, the duration of the ground vehicle route is calculated precisely. Then, the exact waiting time and the exact solution value are obtained by solving MNLP for a certain number of child individuals with the lowest approximate solution value, which is defined as the sum of the exact route time and the approximate waiting time. The best selected child individuals and their exact solution values are added to both the initial population and the training data. After the best children are added to the initial population, the population size is kept constant by removing the worst solution as many times as the number of children added. The training data grows with the input and target values of the children added at each iteration. As new data is added, the learning process is repeated periodically, taking into account the current training data. When the learning phase is completed, new child individual genes are obtained through crossover and these iterations continue until the stopping criteria are met.

The genetic algorithm is based on the principle that individuals with good performance reproduce to form new generations and weak individuals are eliminated through natural selection. The proposed genetic algorithm is used for VRPD solution. In this algorithm, each chromosome represents a ground vehicle route. Chromosomes are composed of genes and the number of genes in a chromosome is determined randomly. VRPD sets a lower bound for the number of genes that can be present in any chromosome. The fact that the number of target regions that can be assigned to the drone depends on an upper limit affects the number of target regions that can be assigned to the ground vehicle. Considering the maximum number of targets that can be assigned to the drone, P_{\max} , at least a $-P_{\max}$ number of target nodes must be assigned to any chromosome. The ground vehicle route always starts and ends at the GCS node. Since there are no GCS nodes at the beginning and end of the ground vehicle route, these nodes are not included in the chromosome structure. This does not affect the operation of the algorithm and the computations.

The initial population is randomly generated, taking into account the presence of targets that cannot be assigned to the drone. In particular, target regions that

cannot be visited by the drone due to package weight or size are taken into account. These target regions are necessarily visited by the ground vehicle and are included in the chromosome representing each ground vehicle route. When creating the initial population, the drone assignments are decided first. First, a random number between 1 and P_{\max} is generated to determine the number of target regions to be assigned to the drone and this number is called " d_r ". Then, in the second step, the following steps are repeated " d_r " times. In the second step, first, a random number is generated in the range aP , which is the number of targets that can be assigned to the drone. In the second step, a random number is generated in the range aP , which is the number of targets that can be assigned to the drone, and this number is called " d_s ". Secondly, " d_s " indicates the number of nodes in KP, the set of targets that can be assigned to the drone. Third, from the set of targets that can be assigned to the drone, KP, the element with index " d_s " is removed, indicating that a target has been removed from the set of targets to be assigned to the drone. Fourthly, aP , the number of targets that can be assigned to the drone, is decreased by one and $aP = aP - 1$ is updated. Finally, this process is repeated " d_r " times from the beginning of the second step and this number is called " d_s ". Secondly, " d_s " indicates the number of nodes in KP, the set of targets that can be assigned to the drone. Third, from the set of targets that can be assigned to the drone, KP, the element with index " d_s " is removed, indicating that a target has been removed from the set of targets to be assigned to the drone. Fourthly, aP , the number of targets that can be assigned to the drone, is decreased by one and $aP = aP - 1$ is updated. Finally, this process is repeated " d_r " times from the beginning of the second step.

As mentioned in step 16 of Algorithm 1, N_{nhn} of the individuals in the initial population are generated according to the nearest neighbor. This process starts by first determining the number of nodes to be assigned to the drone in the solution. Then, the nodes to be assigned to the drone are randomly selected. Then, the customers to be assigned to the ground vehicle are assigned according to the nearest neighborhood and ground vehicle routes are generated. The aim is to represent as short as possible ground vehicle routes and target visits in the initial population.

The crossover operator is included in the proposed algorithm as a technique for generating new children. In this algorithm, the single-point crossover method is particularly preferred. As shown in Fig. 8.3, this form of crossover starts with the random selection of two different parental chromosomes. Then, a random breakpoint is chosen. The genes to the left of the breakpoint are removed from the first parental chromosome and transferred to the first child chromosome. Upon completion of this step, the genes transferred to the first child chromosome are removed from the second parent chromosome. The remaining genes are removed from the second parent chromosome and added to the child chromosome. The reason why these steps are performed sequentially is to prevent the number of genes of the offspring from falling below the lower limit. The same steps are followed to produce the second child and the crossover process is repeated N_c times.

Mutation is used in the genetic algorithm to avoid getting stuck in local best solutions and to provide genetic diversity. In this context, three different mutation operators were applied. In the proposed algorithm, the offspring of the crossover is

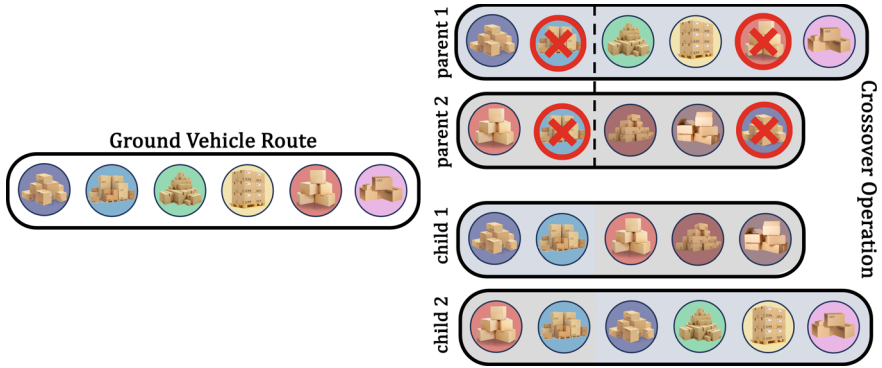


Fig. 8.3 Crossover operation in hybrid genetic algorithm

mutated based on a predetermined probability value. If mutation is to be performed, the type of mutation operator to be used on the child is determined based on the predetermined probabilities. An approximate fitness function is calculated for the new chromosome obtained as a result of mutation, and if the mutation has the effect of increasing the fitness value, the change is considered permanent. However, if there is no improvement in the fitness value, the change made in the mutation phase is undone.

In 1-1 substitution, as shown in Fig. 8.4, the mutation operator, called the first operator, operates by randomly selecting two different positions on the chromosome to be mutated. In this operator, the gene in the first position is placed in the second position, while the gene in the second position is placed in the first position. At the end of these steps, it is seen that two different targets on the relevant ground vehicle route are displaced.

As a result of completing these steps, it is seen that two different targets on the relevant ground vehicle route are replaced. In addition, as shown in Fig. 8.5, in one addition, the second operator aims to assign this target to the ground vehicle by changing the assignment of one of the customers assigned to the drone on the solution. For this operator, a target not found on the chromosome is randomly selected and added to a random position in the ground vehicle route. This process causes the

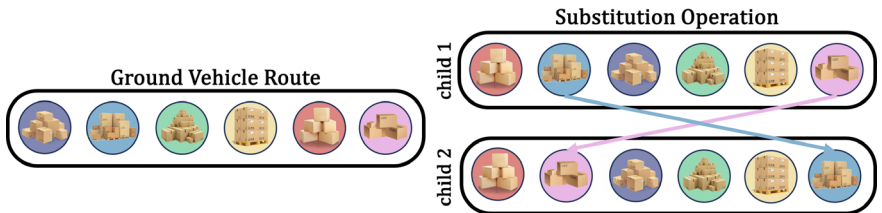


Fig. 8.4 Substitution operator in hybrid genetic algorithm

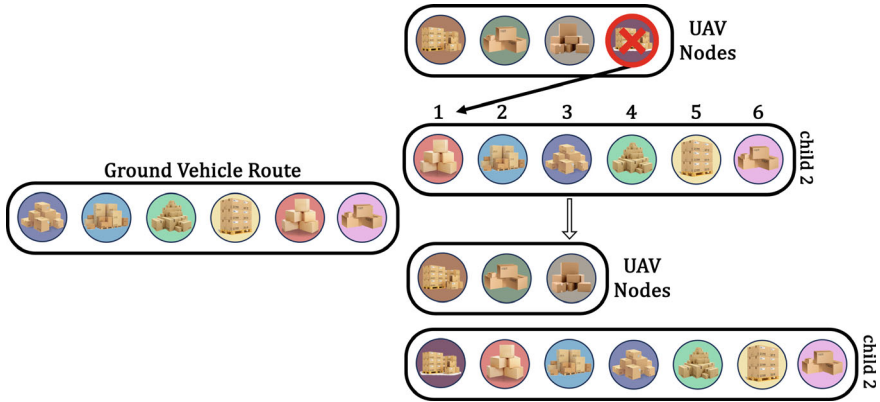


Fig. 8.5 Addition operator in hybrid genetic algorithm

number of targets on the ground vehicle route to increase by one. However, in order for this operator to be applied, more than one target must be assigned to the drone.

However, as shown in Fig. 8.6, in 1 subtraction, the third operator aims to remove one of the targets assigned to the ground vehicle and assign it to the drone. In this operator, a gene is randomly selected from the genes in the chromosome. As a result of this step, the number of targets on the ground vehicle route is reduced by one. However, in order to apply this operator, the ground vehicle should be assigned as small a number of targets as possible. Also, it is important to ensure that the gene to be extracted is not one of the targets that cannot be assigned to the drone, i.e. targets that should be assigned to the ground vehicle. Otherwise, the accuracy of the algorithm is compromised.

In the framework in which the hybrid algorithm is developed, the fitness value is calculated both approximately and precisely in different steps. In contrast to the exact solution approach, instead of generating all possible ground vehicle routes, the hybrid algorithm only generates the routes that can produce good results and calculates the drone dwell times of these routes. However, determining the best drone tour assignments for each ground vehicle route based on the mathematical model discussed in the previous section can require long computation times. Therefore, the approximate waiting times for each ground vehicle route generated by the genetic algorithm are calculated by function estimation. For this purpose, an extreme learning machine is used. The approximate waiting time is combined with the delivery time of the ground vehicle to obtain an approximate fitness value. From the solutions with this approximate fitness value, the exact waiting times are obtained by using the MNLP solution for a certain number of solutions with the lowest approximate fitness value. The exact waiting time and the deployment time of the ground vehicle determine the exact fitness value. In this context, the fitness value of each individual in the population reflects the exact fitness value and the selection process is based on the exact fitness values. In order to improve the performance of the hybrid algorithm, a local search strategy is incorporated into the algorithm. In each iteration, one of

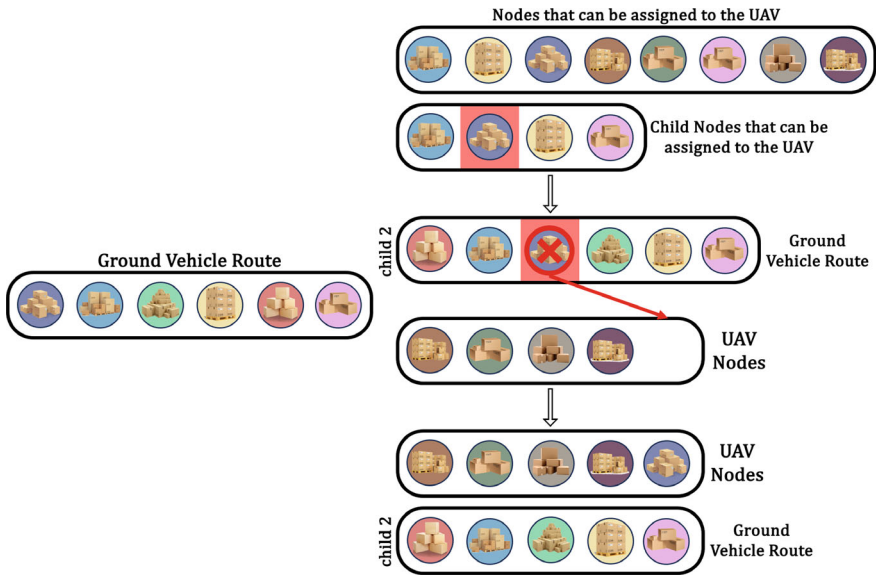


Fig. 8.6 Subtraction operator in hybrid genetic algorithm

three different operators is randomly selected and applied to the best solution. For each operator, the probabilities of selection are determined and the selection process is based on these predetermined probabilities. A new individual generated as a result of the local search joins the population by replacing the worst solution from the population, regardless of whether the fitness value improves or not.

The basis of the proposed hybrid algorithm is function estimation when calculating the approximate fitness values of the solutions generated by the genetic algorithm. This function estimation is performed using an extreme learning machine. The slow running speed of traditional feed-forward learning machines is usually due to the iterative adjustment of parameters between the input and output layers. Therefore, the developed feed-forward and non-iterative extreme learning machine is proved to work faster than traditional learning machines such as back-propagation algorithm and support vector machine. Another feature that distinguishes the extreme learning machine from other traditional methods is that it has a single hidden layer and the parameters of this hidden layer are randomly assigned. In the context of this study, the overlearning machine is preferred for function estimation because it is easy to implement and fast due to its non-iterative nature.

In order to improve the performance of the hybrid algorithm, a local search strategy is implemented within the algorithm. In each iteration, one operator is randomly selected from the three operators to be used in the mutation phase. For each operator, the probability of selection is determined and the selection is performed according to the predetermined probabilities. The new individual obtained as a result of the local

search is added to the population by replacing the worst solution in the population, regardless of whether the fitness value improves or not.

Up to this point in the paper, VRP-D examines the coordination of a single ground vehicle and a single drone. However, in the designed VRP-D scenario, it is important to consider a real-life situation where the delivery ground vehicle transports multiple drones and simultaneously delivers to multiple drones. Therefore, it is important to evaluate this scenario from an operational perspective. The problem with a single ground vehicle and multiple drones making simultaneous deliveries will be referred to as the “Multiple Drone Assisted Vehicle Routing Problem” and abbreviated as “VRP-mD” in the rest of the paper.

When examining the coordination of a single ground vehicle and a single UAV, VRP-D does not take into account real-life scenarios, such as a delivery ground vehicle transporting multiple UAVs and simultaneous distribution by multiple UAVs. However, it is important to consider these scenarios from an operational perspective, as the use of multiple UAVs in the deployment process can offer potential advantages in terms of effectiveness and efficiency. In this context, a scenario where a single ground vehicle and multiple UAVs deploy simultaneously can be referred to as the “Multi-UAV Assisted Vehicle Routing Problem”, abbreviated as “VRP-mD”. This problem implies that distribution operations are more complex and need to be optimized. By studying this new scenario, researchers can focus on developing more effective and efficient distribution strategies in real-world conditions.

In terms of the vehicle fleet, VRP-mD consists of two different vehicle types, ground vehicles and UAVs. The size of the vehicle fleet is expressed as $M + 1$, where M is the maximum number of UAVs that can be transported at the same time. As in VRP-D, UAVs can move with the ground vehicle or separately. The UAVs leave the ground vehicle for a new flight only at customer locations, and after leaving the ground vehicle, they visit a single customer for battery life reasons and have to meet the ground vehicle again at a different customer location. Before each flight, the UAVs’ batteries are replenished and the next customer’s package is loaded. The ground vehicle leaving the UAV may visit one or more customers during this time. During this process, the UAVs that left the ground vehicle first may meet up with the ground vehicle, or other UAVs traveling on the ground vehicle may leave the ground vehicle. The UAVs move independently of each other, i.e. each UAV has to meet the ground vehicle to replenish its battery and load its package, and wait for the ground vehicle if necessary, but there is no need for the UAVs to wait for each other. At a given node, the ground vehicle has to wait for all UAVs that need to meet at that node. Therefore, different departure scenarios are possible at the departure and rendezvous points, depending on the order of arrival of the vehicles. In a scenario with 2 UAVs, the departures in each arrival scenario are as follows:

In Scenario 1 (Ground Vehicle, UAV 1, UAV 2), the ground vehicle arrives at the meeting node first, followed by UAV 1 and UAV 2 respectively. The ground vehicle changes the battery of UAV 1 and loads the next customer’s package. If UAV 1 is going to take off for a new flight from this node, it starts its flight without waiting for UAV 2. If it is a ground vehicle, it has to wait for UAV 2 at the node where it is located. When UAV 2 arrives at this node, after completing its service (battery replacement,

etc.), the ground vehicle and UAV 2 move at the same time. In Scenario 2 (UAV 1, Ground Vehicle, UAV 2), UAV 1 arrives at the rendezvous node first. UAV 1 has to wait for the ground vehicle to change its battery and load the package of its next customer. Only after the arrival of the ground vehicle can it take off for a new flight. After UAV 1 takes off, the ground vehicle continues to wait. After UAV 2 arrives, the ground vehicle and UAV 2 leave the node at the same time, just as in Scenario 1. In Scenario 3 (UAV 1, UAV 2, Ground Vehicle), the ground vehicle arrives after the UAVs. In this case, both vehicles have to wait for the ground vehicle. After the ground vehicle arrives, the batteries of both UAVs are changed, their packages are loaded and the vehicles all leave at the same time. These scenarios clearly illustrate the coordination of vehicles in the VRP-mD problem and the separation and rendezvous order according to different scenarios. In this way, it can be seen that the VRP-mD addresses various scenarios and ensures the coordination of tools in a certain order.

Figure 8.8 explains in detail how a modification to the example solution shown in Fig. 8.7 can disrupt plausibility. In the solution in Fig. 8.7, the tour 0-15-2 assigned to UAV 2 is changed to 0-15-1 in Fig. 8.8. In this case, it is clear that there is some reduction in the ground vehicle's route, because there are fewer stops on the new route (0-1) than on the old route (0-1-2). The probability that the ground vehicle waits for UAV 2 at the end of round 0-15-2 (if the ground vehicle arrives at node 2 after UAV 2), and the probability that this waiting increases if UAV 2 waits at round 0-15-2, depends on the duration of flight 0-15-1. The probability that the ground vehicle waits at node 1 is shown in light gray. If the ground vehicle has to wait at this node, there is a risk of exceeding the battery life. If the ground vehicle is waiting at node 3, the probability of the ground vehicle waiting for the UAV decreases, unlike in the 0-15-2 round. This is indicated by node 3 in Fig. 8.8, shown in dark gray. At other nodes, changes in dwell times can have an impact on the travel time of the vehicles and therefore on the battery life. For example, changes at nodes 3 and 6 can lead to an increase in waiting time at node 7 and an increase in travel times if the ground vehicle is waiting at nodes 4 and 7. This may increase the risk that the ground vehicle will not be able to complete the distance between nodes 3-6 and nodes 6-8 before the battery life runs out, and may disrupt availability. This can happen not only in one round, but also in all three rounds. In the case of UAV 2 waiting, the probability of the ground vehicle waiting varies depending on changes in the ground vehicle's route. This requires all affected nodes to be considered for the calculation of the fitness function, which increases the cost.

It is clear that the increase in the number of UAVs will make it difficult to apply local search-based heuristics for single UAV deployment under the current assumptions, and the feasibility may deteriorate. In this case, the proposed hybrid genetic algorithm based on function estimation with machine learning is suitable for solving VRP-mD under the current assumptions, since it determines the ground vehicle route in the first stage and determines the UAV flights precisely in the second stage. The only modification required to apply the proposed hybrid genetic algorithm to VRP-mD is to adapt the mathematical model solved in the second stage of the algorithm to multiple UAV tour assignments. In this adaptation process, the 2-stage iterative

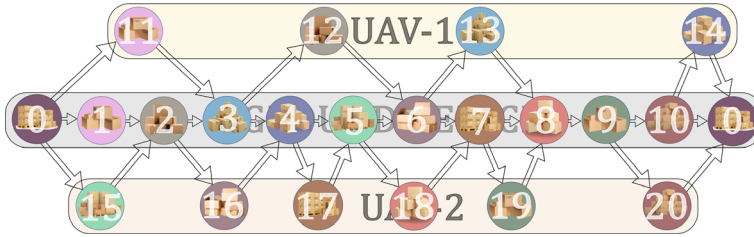


Fig. 8.7 Visualization of the VRP-2D solution

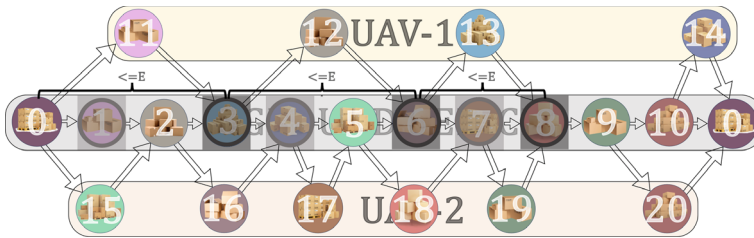


Fig. 8.8 Effects of the modification of the VRP-2D solution

exact solution algorithm is adapted to handle the distribution problem where a single ground vehicle and multiple UAVs make simultaneous deliveries.

In the first step of the 2-step algorithm for the VRP-D problem, the ground vehicle route and hence the customer nodes assigned to the UAV are determined. In this step, once the ground vehicle route is determined, the delivery time is known. However, waiting situations that may be caused by UAV rounds may make the delivery time of the ground vehicle longer than the one determined in the first stage. As a result, the ground vehicle route will take longer to complete than in the first phase. In the second stage, in order to minimize the waiting time of the ground vehicle, the best UAV tour assignments are made considering the ground vehicle route and UAV nodes determined in the first stage. The algorithm starts with the smallest ground vehicle route and then continues by refining the lower and upper bounds. When a shorter ground vehicle route cannot be obtained to the current best solution, the algorithm is terminated by closing the difference between the local lower bound and the global upper bound.

In the case of ground vehicle coordination with a single UAV, waiting times at the warehouse with customers on the ground vehicle route can be taken into account, taking into account that the waits will only be at the meeting points. However, if the ground vehicle coordinates with multiple UAVs, the situation becomes a bit more complicated. In this case, the rendezvous points where the waits take place may coincide with intermediate nodes in other UAV rounds. For example, in the example solution in Fig. 8.7, at nodes 2, 3, 4, 5, 6 and 7, the ground vehicle meets one UAV while the other UAV continues its flight. In this case, the travel time of the ground

vehicle between the departure and rendezvous nodes with the other UAV may be longer. However, these waiting times should not exceed the battery life of the other UAV. For example, due to the waiting time of the ground vehicle at node 2, the time between the departure of UAV 1 from node 0 and the arrival of the ground vehicle at node 3 should not be greater than the battery life of the UAV. Therefore, the mathematical model is expressed in a way that optimizes the rounds of multiple UAVs.

In the multi-UAV scenario, the objective function is defined as the arrival time of the last vehicle to arrive at the depot. Since the ground vehicle has to wait for all UAVs to arrive at a node to move from that node, the objective function in Eq. 8.9 aims to minimize the earliest time at which the ground vehicle can leave the depot.

$$\min[tmax_{N+1}^0] \quad (8.9)$$

In order for the tour assignments in the problem to be feasible assignments, certain conditions must be met. The first of these conditions is given in Eq. 8.10. As can be seen from the equation, each customer assigned to a UAV in the first stage has to be assigned to only one UAV. This condition is important as a fundamental constraint to ensure the appropriateness of tour assignments.

$$\sum_{v=1}^M \sum_{p \in P_s} a_{ip} x_p^v = 1, \quad i \in D_s \quad (8.10)$$

Furthermore, as shown in Eq. 8.11, each node on the ground vehicle route can be selected as the start node of up to one UAV tour on each UAV's route, including the depot. This node can be the start, end or intermediate node in tours assigned to other UAVs. This constraint is intended to coordinate and harmonize tour assignments by ensuring that the starting point of each UAV tour is connected to a specific node on the ground vehicle route.

$$\sum_{p \in P_s} f_{ip} x_p^v \leq 1, \quad i \in C \cup \{0\} \setminus D_s, \quad v = 1, 2, \dots, M \quad (8.11)$$

As expressed in Eq. 8.12, each node on the ground vehicle route can be selected as the end node for at most one UAV tour on each UAV's route, including the depot. This rule is intended to organize and harmonize tour assignments by ensuring that the end point of each UAV tour is connected to a specific node on the ground vehicle route.

$$\sum_{p \in P_s} l_{ip} x_p^v \leq 1, \quad i \in C \cup \{N+1\} \setminus D_s, \quad v = 1, 2, \dots, M. \quad (8.12)$$

Another constraint necessary to ensure the UAVs are assigned to a feasible tour is given in Eq. 8.13. This constraint ensures that each UAV does not start another

tour before completing a tour. Given “ n_s ” customers on the ground vehicle route, when a UAV tour is selected that starts at a customer at position “ i ” of the ground vehicle route and ends at a customer at position “ j ”, a UAV tour cannot be selected that starts or ends at any “ k ” position between these two positions. This constraint prevents overlapping UAV tours and ensures that tour assignments are appropriate and logical.

$$\sum_{p \in P_s} f_{m_k p} x_p^v + \sum_{p \in P_s} l_{m_k p} x_p^v \leq 2 \left(1 - \sum_{p \in P_s} f_{m_i p} l_{m_j p} x_p^v \right)$$

$$i = 0, 1, \dots, n_s - 1, \quad j = i + 2, \dots, n_s + 1, \quad v = 1, 2, \dots, M,$$

$$k \in H \quad H = \{h | i < h < j\}, \quad i \neq j \quad (8.13)$$

Equation 8.14 calculates the arrival time of the ground vehicle at the next node on the route, departing from the node at position “ i ”. In this equation, if any flight has started at the node at position “ j ”, a time equal to S_L (Service Time) will be spent. This is important to determine the arrival time by taking into account the service time between the nodes on the ground vehicle’s route.

$$t_{m_j+1}^k \geq tmax_{m_j}^k + \text{dist}_{m_{j+1}, m_j} + S_L \sum_{p \in P_s} f_{m_i p} x_p^v,$$

$$j = 0, \dots, n_s, \quad k = 0, \quad v = 1, \dots, M \quad (8.14)$$

Using the method in Eq. 8.15, when vehicle “ v ” is assigned a tour starting at node “ i ” and ending at node “ j ”, the arrival time at “ j ” is calculated by adding the flight time to the earliest time at which vehicle v can take off from node i . At the departure node, as specified in Eq. 8.15, as much time as S_L (Service Time) will be spent. This is an important constraint used when calculating the arrival time, taking into account the processing time at the start and end nodes of each round.

$$t_{m_j}^k \geq tmax_{m_i}^v + (d_p + S_L) \sum_{p \in P_s} f_{m_i p} l_{m_j p} x_p^v,$$

$$j = 1, \dots, n_s + 1, \quad i < j, \quad i \cup 0, \quad v = 1, 2, \dots, M \quad (8.15)$$

Equation 8.16 states that the ground vehicle must wait for the arrival of each UAV it will meet at this node before leaving the node at position “ j ”. The ground vehicle can leave this node after spending the S_R (Service Waiting) time waiting for the arrival of the UAVs.

$$tmax_{m_j}^k \geq t_{m_j}^v + S_R \sum_{p \in P_s} l_{m_j p} x_p^v, \quad j = 1, \dots, n_s + 1, \quad k = 0, \quad v = 1, 2, \dots, M$$

$$(8.16)$$

Equation 8.17 states that the departure time of vehicles from any node cannot be earlier than the arrival time at that node. According to this rule, if a rendezvous takes place at a node at position “j”, the S_R time must be added to the departure time of the vehicle from that node. UAVs cannot move for a new flight from any node on the ground vehicle route before meeting the ground vehicle.

$$t\max_{m_j}^k \geq t_{m_j}^v + S_R \sum_{p \in P_s} l_{m_j p} x_p^v, \quad j = 1, \dots, n_s + 1, \quad v \in V \quad (8.17)$$

Therefore, as expressed in Eq. 8.18, in order for the vehicle to leave this node, it must first wait for the ground vehicle to arrive at that node and then spend the S_R time.

$$t\max_{m_j}^v \geq t_{m_j}^k + S_R \sum_{p \in P_s} l_{m_j p} x_p^v, \\ j = 1, \dots, n_s + 1, \quad k = 0, \quad v = 1, 2, \dots, M \quad (8.18)$$

As indicated in Eq. 8.19, the vehicles leave the depot at the same time and at time 0. In cases where the UAVs are flying from a starting point (position i) to an end point (position j) on the ground vehicle route, they must meet the ground vehicle before running out of battery.

$$t_{m_j}^k - t\max_{m_i}^v \leq E + M \left(1 - \sum_{p \in P_s} f_{m_i p} l_{m_j p} x_p^v \right), \\ j = 1, \dots, n_s + 1, \quad i < j, \quad i \cup 0, \quad k = 0, \quad v = 1, 2, \dots, M \quad (8.19)$$

In cases where the UAV starts its flight from the node at position i, the ground vehicle can continue to wait for another UAV at the same node. If the UAV reaches the rendezvous point before the arrival of the ground vehicle, it has to wait for the ground vehicle in the air. Therefore, it is not enough for the UAV’s flight to be shorter than the battery lifetime; the time between the UAV’s take-off and its rendezvous with the ground vehicle, i.e. the time in the air, must also be shorter than the battery lifetime. This constraint also limits the travel time of the ground vehicle between the UAV and its departure and reunion nodes. Equation 8.20 represents the variable definition used for the tour assignments to the UAVs. This equation represents a mathematical expression used in the problem to assign UAVs to specific tours.

$$x_p^v \in \{0, 1\}, \quad p \in P_s, \quad v = 1, 2, \dots, M \quad (8.20)$$

The decision variable definitions for arrival times at nodes in the ground vehicle route are presented in Eq. 8.21. This equation provides a mathematical description of the variables that express when vehicles will arrive at specific nodes.

$$t_{m_j}^v \geq 0, j = 1, \dots, n_s + 1, v \in V \quad (8.21)$$

The decision variable definitions for the exit times from the nodes in the ground vehicle route are found in Eq. 8.22 This equation provides the mathematical expression of the variables that specify when the vehicles will leave certain nodes.

$$tmax_{m_j}^v \geq 0, j = 0, \dots, n_s + 1, v \in V \quad (8.22)$$

8.4 Results

In this section, in order to evaluate the performance of the heuristic method, comparison results with the studies in the literature are presented. Along with the obtained results, the parameter selection and related assumptions used in the numerical study are also explained in detail. The packet loading times at the departure nodes and battery replacement times at the rendezvous nodes are neglected in the solutions obtained using the proposed exact solution algorithm and the hybrid genetic algorithm. Given that the vehicles wait for each other at the rendezvous node and depart from this node at the same time, these service times are taken into account when generating each UAV tour produced in the mathematical model solution phase of the hybrid genetic algorithm. In this context, service times are added to the ground vehicle waiting time at each meeting node. In the example problems, customers are distributed in a homogeneous area of 32.1868 km \times 32.1868 km. Ground vehicle speed is assumed as 56.32 km/h, UAV speed as 80.47 km/h and battery life as 24 min. Furthermore, the service times were set as 40 s at the departure node and 30 s at the joining node. The distances were calculated considering the Euclidean scale. The hybrid genetic algorithm was coded in MATLAB and IBM ILOG CPLEX Optimization Studio was used for the mathematical models. All experiments were performed on a laptop with an Intel Core i7 processor and 16 GB RAM.

The results present the results of experiments on distribution networks with different number of objectives. The ‘‘Purpose’’ column in Fig. 8.9a shows the average of the best results obtained after 10 experiments with the proposed hybrid genetic algorithm and the annealing simulation. The ‘‘purpose_{TSP}’’ column shows the distribution time obtained with the TSP solution. In addition, in Fig. 8.9, the healing percentages are presented for three different methods. The first column (healing_{HGA}) and the second column (healing_{AS}) represent the improvement of the proposed algorithm and the annealing simulation method compared to the travelling salesman person. The last column shows the differences between the improvements achieved by both methods.

Figure 8.9 is sorted by the number of customers. It is clear from these results that the Hybrid Genetic Algorithm (HGA) achieved the best result in four out of five problem cases and outperformed the Annealing Simulation (AS) method in the remaining one problem. The AS method achieved the best result in three problem

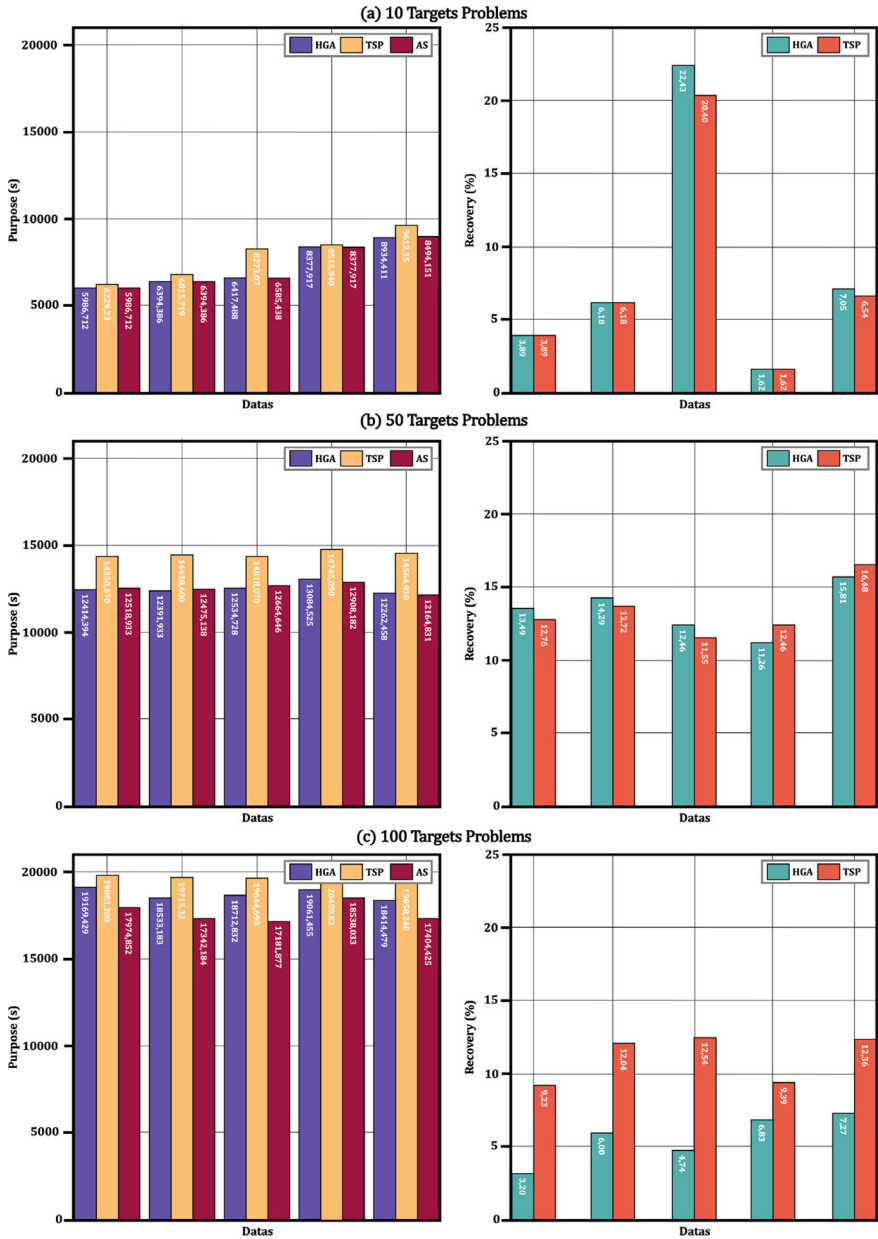


Fig. 8.9 Comparison of hybrid genetic algorithm with the a purpose and b recovery rates of annealing simulation

cases. Moreover, the average solution time of HGA is less than half a minute. These results show that the performance of HGA is superior to that of AS, especially for large data sets.

When comparing the Hybrid Genetic Algorithm (HGA) and annealing simulation (AS) methods for 50-target scenarios in medium-sized problems, HGA achieved better results in three of the five different problem cases, while AS achieved better results in the other two cases. Using both methods, the drone-assisted deployment improved the deployment time by more than 10% compared to the TSP solution. In this context, the average solution time of HGA was observed to be approximately 300 s for problems with 50 targets. Figure 8.10 shows the average time (in seconds) and total time (in seconds) to achieve the best result in the experiments with both methods.

It is seen that in large-sized problems and scenarios with 100 targets, HGA's solution times reach up to 30 min and exceed this time in some problems. This increases the difficulty of completing the iterations required to achieve better results. Indeed, the results obtained after 800 iterations show longer deployment times compared to AS. However, on average, 6% shorter deployment times are achieved compared to the TSP solution. The increased solution times of HGA on large datasets reflect the difficulty of addressing more complex problems, showing that drone-assisted deployment provides a distinct advantage over TSP in such scenarios.

During the parameter selection during the repetitions where these results were obtained, during the solution of problems with 10 and 50 customers; population size (N_p) is 50, the number of individuals formed according to the nearest neighbor in the population (N_{nhn}) is 10, the size of the initial training data (N_t) is 50, the number of crossovers to be applied in each iteration (N_c) is 25, the number of iterations (N_{gn}) is 800, number of best children (N_b) 5, number of randomly generated individuals to be included in the population at each iteration (N_{rhn}) 2, number of individuals generated according to the nearest neighbor to be included in the population at each iteration (N_{nhnb}) 3, mutation probability (M_p) was determined as 0.5 and the number of nodes in the hidden layer was determined as 100. Similarly, when solving problems with 100 customers: population size (N_p) is 50, the number of individuals formed according to the nearest neighbor in the population (N_{rhn}) is 10, the size of the initial training data (N_t) is 50, the number of crossovers to be applied in each iteration (N_c) is 25, number of iterations (N_{gn}) 800, number of best children (N_b) 5, number of randomly generated individuals to be included in the population at each iteration (N_{rhn}) 1, number of individuals generated according to the nearest neighbor to be included in the population at each iteration (N_{nhnb}) was determined as 2, mutation probability (M_p) was determined as 0.5 and the number of nodes in the hidden layer was determined as 100. It is clearly seen here that as the problem size increases, solution times increase dramatically. During the learning process of the algorithm, retraining the data with the exact solution of new individuals puts an additional burden on the algorithm in terms of time. Therefore, instead of repeating learning during each iteration, the learning process is restarted every 20 iterations.

The evaluation of the heuristic results of VRP-mD is based on an experimental preparation similar to VRP-D. However, the 10-customer problems are too small

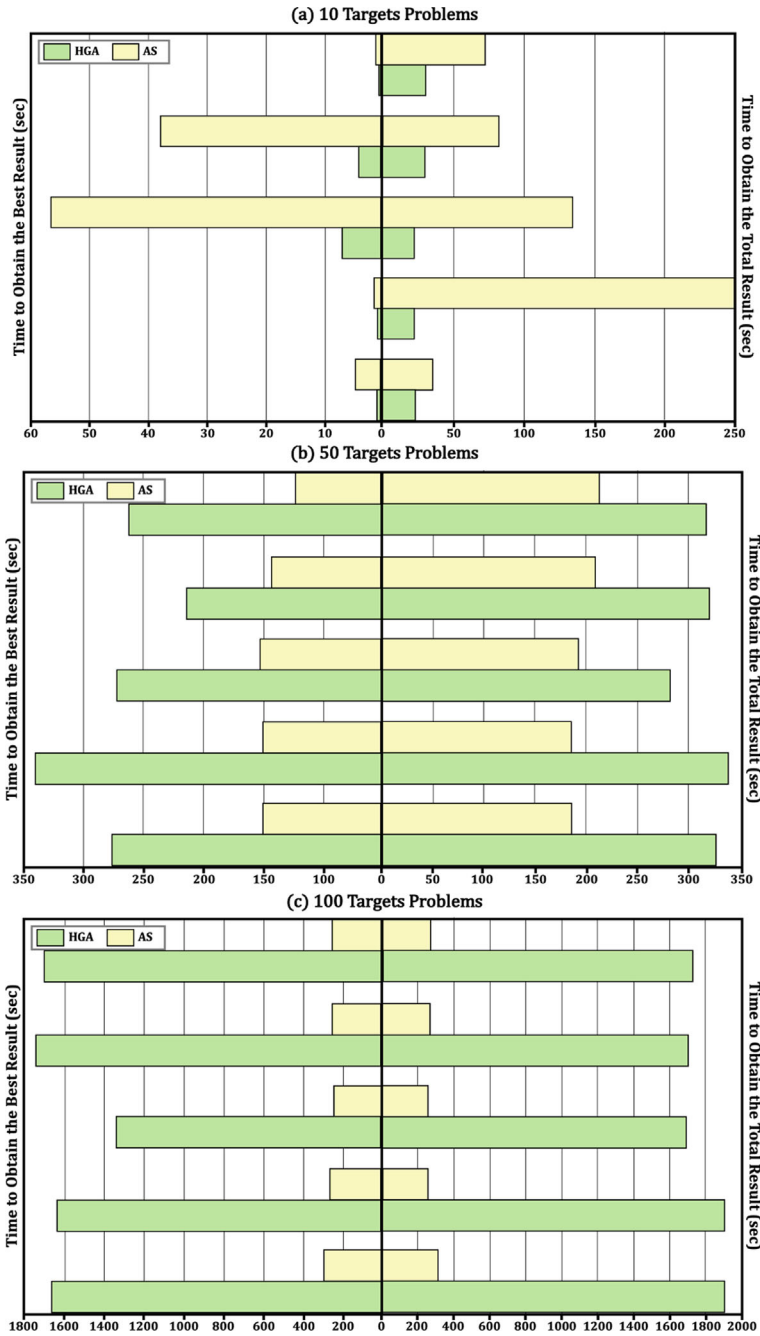


Fig. 8.10 Comparison of the times to obtain the best and total results of the hybrid genetic algorithm in annealing simulation

and unrealistic for the multi-UAV deployment version. Similarly, problems with 50 customers take considerably longer to solve compared to the solution times of the single UAV version. For this reason, the numerical analysis of VRP-mD is performed on 30-customer sample problems of a more appropriate size. For this purpose, 30 customer instances with x and y coordinates uniformly distributed between -16.0934 and $+16.0934$ were created. In these instances, the values of the assumed numerical parameters were used.

Figure 8.11 shows the distribution times and TSP solutions obtained by solving the example problems with 30 customers with 1, 2 and 3 UAVs. For each problem, the shared values are the average of the values obtained in 10 iterations of that problem. The objective column shows the deployment times (s), while the T_{best} and T_{total} columns indicate the time to obtain the best solution and the total solution time, respectively. The $Improvement_{TSP}$ column presents the percentage improvement in deployment times compared to the TSP solution in each scenario. The solution times increase rapidly in proportion to the number of UAVs. For example, in the case of 1 UAV, the time to reach the solution is 1 min, while this time increases to 1 h for 2 UAVs and 2.5 h for 3 UAVs.

The rate of increase in the improvement percentages obtained according to the TSP solution decreases with the number of UAVs. This result can be clearly seen in Table 8.1. In this context, it is seen that even the most inefficient solution can achieve 10% improvement, while 35% improvement can be achieved with increasing number of UAVs for a similar problem. Especially with the addition of the 3rd UAV to the distribution network, very small improvement rates are observed in problems 1 and 2 compared to the scenario with 2 UAVs. This situation is similarly observed in the utilization rates of UAVs.

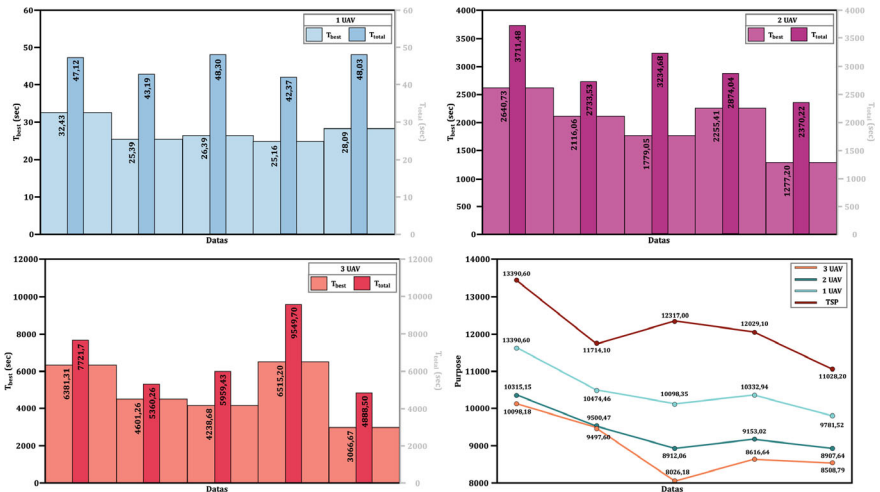


Fig. 8.11 Heuristic solutions with 1, 2 and 3 UAVs

Table 8.1 Percent improvement of the heuristic solution compared to the TSP solution

Datas	1 UAV (%)	2 UAV (%)	3 UAV (%)
I	13.47	22.97	24.59
II	10.58	18.90	19.26
III	18.01	27.64	34.84
IV	14.10	23.91	28.39
V	11.30	19.23	22.85

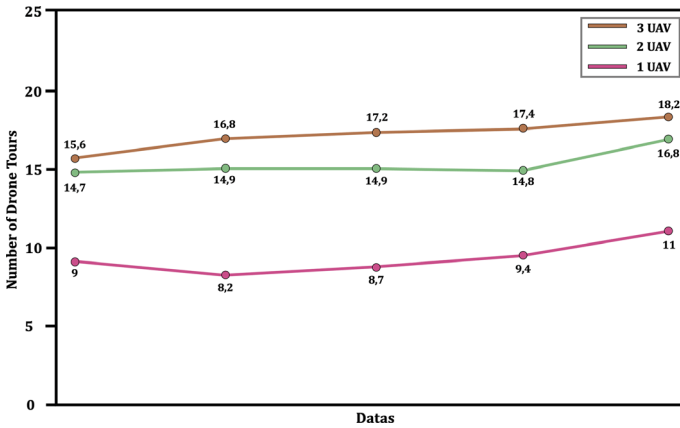


Fig. 8.12 Average utilization rates of UAVs

Figure 8.12 presents the UAV utilization rates obtained with different numbers of UAVs in the 30-client problem. According to Fig. 8.12, the UAV utilization rate decreased in the case of 3 UAVs compared to the case of 2 UAVs. This decrease in the UAV utilization rate is thought to be due to the increase in the number of UAVs and the increase in the waiting time of the ground vehicle for the UAVs.

The results in Fig. 8.12 show that the increase in the number of UAVs reduces the effective utilization of UAVs. This finding raises the question whether the increase in the number of UAVs can be compensated by an increase in battery life. For this purpose, the VRP-UAV problem with 30 customers was re-solved under the assumption that the battery life is twice as long (40 min). Figure 8.13 shows the average values of the deployment times obtained with different combinations of UAV numbers and battery life. According to the findings, only one of the 5 different problem instances, the combination of 1 UAV and 40 min battery life, provides similar deployment times to the combination of 2 UAVs and 20 min battery life.

Table 8.2 shows the average of the distribution times obtained at the end of the objective function performed in 10 iterations at 20 battery lifetimes for 30 delivery problems using 1, 2 and 3 UAVs.

This result shows that in addition to the number of UAVs, battery life and hence battery technology also plays an important role in the simultaneous deployment

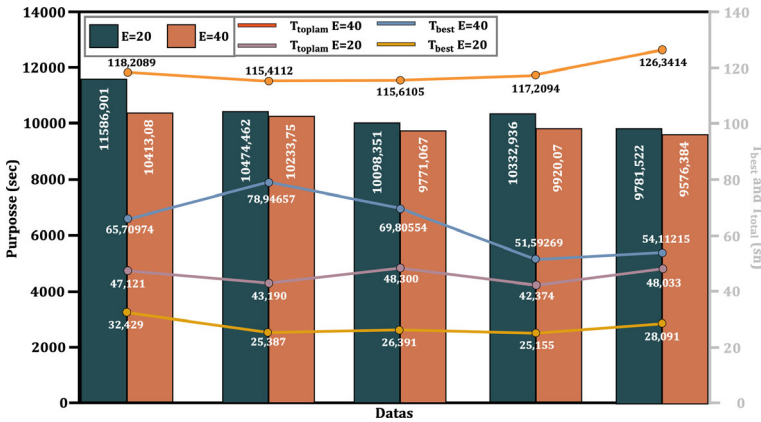


Fig. 8.13 Deployment times (s), various combinations of UAV numbers and battery life

Table 8.2 For 30 deployment problems, the deployment time for different UAVs with the same battery life

Datas	1 UAV		2 UAV		3 UAV	
	Time (h)	Recovery (%)	Time (h)	Recovery (%)	Time (h)	Recovery (%)
I	3.21	10.90	2.86	11.78	2.80	12.77
II	2.90	9.31	2.63	11.49	2.62	9.65
III	2.80	11.78	2.47	8.85	2.22	20.71
IV	2.87	11.49	2.54	8.85	2.39	16.72
V	2.71	8.85	2.47	8.85	2.36	12.91

problem of ground vehicles and UAVs. For a more in-depth analysis, there is clearly a need for studies that also consider costs.

8.5 Conclusion

This paper aims to improve the methods in the existing literature and fill the knowledge gaps in this area by addressing the simultaneous ground vehicle and drone delivery problem. The study aims to develop an exact solution to the VRP-D problem, which considers the scenario where ground vehicles and a single drone are used. This is achieved with a genetic algorithm based on function estimation with machine learning that maintains a two-stage structure. The numerical study was compared with the annealing simulation method using the same assumptions and showed that the genetic algorithm based on machine learning gives better results in solving small and medium-sized problems. However, for large-scale problems, the genetic algorithm based on machine learning was slower than the other method. Although VPR-D is still

considered as a new problem, it is attracting attention from the academic and business world and the number of studies in this field is increasing rapidly. Based on the results of this study, several important roadmaps for future research can be suggested. First, the focus should be on solving large-scale VDP-D problems more effectively, as such problems play an important role in real-world applications. Furthermore, studying various deployment scenarios involving different combinations of UAVs and ground vehicles can help to understand the potential of this technology from a broader perspective. Furthermore, studies that more closely examine how VRP-D can be adapted to different application domains and its environmental impacts can help to better understand the future use of this technology. These recommendations can help determine the direction of future studies in the field of VRP-D. These results were obtained during the parameter selection during the iterations when solving problems with 30 customers; population size (N_p) 50, number of individuals in the population generated by nearest neighbor (N_{nhn}) 10, initial training data size (N_t) 50, number of crossovers to be applied in each iteration (N_c) 25, number of iterations (N_{gn}) 500, number of best children to be included in the population at the end of each iteration (N_b) 3, number of randomly generated individuals to be included in the population at each iteration (N_{rhn}) 1, number of nearest neighbor generated individuals to be included in the population at each iteration (N_{nhnb}) 2, mutation probability (M_p) 0.5 and the number of nodes in the hidden layer is set to 100.

Appendix 1: Nomenclature

i, j	Node Index
k	Position Index
p	UAV Route Index
v	Vehicle Index
s	Iteration Index
C	Customer Cluster
V	Vehicle Cluster (Ground Vehicle and UAVs, $M + 1$)
N	Number of Customers in the Distribution Network
M	Number of Uavs in the Distribution Network
E	Duration Of Time The UAV Can Stay In The Air Without Running Out of Battery (Battery Life)
s	Iteration Index
D_s	Cluster of Customers Assigned to the UAV in Iteration “s”
P_s	The Cluster of Tours that the UAV can Perform in Iteration “s”
d_p	“P” Round Completion Time
f_{ip}	“P” Round is Given 1 If It Starts With Node “i”, 0 Otherwise
a_{ip}	In round”p”, The UAV is Given 1 If It Delivers to Customer “i” and 0 Otherwise
l_{ip}	“P” Round is Given 1 If It Ends With Node “i”, 0 Otherwise
D_{max}	The Largest Number of Clients that can be Assigned to UAV

N	Number of Customers in the Distribution Network
t_i	Time of Arrival of the Ground Vehicle at the “i” Node
m_k	Customer Assigned to Position “k” on the Ground Vehicle Route
n_s	Number of Customers on the Ground Vehicle Route Selected in Iteration “s”
\bar{r}_s	Ground Vehicle Route Determined in the First Stage of Iteration “s”
$\{0, N + 1\}$	Start and End Point of the Station
x_p	1 If Tour “p” is Assigned to the UAV, 0 Otherwise
w_i	Waiting Time of the Ground Vehicle for the UAV at Node “i”
t_i	Arrival Time of the Vehicle at Node ‘i’ on the Truck Route
H	Hidden Layer Output Matrix
h	Number of nodes in the hidden layer
$dist_{ij}$	Travel Time of the Truck Between Nodes “i” and “j”
t_{max}	Earliest Departure Time of the Vehicle from Node “I” on the Ground Vehicle Route
S_L	Service Time
S_R	Service Waiting
N_p	Population Size
N_{nhn}	Nearest Neighbor in the Population
N_t	Initial Training Data
N_c	Number of Crossovers to be Applied in each Iteration
N_{gn}	Number of Iterations
N_b	Number of Best Children
N_{nhn}	Number of Randomly Generated Individuals to be Included in the Population at Each Iteration
N_{nhnb}	Nearest Neighbor to be Included in the Population at Each Iteration
M_p	Mutation Probability
N_p	Population Size
N_{rhn}	Number of Individuals Formed According to the Nearest Neighbor in the Population
N_t	Initial Training Data

Appendix 2: List of Acronyms

UAV	Unmanned Aerial Vehicle
TSP	Travelling Salesmen Person
VRP	Vehicle Routing Problem
VRP-D	Vehicle Routing Problem with Drone
MNLP	Mixed-Integer Nonlinear Programming
GCS	Ground Control Station

References

1. Mohsan, S.A.H., Othman, N.Q.H., Li, Y., Alsharif, M.H., Khan, M.A.: Unmanned aerial vehicles (UAVs): practical aspects, applications, open challenges, security issues, and future trends. *Intel. Serv. Robot.* **16**(1), 109–137 (2023). <https://doi.org/10.1007/s11370-022-00452-4>
2. Fennelly, L.J., and Perry, M.A.: Unmanned aerial vehicle (drone) usage in the 21st century. In: *The Professional Protection Officer*, 2nd edn., pp. 183–189. Butterworth-Heinemann (2020). <https://doi.org/10.1016/B978-0-12-817748-8.00050-X>
3. Wang, H., Cheng, H., Hao, H.: The use of unmanned aerial vehicle in military operations. In: Long, S., Dhillon, B.S. (eds.) *Man-Machine-Environment System Engineering (MMESE)*, Lecture Notes in Electrical Engineering, vol. 645, pp. 939–945. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-6978-4_108
4. Norasma, C.Y.N., Fadzilah, M.A., Roslin, N.A., Zanariah, Z.W.N., Tarmidi, Z., Candra, F.S.: Unmanned aerial vehicle applications in agriculture. *IOP Conf. Ser.: Mater. Sci. Eng.* **506**(1), 1–10 (2019). <https://doi.org/10.1088/1757-899X/506/1/012063>
5. Mohamed, N., Al-Jaroodi, J., Jawhar, I., Idries, A., Mohammed, F.: Unmanned aerial vehicles applications in future smart cities. *Technol. Forecast. Soc. Chang.* **153**, 1–15 (2020). <https://doi.org/10.1016/j.techfore.2018.05.004>
6. Sanjana, P., and Prathilothamai, M.: Drone design for first aid kit delivery in emergency situation. In: *6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India, pp. 215–220 (2020). <https://doi.org/10.1109/ICACCS48705.2020.9074487>
7. Tao, L., Haitao, H.: Development of the use of unmanned aerial vehicles (UAVs) in emergency rescue in china. *Risk Manage. Healthcare Policy* **14**, 4293–4299 (2022). <https://doi.org/10.2147/RMHP.S323727>
8. Bakirci, M., Ozer, M.M.: Enhancing ground vehicle route planning with multi-drone integration. In: Seyman, M.N. (eds) *2nd International Congress of Electrical and Computer Engineering. ICECENG 2023. EAI/Springer Innovations in Communication and Computing*, pp. 103–117, Springer, Cham (2024). https://doi.org/10.1007/978-3-031-52760-9_8
9. Idrissi, M., Salami, M., Annaz, F.: A review of quadrotor unmanned aerial vehicles: applications, architectural design and control algorithms. *J. Intell. Rob. Syst.* **104**(2), 1–33 (2022). <https://doi.org/10.1007/s10846-021-01527-7>
10. Jha, S.K., Prakash, S., Rathore, R.S., Mahmud, M., Kaiwartya, O., Lloret, J.: Quality-of-service-centric design and analysis of unmanned aerial vehicles. *Sensors.* **22**(15), 1–18 (2022). <https://doi.org/10.3390/s22155477>
11. Darvishpoor, S., Roshanian, J., Raissi, A., Hassanalian, M.: Configurations, flight mechanisms, and applications of unmanned aerial systems: a review. *Prog. Aerosp. Sci.* **121**, 1–59 (2020). <https://doi.org/10.1016/j.paerosci.2020.100694>
12. Tan, M., Tang, A., Ding, D., Xie, L., Huang, C.: Autonomous air combat maneuvering decision method ofUCAV based on LSHADE-TSO-MPC under enemy trajectory prediction. *Electronics* **11**(20), 1–25 (2022). <https://doi.org/10.3390/electronics11203383>
13. Ruan, W., Duan, H., Deng, Y.: Autonomous maneuver decisions transfer learning pi-geon-inspired optimization forUCAVs in dogfight engagements. *IEEE/CAA Journal of Automatica Sinica* **9**(9), 1639–1657 (2022). <https://doi.org/10.1109/JAS.2022.105803>
14. Yue, L., Xiaohui, Q., Xiaodong, L., Qunli, X.: Deep reinforcement learning and its application in autonomous fitting optimization for attach areas ofUCAVs. *J. Syst. Eng. Electron.* **31**(4), 734–742 (2020). <https://doi.org/10.23919/JSEE.2020.000048>
15. Namian, M., Khalid, M., Wang, G., Turkan, Y.: Revealing safety risks of unmanned aerial vehicles in construction. *Transp. Res. Rec.* **2675**(11), 334–347 (2021). <https://doi.org/10.1177/03611981211017134>
16. Potter, B., Valentino, G., Yates, L., Benzing, T., Salman, A.: Environmental monitoring using a drone-enabled wireless sensor network. In: *Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA, pp. 1–6 (2019). <https://doi.org/10.1109/SIEDS.2019.8735615>

17. Yakushiji, K., Fujita, H., Murata, M., Hiroi, N., Hamabe, Y., Yakushiji, F.: Short-range transportation using unmanned aerial vehicles (UAVs) during disasters in japan. *Drones* **4**(4), 1–8 (2020). <https://doi.org/10.3390/drones4040068>
18. Salmoral, G., Rivas Casado, M., Muthusamy, M., Butler, D., Menon, P.P., Leinster, P.: Guidelines for the use of unmanned aerial systems in flood emergency response. *Water* **12**(2), 1–22 (2020). <https://doi.org/10.3390/w12020521>
19. Nikhil, N., Shreyas, S.M., Vyshnavi G., Yadav, S.: Unmanned aerial vehicles (UAV) in disaster management applications. In: 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 140–148 (2020). <https://doi.org/10.1109/ICSSIT48917.2020.9214241>
20. Al-Kaff, A., Madridano, Á., Campos, S., García, F., Martín, D., de la Escalera, A.: Emergency support unmanned aerial vehicle for forest fire surveillance. *Electronics* **9**(2), 1–14 (2020). <https://doi.org/10.3390/electronics9020260>
21. Hammad, A.W.A., da Costa, B.B.F., Soares, C.A.P., Haddad, A.N.: The use of unmanned aerial vehicles for dynamic site layout planning in large-scale construction projects. *Buildings* **11**(12), 1–17 (2021). <https://doi.org/10.3390/buildings11120602>
22. Olson, D., Anderson, J.: Review on unmanned aerial vehicles, remote sensors, imagery processing, and their applications in agriculture. *Agron. J.* **113**(2), 971–992 (2021). <https://doi.org/10.1002/agj2.20595>
23. Syed, F., Gupta, S.K., Hamood Alsamhi, S., Rashid, M., Liu, X.: A survey on recent optimal techniques for securing unmanned aerial vehicles applications. *Trans. Emerg. Tel. Tech.* **32**(7), 1–34 (2021). <https://doi.org/10.1002/ett.4133>
24. Gordan, M., Ismail, Z., Ghaedi, K., Ibrahim, Z., Hashim, H., Ghayeb, H.H., and Talebkah, M.: A brief overview and future perspective of unmanned aerial systems for in-service structural health monitoring. *Eng. Adv.* **1**(1), 9–15 (2021). <https://doi.org/10.26855/ea.2021.06.002>
25. Huttunen, M.: Civil unmanned aircraft systems and security: the European approach. *J. Transp. Secur.* **12**, 83–101 (2019). <https://doi.org/10.1007/s12198-019-00203-0>
26. Ewertowski, M.W., Tomczyk, A.M., Evans, D.J.A., Roberts, D.H., Ewertowski, W.: Operational framework for rapid, very-high resolution mapping of glacial geomorphology using low-cost unmanned aerial vehicles and structure-from-motion approach. *Remote Sens.* **11**(1), 1–19 (2019). <https://doi.org/10.3390/rs11010065>
27. Liu, X., Yang, Y., Ma, C., Li, J., Zhang, S.: Real-time visual tracking of moving targets using a low-cost unmanned aerial vehicle with a 3-axis stabilized gimbal system. *Appl. Sci.* **10**(15), 1–27 (2020). <https://doi.org/10.3390/app10155064>
28. Kawamura, K., Asai, H., Yasuda, T., Khanthavong, P., Soisouvanh, P., Phongchanmixay, S.: Field phenotyping of plant height in an upland rice field in Laos using low-cost small unmanned aerial vehicles (UAVs). *Plant Prod. Sci.* **23**(4), 452–465 (2020). <https://doi.org/10.1080/1343943X.2020.1766362>
29. Morita, S., Konert, A., Smereka, J., Szarpak, L.: The use of drones in emergency medicine: practical and legal aspects. *Emerg. Med. Int.* **2019**, 1–5 (2019). <https://doi.org/10.1155/2019/3589792>
30. Valsan, A., Parvathy, B., GH, V.D., Unnikrishnan, R. S., Reddy, P. K., Vivek, A.: Unmanned aerial vehicle for search and rescue mission. In: 4th International Conference on Trends in Electronics and Informatics (ICOEI) (48184), IEEE, Tirunelveli, India, pp. 684–687 (2020). <https://doi.org/10.1109/ICOEI48184.2020.9143062>
31. Bakirci, M., Ozer, M.M. (2023). Drone-assisted path planning optimization for mobile robots in dynamic scenarios. In: 2023 IEEE 7th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC). pp. 106–111. Kyiv, Ukraine. <https://doi.org/10.1109/MSNMC61017.2023.10329084>
32. Wang, D.B., Israr, A., Abro, G.E.M., Sadiq Ali Khan, M., Farhan, M., Bin Mohd Zulkifli, Su.A.: Internet of things (IoT)-enabled unmanned aerial vehicles for the inspection of construction sites: aA vision and future directions. *Math. Prob. Eng.* 1–15 (2021). <https://doi.org/10.1155/2021/9931112>

33. Tovar-Sánchez, A., Román, A., Roque-Atienza, D., Navarro, G.: Applications of unmanned aerial vehicles in Antarctic environmental research. *Sci. Rep.* **11**(1), 1–8 (2021). <https://doi.org/10.1038/s41598-021-01228-z>
34. Kyrkou, C., Theocharides, T.: EmergencyNet: efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **13**, 1687–1699 (2020). <https://doi.org/10.1109/JSTARS.2020.2969809>
35. Ahmed, F., Mohanta, J.C., Keshari, A., Yadav, P.S.: Recent advances in unmanned aerial vehicles: a review. *Arab. J. Sci. Eng.* **47**(7), 7963–7984 (2022). <https://doi.org/10.1007/s13369-022-06738-0>
36. Saeed, F., Mehmood, A., Majeed, M.F., Maple, C., Saeed, K., et al.: Smart delivery and retrieval of swab collection kit for COVID-19 test using autonomous Unmanned Aerial Vehicles. *Phys. Commun.* **48**, 1–16 (2021). <https://doi.org/10.1016/j.phycom.2021.101373>
37. Monica Dev, M., Hema, R.: A safe road to health: Medical services using unmanned aerial vehicle. In: Gupta, D., Khanna, A., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds.) *International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing*, vol. 1165, pp. 367–375. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-5113-0_27
38. Ranjan, A., Sahu, H.B., Misra, P., Panigrahi, B.: Leveraging unmanned aerial vehicles in mining industry: research opportunities and challenges. In: Al-Turjman, F. (eds.) *Unmanned Aerial Vehicles in Smart Cities, Unmanned System Technologies*, pp. 107–132. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-38712-9_7
39. Anand, R., Muneshwara, M.S., Shivakumara, T., Swetha, M.S., Anil, G.N.: Emergency medical services using drone operations in natural disaster and pandemics. In: Ranganathan, G., Fernando, X., Shi, F. (eds.) *Inventive Communication and Computational Technologies. Lecture Notes in Networks and Systems*, vol. 311, pp. 227–239. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-5529-6_19
40. Tkáč, M., Mésároš, P.: Utilizing drone technology in the civil engineering. *Sel. Sci. Pap. J. Civil Eng.* **14**(1), 27–37 (2019). <https://doi.org/10.1515/sspjce-2019-0003>
41. Muneem, I.A., Fahim, S.M., Khan, F.R., Emon, T.A., Islam M.S., Khan, M.M.: Research and development of multipurpose unmanned aerial vehicle (flying drone). In: 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, pp. 402–406 (2021). <https://doi.org/10.1109/UEMCON53757.2021.9666736>
42. Eichleay, M., Evens, E., Stankevitz, K., Parker, C.: Using the unmanned aerial vehicle delivery decision tool to consider transporting medical supplies via drone. *Glob. Health: Sci. Practice* **7**(4), 500–506 (2019). <https://doi.org/10.9745/GHSP-D-19-00119>
43. Otero Arenzana, A., Escribano Macias, J.J., Angeloudis, P.: Design of hospital delivery networks using unmanned aerial vehicles. *Escribano Macias Jose Javier* **2674**(5), 405–418 (2020). <https://doi.org/10.1177/0361198120915891>
44. Škrinjar, J.P., Škorput, P., Furdić, M.: Application of unmanned aerial vehicles in logistic processes. In: Karabegović, I. (eds.) *New Technologies, Development and Application (NT 2018), Lecture Notes in Networks and Systems*, vol. 42, pp. 359–366. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-90893-9_43
45. Engesser, V., Rombaut, E., Vanhaverbeke, L., Lebeau, P.: Autonomous delivery solutions for last-mile logistics operations: a literature review and research agenda. *Sustainability* **15**(3), 1–17 (2023). <https://doi.org/10.3390/su15032774>
46. Raivi, A.M., Huda, S.M.A., Alam, M.M., Moh, S.: Drone routing for drone-based delivery systems: a review of trajectory planning, charging, and security. *Sensors* **23**(1463), 1–26 (2023). <https://doi.org/10.3390/s23031463>
47. Khosravi, M., Enayati, S., Saeedi, H., Pishro-Nik, H.: Multi-purpose drones for coverage and transport applications. *IEEE Trans. Wirel. Commun.* **20**(6), 3974–3987 (2021). <https://doi.org/10.1109/TWC.2021.3054748>

48. Das, D.N., Sewani, R., Wang, J., Tiwari, M.K.: Synchronized truck and drone routing in package delivery logistics. In *IEEE Trans. Intell. Transp. Syst.* **22**(9), 5772–5782 (2021). <https://doi.org/10.1109/TITS.2020.2992549>
49. Najy, W., Archetti, C.: Collaborative truck-and-drone delivery for inventory-routing problems. *Trans. Res. Part C: Emerg. Technol.* **146**, 1–20 (2023). <https://doi.org/10.1016/j.trc.2022.103791>
50. Rinaldi, M., Primatesta, S., Bugaj, M., Rostáš, J., Guglieri, G.: Development of heuristic approaches for last-mile delivery tsp with a truck and multiple drones. *Drones* **7**(7), 1–32 (2023). <https://doi.org/10.3390/drones7070407>
51. Montaña, L.C., Malagon-Alvarado, L., Miranda, P.A., Arboleda, M.M., Solano-Charris, E.L., et al.: A novel mathematical approach for the truck-and-drone location-routing problem. *Procedia Comput. Sci.* **200**, 1378–1391 (2022). <https://doi.org/10.1016/j.procs.2022.01.339>
52. Santillán, C.G., Reyes, L.C., Rodríguez, M.L.M., Barbosa, J.J.G., López, O.C., Zarate, G.R., Hernández, P.: Variants of VRP to optimize logistics management problems. In: *Logistics Management and Optimization through Hybrid Artificial Intelligence Systems*, pp. 207–237. IGI Global (2012). <https://doi.org/10.4018/978-1-4666-0297-7.ch008>
53. Imran, N.M., Mishra, S., Won, M.: A-VRPD: Automating drone-based last-mile delivery using self-driving cars. *IEEE Trans. Intell. Transp. Syst.* **24**(9), 9599–9612 (2023). <https://doi.org/10.1109/TITS.2023.3266460>
54. Prawira, H.A., Santosa, B.: Development of particle swarm optimization and simulated annealing algorithms to solve vehicle routing problems with drones. *PROZIMA (Prod. Optim. Manuf. Syst. Eng.)* **5**(1), 1–12 (2021). <https://doi.org/10.21070/prozima.v5i1.1398>
55. Yang, J., Yang, H., He, Z., Zhao Q., Shi, Y.: Solving vehicle routing problem with drones based on a bi-level heuristic approach. In: *International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2906–2911. IEEE, Prague, Czech Republic (2022). <https://doi.org/10.1109/SMC53654.2022.9945266>
56. Sitek, P., Wikarek, J., Jagodziński, M.: A proactive approach to extended vehicle routing problem with drones (EVRPD). *Appl. Sci.* **12**(16), 1–21 (2022). <https://doi.org/10.3390/app12168255>
57. Tamke, F., Buscher, U.: A branch-and-cut algorithm for the vehicle routing problem with drones. *Transp. Res. Part B: Methodol.* **144**, 174–203 (2021). <https://doi.org/10.1016/j.trb.2020.11.011>
58. Kitjacharoenchai, P., Min, B.-C., Lee, S.: Two echelon vehicle routing problem with drones in last mile delivery. *Int. J. Prod. Econ.* **225**, 1–14 (2020). <https://doi.org/10.1016/j.ijpe.2019.107598>
59. Han, Y., Li, J., Liu, Z., Liu, C., Tian, J.: Metaheuristic algorithm for solving the multi-objective vehicle routing problem with time window and drones. *Int. J. Adv. Rob. Syst.* **17**(2), 1–14 (2020). <https://doi.org/10.1177/1729881420920031>
60. Nguyen, M.A., Dang, G.T.H., Hà, M.H., Pham, M.T.: The min-cost parallel drone scheduling vehicle routing problem. *Eur. J. Oper. Res.* **299**(3), 910–930 (2022). <https://doi.org/10.1016/j.ejor.2021.07.008>
61. Ahn, N., Kim, S.: Optimal and heuristic algorithms for the multi-objective vehicle routing problem with drones for military surveillance operations. *J. Ind. Manage. Optim.* **18**(3), 1651–1663 (2022). <https://doi.org/10.3934/jimo.2021037>
62. Liu, Y.-Q., Han, J., Zhang, Y., Li, Y., Jiang, T.: Multivisit drone-vehicle routing problem with simultaneous pickup and delivery considering no-fly zones. *Discret. Dyn. Nat. Soc.* **2023**, 1–21 (2023). <https://doi.org/10.1155/2023/1183764>
63. Li, H., Chen, J., Wang, F., Zhao, Y.: Truck and drone routing problem with synchronization on arcs. *Nav. Res. Logist.* **69**(6), 884–901 (2022). <https://doi.org/10.1002/nav.22053>
64. Perwira Redi, A.A.N., Liperdá, R.I., Sopha, B.M., Sri Asih, A.M., Sekarintyas N.N., Astiana, H.B.: Relief mapping assessment using two-echelon vehicle routing problem with drone. In: *6th International Conference on Science and Technology (ICST)*, pp. 1–5. IEEE, Yogyakarta, Indonesia (2020). <https://doi.org/10.1109/ICST50505.2020.9732812>
65. Chen, C., Demir, E., Huang, Y.: An adaptive large neighborhood search heuristic for the vehicle routing problem with time windows and delivery robots. *Eur. J. Oper. Res.* **294**(3), 1164–1180 (2021). <https://doi.org/10.1016/j.ejor.2021.02.027>

66. Lu, Y., Yang, C., Yang, J.: A multi-objective humanitarian pickup and delivery vehicle routing problem with drones. *Ann. Oper. Res.* **319**(1), 291–353 (2022). <https://doi.org/10.1007/s10479-022-04816-y>
67. Moeini, M., Wendt, O., Schummer, M.: A bi-objective routing problem with trucks and drones: Minimizing mission time and energy consumption. In: Gervasi, O., et al. (eds.) *Computational Science and Its Applications—ICCSA 2023 Workshops*. ICCSA 2023. *Lecture Notes in Computer Science*, vol. 14106, pp. 291–308. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-37111-0_21
68. Ahmadi, E., Wicaksono, H., and Valilai, O.F.: Extending the last mile delivery routing problem for enhancing sustainability by drones using a sentiment analysis approach. *International Conference on Industrial Engineering and Engineering Management (IEEM)*. pp. 207–212. IEEE, Singapore, Singapore (2021). <https://doi.org/10.1109/IEEM50564.2021.9672856>
69. Konstantakopoulos, G.D., Gayialis, S.P., Kechagias, E.P.: Vehicle routing problem and related algorithms for logistics distribution: a literature review and classification. *Oper. Res. Int. Journal* **22**(3), 2033–2062 (2022). <https://doi.org/10.1007/s12351-020-00600-7>
70. Rojas Vilorio, D., Solano-Charris, E.L., Muñoz-Villamizar, A., Montoya-Torres, J.R.: Unmanned aerial vehicles/drones in vehicle routing problems: a literature review. *Int. Trans. Oper. Res.* **28**(4), 1626–1657 (2021). <https://doi.org/10.1111/itor.12783>
71. Mor, A., Speranza, M.G.: Vehicle routing problems over time: a survey. *Ann. Oper. Res.* **314**(1), 255–275 (2022). <https://doi.org/10.1007/s10479-021-04488-0>
72. Sakthivel, M., Kant Gupta, S., Karras, D.A., Khang, A., Kumar Dixit, C. Haralayya, B.: Solving vehicle routing problem for intelligent systems using delaunay triangulation. In: *International Conference on Knowledge Engineering and Communication Systems (ICKES)*, pp. 1–5. IEEE, Chickballapur, India (2022). <https://doi.org/10.1109/ICKES56523.2022.10060807>
73. Ramasamy S., Mondal, M.S., Reddinger, J.-P.F., Dotterweich, J.M., Humann, J.D.: Heterogeneous vehicle routing: comparing parameter tuning using genetic algorithm and bayesian optimization. In: *International Conference on Unmanned Aircraft Systems (ICUAS)*, Dubrovnik, Croatia, pp. 104–113 (2022). <https://doi.org/10.1109/ICUAS54217.2022.9836044>
74. Mirjalili, S., Song Dong, J., Sadiq, A.S., Faris, H.: Genetic algorithm: Theory, literature review, and application in image reconstruction. In: Mirjalili, S., Song Dong, J., Lewis, A. (eds) *Nature-Inspired Optimizers*. *Studies in Computational Intelligence*, Springer, Cham, 811, 69–85 (2020). https://doi.org/10.1007/978-3-030-12127-3_5
75. Gen, M., Lin, L.: Genetic algorithms and their applications. In: Pham, H. (eds.) *Springer Handbook of Engineering Statistics*. *Springer Handbooks*, pp. 635–674. Springer, London, (2023). https://doi.org/10.1007/978-1-4471-7503-2_33
76. Sohail, A.: Genetic algorithms in the fields of artificial intelligence and data sciences. *Ann. Data Sci.* **10**(4), 1007–1018 (2023). <https://doi.org/10.1007/s40745-021-00354-9>
77. Reddy, G.T., Reddy, M.P.K., Lakshmana, K., Rajput, D.S., Kaluri, R., et al.: Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evol. Intel.* **13**(2), 185–196 (2020). <https://doi.org/10.1007/s12065-019-00327-1>
78. Katoch, S., Chauhan, S.S., Kumar, V.: A review on genetic algorithm: past, present, and future. *Multimed. Tools Appl.* **80**(5), 8091–8126 (2021). <https://doi.org/10.1007/s11042-020-10139-6>
79. Thida San, K., Chang, Y.S.: Drone-based delivery: a concurrent heuristic approach using a genetic algorithm. *Aircraft Eng. Aerosp. Technol.* **94**(8), 1312–1326 (2022). <https://doi.org/10.1108/AEAT-07-2020-0138>
80. Ochelska-Mierzejewska, J., Ponszewska-Marańda, A., Marańda, W.: Selected genetic algorithms for vehicle routing problem solving. *Electronics* **10**(24), 1–34 (2021). <https://doi.org/10.3390/electronics10243147>
81. Bakirci, M., Cetin, M.: Improving position-time trajectory accuracy in vehicle stop-and-go scenarios by using a mobile robot as a testbed. *J. Control Eng. Appl. Inf.* **25**(3), 35–44 (2023). <https://doi.org/10.61416/ceai.v25i3.8365>
82. Ahmed, Z.H., Hameed, A.S., Mutar, M.L.: Hybrid genetic algorithms for the asymmetric distance-constrained vehicle routing problem. *Math. Probl. Eng.* **2022**, 1–20 (2022). <https://doi.org/10.1155/2022/2435002>

83. Bakirci, M.: Data-driven system identification of a modified differential drive mobile robot through on-plane motion tests. *Electrica* **23**(3), 619–633 (2023). <https://doi.org/10.5152/electrica.2023.22164>
84. Bakirci, M., Toptas, B.: Kinematics and autoregressive model analysis of a differential drive mobile robot. In: (IEEE) 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, pp. 1–6 (2022). <https://doi.org/10.1109/HORA55278.2022.9800071>
85. Benarbia, T., Kyamakya, K.: A literature review of drone-based package delivery logistics systems and their implementation feasibility. *Sustainability* **14**(1), 1–15 (2021). <https://doi.org/10.3390/su14010360>
86. Rave, A., Fontaine, P., Kuhn, H.: Drone location and vehicle fleet planning with trucks and aerial drones. *Eur. J. Oper. Res.* **308**(1), 113–130 (2023). <https://doi.org/10.1016/j.ejor.2022.10.015>
87. Khosravi, M., Pishro-Nik, H.: Unmanned aerial vehicles for package delivery and network coverage. In: 91st Vehicular Technology Conference (VTC2020-Spring), pp. 1–5. IEEE, Antwerp, Belgium (2020). <https://doi.org/10.1109/VTC2020-Spring48590.2020.9129495>
88. Dinelli, C., Racette, J., Escarcega, M., Lotero, S., Gordon, J., Montoya, J., Dunaway, C., Androulakis, V., Khaniani, H., Shao, S., Roghanchi, P., Hassanaliam, M.: Configurations and applications of multi-agent hybrid drone/unmanned ground vehicle for underground environments: a review. *Drones* **7**(136), 1–54 (2023). <https://doi.org/10.3390/drones7020136>
89. Ribeiro, R.G., Cota, L.P., Euzébio, T.A.M., Ramírez, J.A., Guimarães, F.G.: Unmanned-aerial-vehicle routing problem with mobile charging stations for assisting search and rescue missions in postdisaster scenarios. *IEEE Trans. Syst. Man Cybern. Syst.* **52**(11), 6682–6696 (2022). <https://doi.org/10.1109/TSMC.2021.3088776>
90. Jiang, J., Dai, Y., Yang, F., Ma, Z.: A multi-visit flexible-docking vehicle routing problem with drones for simultaneous pickup and delivery services. *Eur. J. Oper. Res.* **312**(1), 125–137 (2024). <https://doi.org/10.1016/j.ejor.2023.06.021>
91. Mara, S.T.W., Elsayed, S., Essam, D., Sarker, R.: Vehicle routing problem for an integrated electric vehicles and drones system. In: Martins, A.L., Ferreira, J.C., Kocian, A., Tokkozhina, U. (eds.) *Intelligent Transport Systems. INTSYS 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 486, pp. 197–214. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-30855-0_14

Chapter 9

Automation of Topic Generation in Government Information Requests in Mexico



Hermelando Cruz-Pérez , Alejandro Molina-Villegas ,
and Edwin Aldana-Bobadilla 

Abstract In Mexico, legislation guarantees public access to information, empowering citizens to request data from the government. This research delves into the National Transparency Platform's extensive archive, which includes over 2 million requests for information, with the goal of discerning the primary interests of citizens in government actions from 2003 to 2020. Through the analysis of 2,518,875 requests, Genetic Algorithms were employed to fine-tune three crucial hyperparameters of the Latent Dirichlet Allocation (LDA) model: alpha, beta, and the number of topics. This optimization aimed at enhancing the model's accuracy in topic identification, measured by the coherence metric of the topics identified. Additionally, Generative Pre-trained Transformer (GPT) technology facilitated the automatic generation of titles and descriptions for these topics. The investigation revealed 4131 topics of public interest throughout the Mexican Republic, with significant emphasis on environmental management, public policies, the response to the COVID-19 health crisis, labor issues, and education in 2020. These findings underscore the critical role of proactive transparency and the provision of open data in advancing the analysis of vast quantities of government data. This study paves the way for future data-driven decision-making and policy development research. It highlights the profound influence of sophisticated data analysis in promoting government transparency and stimulating citizen engagement. Using genetic algorithms to refine the LDA model and large language model technology for content generation, this study innovates analyzing public information requests, contributing significantly to improving governmental transparency.

H. Cruz-Pérez (✉) · A. Molina-Villegas
Conahcyt-Centro de Investigación en Ciencias de Información Geoespacial,
Mexico City 14240, Mexico
e-mail: hacruz@centrogeo.edu.mx

A. Molina-Villegas
e-mail: amolina@centrogeo.edu.mx

E. Aldana-Bobadilla
Conahcyt-Centro de Investigación y de Estudios Avanzados del Instituto Politécnico
Nacional-Unidad Tamaulipas, Ciudad Victoria, Tamaulipas 87130, Mexico
e-mail: edwyn.aldana@cinvestav.mx

Keywords Genetic algorithm · Latent Dirichlet Allocation · Governmental transparency · Open data · Data processing

9.1 Introduction

In an era where government data swells at an unprecedented rate, the skill to meticulously parse and derive insights from these extensive datasets is a cornerstone for advancing public understanding and underpinning data-driven policy-making. In the governmental sphere, this reality is especially notable, as public administrations generate immense amounts of data in their daily interactions with citizens. A clear example is the numerous requests directed to the Mexican government through the National Transparency Platform (NTP).

This study focuses on the application of advanced data processing techniques. Specifically, a genetic algorithm has been developed to automate the selection of key hyperparameters, including Alpha, Beta, and the number of topics, which are crucial for the efficient training of the LDA (Latent Dirichlet Allocation) model. This model is fundamental in the thematic analysis of large datasets, enabling the extraction of meaningful topics from unstructured text data. Further, we have applied optimization techniques to the vocabulary to enhance the quality of topic analysis. This includes Text Mining techniques to ascertain topic similarity across different geographic areas. Additionally, in line with the principles of Zipf's Law, our approach to vocabulary optimization has significantly improved the precision of our results, confirming the law's relevance in the context of natural language processing and topic modeling.

With this study, we aim to establish a solid framework for future research and practical applications in the processing and analysis of large data sets, thus marking a milestone in the management of public information and its efficient interpretation.

9.1.1 *Background of Access to Information*

The history of access to public information has its roots in Sweden, where, in 1766, the Freedom of the Press Act and the right to access public records were enacted. Anders Chydenius, a Swedish-Finnish priest who also played roles as a member of parliament and economist, drove this pioneering legislation. This event marked a significant milestone in the administrative centralization of Sweden, laying the groundwork for future developments in this field [27].

However, the modern concept of access to information underwent a fundamental change after the conclusion of World War II. In 1948, the Universal Declaration of Human Rights was proclaimed, which revisited the notion of freedom of expression. This declaration established that no citizen should be disturbed because of their opinions and fundamentally recognized the rights of citizens to seek, disseminate, and receive information. These historical events laid the foundations for the current con-

cept of access to information, which is intrinsically linked to the right of individuals to seek, disseminate, and receive information by any means available. Today, information is considered an indispensable resource for the exercise of other fundamental rights [14].

In Mexico, freedom of expression is a fundamental right recognized in the Constitution of 1917. Article 6 states that the expression of ideas is protected, except when it affects morality, the rights of others, incites crime, or disturbs public order. This balanced approach seeks to promote freedom of expression while protecting other important values [10].

However, it was in December 1977, during López Portillo's political reforms, that the phrase "The right to information is guaranteed by the State" was added to Article Six of the Constitution, marking a milestone in the recognition of this fundamental right in Mexico [14]. Despite this significant inclusion, there persisted a notable ambiguity concerning the practical implications of this enshrined right. Citizens could access governmental information, but the Supreme Court of Justice of the Nation considered that this reform did not fully guarantee access to information generated by the government. This process reflects the evolution of the right to information in Mexico and the ongoing need to protect it effectively.

The landscape of transparency in Mexico underwent a pivotal transformation on April 30, 2002, when Congress unanimously enacted the Federal Law of Transparency and Access to Governmental Information (LFTAIG). This landmark legislation introduced institutions and procedures enabling individuals to request federal governmental information from designated entities, marking a significant milestone in the country's commitment to open governance. Additionally, a one-month deadline was established for these entities to provide the requested information. The Federal Institute for Access to Information (IFAI) was designated as the entity responsible for ensuring the right to access information [8, 15].

The most significant national milestone was the approval in 2015 of the General Law of Transparency and Access to Public Information (LGTAIP). Its main objective is to promote transparency in institutions and strengthen participation and accountability [25]. Furthermore, this law changed the name of IFAI to INAI to enhance its role as a national-level guarantor institution.

9.1.2 Role of the National Institute for Transparency, Access to Information, and Personal Data Protection (INAI)

The National Institute of Transparency, Access to Information and Protection of Personal Data (INAI) in Mexico plays a key role as an independent body in promoting government transparency and safeguarding personal information. Its main function is to ensure public access to information and to oversee the proper handling of personal data, fostering accountability of authorities, and ensuring the privacy of citizens.

According to the General Law of Transparency and Access to Public Information [11], the (INAI) performs the following functions:

- Resolve citizen complaints who submit requests for review against obligated subjects to safeguard the right to access information.
- Resolve disputes from citizens who submit resources of non-compliance against decisions made by local guarantor organizations; when these contravene the principle of maximum publicity of information.
- Establish sanctions and take appropriate measures in cases where the rights of access to public information and the protection of personal data are violated.
- Ensure that personal data in the possession of government entities or companies are properly safeguarded and secure, and that they are not shared without the knowledge, explicit consent, and approval of their owners.

The purpose of the INAI in Mexico is to ensure that all Mexican citizens can exercise their right to access information while being protected from any undue interference by other governmental actors. The INAI conforms to policies proposed by international organizations that seek to promote democratic governance through inclusive political institutions [9].

The General Law of Transparency and Access to Governmental Public Information (LGTAIPG), published on May 5 strengthens the autonomy of the INAI by granting it can review resolutions of local guarantors bodies in cases of controversy. Additionally, it consolidates the position of the INAI as the highest authority on the matter and prevents authorities from challenging its decisions before administrative or judicial bodies. The law also expands the powers of the guarantor bodies by allowing the INAI to resolve review requests filed by individuals against state guarantor bodies, as well as to take coercive measures and sanctions when necessary [4, 17].

The INAI is committed to promoting and contributing to the creation of a state in Mexico that is more transparent and accessible. The aim is to restore the public's trust in the authorities and to promote active participation of society in all processes related to the formulation of public policies. By working together to address public challenges, the likelihood of democracy being effective and generating positive results is increased [26].

9.1.3 The Importance of the National Transparency Platform (PNT)

The National Transparency Platform (PNT) is the digital strategy established by the Mexican Government to make transparent the operations of more than 8000 Obligated Subjects (SO) across the country. The PNT is a fundamental part of the information dissemination work coordinated by the National Institute for Transparency, Access to Information, and Protection of Personal Data (INAI), which is responsible for liaising with and linking the 32 institutes and/or commissions that, in turn, work in coordination with the SOs of each state.

The PNT plays an essential role in the Mexican Government's efforts to promote transparency in the management of more than 8000 Obligated Subjects (SO) in all states of the Mexican Republic. This digital platform has been designed to ensure that relevant information is available to the public, promoting greater accountability at all levels of government.

The INAI plays a central role in the implementation and coordination of the PNT. This body is responsible for establishing links with the 32 state institutes and/or commissions, who, in turn, collaborate closely with the SOs in each state. Together, they work synchronously to ensure that public information is accessible, verifiable, and understandable for all citizens. The PNT not only seeks to ensure transparency in the management of public resources but also promotes access to relevant information in areas such as government procurement, public finances, government policies, and other topics of public interest. This allows citizens to be better informed and actively participate in decision-making, thus strengthening democracy and citizen control over government actions.

It is crucial to highlight that the National Transparency Platform (PNT) consists of four key systems, each playing specific roles in establishing a comprehensive and highly interoperable digital space. Firstly, the Information Access Request System (SISAI) connects users with over eight thousand public institutions, allowing them to request public and personal information. In addition to facilitating this interaction, SISAI also manages the registration of requests and ensures compliance with deadlines established by law. In cases of potential delays or omissions, this system presents complaint resources to expedite the process [1].

Complementing the Information Access Request System, the PNT includes three other essential systems: the Management System of Means of Contestation, the System of Transparency Obligations Portals, and the Communication System between Guarantor Bodies and Obligated Subjects (SO). These systems collaborate to establish an efficient and highly integrated cyberspace that simplifies the participation of all actors involved in access to information and governmental transparency.

9.2 State of the Art

Some significant works stand out in information request analysis using advanced techniques. [2] used supervised Latent Dirichlet Allocation (LDA) techniques to predict the Mexican government's response capability to federal information requests between 2003 and 2015. Similarly, [3] applies sLDA to investigate the Mexican government's response (or lack thereof) to public information requests during the same period, identifying the topics most linked to governmental responses.

Another author [6] analyzes a large dataset consisting of hundreds of thousands of information requests, comparing similar cases in terms of dependencies, themes, and timing and considering the complexity and sensitivity of the requests. This study reveals a higher success rate for requests from regions with strong support for ruling parties, especially on issues of public relevance. This suggests an attempt to mitigate

political risks rather than favor supporters. Similarly, [5] examines over a million information requests, using unsupervised methods to categorize them based on their thematic diversity. This analysis demonstrates the variability of requests, from public transparency to more private and micropolitical interests, and highlights public concern for environmental impact and violence.

Our methodology is distinguished by proposing a new automated approach to identify topics with the optimal values of alpha and beta parameters based on their coherence. In addition, we conduct differentiated analyses by state and year.

Regarding the automation of hyperparameters using another type of information, we can mention [20], which focuses on optimizing LDA configurations to surpass the results of the conventional LDA model. The work introduces and applies the SA-LDA algorithm, which leverages Simulated Annealing (SA) to determine the optimal values of LDA parameters. Analogously, [29] explores the improvement of LDA model parameters through the development of a parallel differential evolution algorithm that incorporates two cost functions, LDADE and Word2Vec, emphasizing the ongoing commitment to refining techniques for thematic modeling.

9.3 Theoretical Framework: Automation of Topic Generation

The Automation of Topic Generation using Genetic Algorithms for Hyperparameters in Latent Dirichlet Allocation (LDA) Modeling represents a fascinating and advanced approach in the field of natural language processing and text mining. This methodology combines the robustness of LDA Topic Analysis with the efficiency of genetic algorithms, offering a synergistic approach that enhances the capability to uncover latent topics within large text corpora. LDA is a widely used topic modeling technique for uncovering hidden structures or topics within large text data sets. However, the performance of this model depends greatly on the optimization of its hyperparameters, mainly alpha, beta, and the number of topics (topic).

In this context, genetic algorithms, which are search and optimization methods inspired by natural selection and genetics present themselves as a powerful solution. These algorithms iteratively adjust and fine-tune the values of the LDA hyperparameters, seeking the optimal combination that maximizes the coherence and relevance of the generated topics. The process involves generating a population of LDA models with different hyperparameters configurations, evaluating their performance, and applying genetic operations such as selection, crossover, and mutation to evolve the hyperparameter configurations over several generations.

This automatic approach not only improves the quality of the generated topics but also frees users from the tedious and often subjective task of manually adjusting hyperparameters. In this section, in addition to presenting a solid theoretical framework, we will delve into the key concepts that support this approach, providing a more complete understanding of its functioning and applicability.

9.3.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is an unsupervised topic analysis algorithm used to identify underlying themes in a set of documents. The algorithm functions by assuming that each document is composed of a mixture of latent topics, and that each topic is comprised of a set of words [7]. The idea behind LDA is to find a representation of documents in terms of a probability distribution over a finite set of latent topics, where each topic is represented by a probability distribution over the vocabulary. To generate a new document, LDA first selects a mix of topics according to the topic distribution of the document. Then, for each word in the document, a topic is selected according to the chosen mix, and a word from that topic is chosen according to its probability distribution in the vocabulary.

In [7], Blei et al. introduced the concept of Latent Dirichlet Allocation by employing the Dirichlet distribution, as delineated in Eq. (9.1), as the prior distribution of the two generative models: topics in documents and words in topics. In Eq. (9.1), the Dirichlet distribution's parameters are leveraged to control the mixture of topics within documents. Here, Γ denotes the generalization of the factorial function to the real number domain, \mathbb{R}^n , effectively extending $n!$ to non-integer dimensions. The variable T represents the total number of topics within the document corpus, while α serves as the hyperparameter influencing the distribution's shape, effectively dictating the topic mixture's concentration and diversity.

The Dirichlet distribution, defined by Eq. (9.1) below, plays a pivotal role in the generative process of LDA, specifying the prior probabilities of topic distributions within documents. The term p_j , within the product, signifies the probability assigned to each topic j , with the exponent $\alpha - 1$ modulating the influence of each topic based on the hyperparameter α . The objective is to find the probability values of the distribution so that the original documents can be generated.

$$Dir(\alpha) = \frac{\Gamma(\alpha T)}{\Gamma(\alpha)^T} \prod_{j=1}^T p_j^{\alpha-1} \quad (9.1)$$

In equation for the Dirichlet distribution within LDA:

- $Dir(\alpha)$: Dirichlet distribution parameterized by α .
- α : Controls uniformity across topics; higher values mean more evenly distributed topics.
- T : Number of topics.
- Γ : Gamma function, for normalization.
- p_j : Probability of topic j .
- \prod : Product of topic probabilities, influenced by α .

9.3.2 Topic Evaluation

In topic modeling, coherence assessment initiates with identifying “pivotal words” for a fixed corpus. The “pivotal words” are selected based on frequency and thematic pertinence. The interrelations among these words are scrutinized using metrics like semantic similarity, thereby establishing a foundation for thematic cohesion. Subsequently, coherence is quantified by evaluating aspects such as the semantic proximity of words and their distribution across the corpus. This analytical procedure was integrated using statistical methodologies, including the arithmetic mean or the median, to formulate a comprehensive coherence score for the theme [24].

As delineated in [18], the role of coherence in topic modeling is instrumental for ascertaining the interpretability of themes by human analysts. By characterizing themes through the most likely words, the coherence score gauges the degree of similarity among these words, serving as a crucial indicator of thematic coherence. One of the most commonly used metrics is the UMass coherence score Eq. 9.2. This metric calculates the frequency with which two words appear together in the corpus.

$$C_{UMass} = \sum_{i=2}^N \sum_{j=1}^{i-1} \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)} \quad (9.2)$$

The components of the UMass coherence equation are delineated below, elucidating their roles in assessing topic coherence within the realm of topic modeling:

- C_{UMass} : The UMass coherence final score evaluates semantic coherence among topic words.
- N : The count of unique words in the evaluated topic.
- w_i, w_j : The topic’s indexed words for pairwise comparison.
- $D(w_i, w_j)$: Document frequency of both w_i and w_j , indicating co-occurrence.
- ϵ : A negligible positive value to prevent undefined logarithms.
- $D(w_j)$: Document frequency of w_j , the occurrence count of w_j .
- \log : Logarithmic function, normalizing the coherence score.
- \sum : Sum over word pairs.

Crucially, C_{UMass} quantifies thematic cohesion by summing over word pairs to evaluate their co-occurrence ($D(w_i, w_j)$) and individual frequencies ($D(w_j)$), adjusted by ϵ to ensure mathematical stability. This calculation, underpinned by the logarithmic ratio, offers insights into word semantic connections. The aggregate coherence score, derived from averaging these pairwise scores, reflects the topical harmony anchored in the corpus frequencies that originally informed the topic models. This approach ensures a nuanced, context-sensitive analysis of topics [28].

9.3.3 Zipf's Law in Vocabulary Refinement

The frequency distribution of words in human languages follows a fundamental and enthralling phenomenon within linguistic studies, epitomized by Zipf's Law presented in Formula (9.3). This distribution follows a systematic and universally acknowledged pattern: specific words occur with high frequency—such as “a”, “the”, and “you”—dominating functional words, whereas terms related to document semantics surface sporadically [21]. This pattern exhibits remarkable consistency across diverse natural languages and contexts, significantly influencing fields like natural language processing. We used such properties to automatize the distinction between functional words and useful terms for topic modeling.

Zipf's Law articulates this distribution as:

$$f(r) \propto \frac{1}{r^\alpha} \tag{9.3}$$

where:

- $f(r)$ represents the frequency of the term with rank r .
- α is an exponent, which in natural languages typically is close to 1.
- \propto denotes proportionality, indicating that the frequency of a word is inversely proportional to its rank in the frequency list, raised to the power of α .

9.3.4 Genetic Algorithm

Genetic algorithms are search procedures that are based on the principles of natural selection and genetics. These algorithms combine the survival of the fittest chain structures with certain innovative aspects of human search. In each generation, a new set of artificial “creatures” (chains) is generated using fragments from the fittest of the previous generation. Occasionally, new parts are incorporated for a better fit [12].

In the 1960s, John Holland pioneered genetic algorithms (GA) and collaborated with his students and colleagues at the University of Michigan during the 1960s and 1970s. Unlike approaches such as evolution strategies and evolutionary programming, Holland's original goal was not to design specific algorithms to solve particular problems. His primary focus was conducting a formal investigation of the adaptation process, as it manifests in nature, and developing methods to incorporate natural adaptation mechanisms into computer systems. Holland's approach and contributions are extensively discussed in Melanie Mitchell's work, “An introduction to genetic algorithms” where it is well explained that genetic algorithms guaranteed function optimization despite their random nature [16].

Genetic algorithms stand out compared to their conventional counterparts in the search for robustness, in four fundamental aspects. First, they employ a coded rep-

resentation of the parameter set instead of individual parameters. Second, instead of focusing on a single point, they explore a population of points. Additionally, they rely exclusively on the information of the objective function (Payoff) without using derivatives or other additional knowledge. Lastly, they guide their search process through probabilistic transition rules instead of deterministic rules [12].

According to [32], a genetic algorithm designed to address a specific problem consists of several essential components. First, a genetic representation is required to define how to encode potential solutions to the problem. Subsequently, a procedure is established for generating an initial the population of solutions, serving as a starting point for the algorithm. An evaluation function acting as a virtual environment plays a crucial role in rating the solutions in terms of their suitability, guiding the search. In addition, genetic operators are used that alter the composition of future generations, facilitating the evolution and adaptation of solutions throughout the process. Lastly, the values of various parameters of the genetic algorithm, such as the size of the population and the probabilities of applying the genetic operators are adjusted to optimize its performance in the specific context of the problem to be solved. In this sense, most genetic algorithms (GAs) share a series of fundamental components, such as populations of chromosomes, selection based on the fitness of individuals, crossover processes that generate new generations of offspring, and the introduction of random mutations in these offspring [16].

9.4 Optimizing Topic Generation in LDA: A Genetic Algorithm Approach for Automated Hyperparameter Tuning

This section details the methodological approach used to automate the hyperparameter tuning, which produced many topics found in the National Transparency Platform.

To clarify the methodological process used, Fig. 9.1 schematizes the main phases and the sequence followed by the proposal. This process ranges from collecting information requests, data preprocessing, and vocabulary optimization through Zipf's law to automating the LDA model's hyperparameters. The latter allows for identifying the main results: the prevalent topics in the citizens' information requests. Each stage is explained in detail in the rest of the section.

9.4.1 Information Gathering

The dataset for this study was meticulously compiled from the open data section of the National Transparency Platform, which focuses on federal entities obligated to adhere to transparency requirements. This research method deliberately omitted

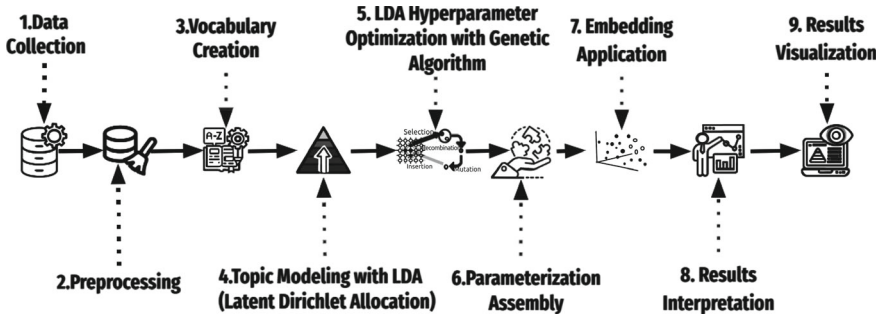


Fig. 9.1 Methodological process



Fig. 9.2 Infomex Platform and National Transparency Platform

data from non-federal entities, including state and local governments, legislative and judicial branches, political parties, unions, and similar bodies, to maintain a federal scope. The data were accessible in various formats, such as SQL, XML, CSV, and JSON, facilitating comprehensive analysis. The time frame for data collection spanned from 2003 to 2020, during which data retrieval initially utilized Infomex green [30], the precursor to the National Transparency Platform, until 2015. Subsequently, data collection shifted to the National Transparency Platform [31] for the remainder of the period Fig. 9.2 shows the two platforms. Throughout this extensive timeframe, approximately 2,518,875 public information requests were submitted to various federal departments, underscoring the vast scope of data analyzed.

The database contains full details on citizen inquiries sent to various governmental entities and their responses. It includes around 20 attributes; for more detailed information, see Table 9.1. This table provides an example of a request related to COVID-19. In this instance, the request involves seeking information on the availability of free tests, existing treatments and their costs, as well as the procedures for receiving medical care in Mérida, Yucatán, Mexico. This instance highlights the

Table 9.1 Attributes of the INAI requests database

INAI attribute	Description	Example
Folio	Unique 13-digit request identifier.	310572322000377
Fecha de solicitud	Date and time when the request was filed.	05/07/2022, 14:23
Dependencia	Name of the government department the request is sent to.	Servicios De Salud De Yucatán
Status	Current status of the request, e.g., in process.	Terminada
Medio de entrada	How the request was submitted: electronic or manual.	Electrónica
Tipo de solicitud	Nature of the request: public info, personal data, data correction.	Información pública
Descripción	Applicant's explanation of requested information.	Detalles de pruebas y tratamientos de COVID-19 en Mérida, Yucatán.
Otros datos	Additional details to aid information location.	N/A
Archivo adjunto de solicitud	URL for supplementary files provided by the applicant.	N/A
Medio de entrega	Preferred method for receiving the requested information.	Electrónico a través del sistema PNT
Fecha límite de respuesta	Deadline for the department to respond.	20/07/2022
Respuesta	Type of response provided by the government department.	Entrega de información vía PNT
Archivo de respuesta	Digital file provided by the department as part of the response.	N/A
Fecha de respuesta	Date the department responded.	14/07/2022
País	Country of the applicant's location.	México
Estado	State of the applicant's location within the country.	Yucatán
Municipio	Municipality of the applicant's location.	Mérida
Código Postal	Postal code of the applicant's location.	97098
Sector	Government sector of the addressed department.	Descentralizados

diverse and complex nature of unstructured information and the importance of understanding the specific needs and queries of citizens. In this research, we analyze and offer a detailed description of the requests submitted by interested parties. These requests, expressed through texts provided by the applicants, represent unstructured information.

9.4.2 INAI Information Request Preprocessing

The preparation of descriptions is a pivotal step in data management, involving the rigorous cleaning, transformation, and organization of the data to set the stage for further analysis and modeling. This process starts with creating JSON files, carefully categorized by federal entity and year. These files are crafted based on our database’s “state” column classification, excluding descriptions that fail to specify a particular federal entity, thereby focusing solely on those that explicitly identify one.

Subsequent steps include data cleaning removing special characters, numbers, and punctuation that could detract from the text’s clarity. Following this, tokenization is performed, a crucial procedure in text processing that is especially vital for data analysis and natural language processing (NLP) [23]. For instance, after cleaning preprocessing and tokenization, the sentence “I request information about COVID vaccines.” would dissect into tokens: [“I”, “request”, “information”, “about”, “COVID”, “vaccines”]. This division into smaller units simplifies the later stages of analysis and manipulation of the text, enabling more effective and efficient processing.

Moreover, this stage involves identifying and removing stop words that contribute limited informative value. A custom word filter is employed for this task, designed to exclude general terms, prepositions, polite expressions, legal terminology, and typical phrases found in descriptions. This filter is critical for refining the content, allowing the analysis to concentrate on the text’s most pertinent elements, and ensuring the analysis remains focused and significant.

9.4.3 Vocabulary Optimization

During this phase, we constructed a specialized vocabulary from terms in texts. We evaluated how often each term appears in our dataset and prioritized the most common ones, previous filtering of functional words. Zipf’s law enabled us to streamline our vocabulary. According to Zipf’s law, the frequency f of a word is inversely proportional to its rank r in the frequency table, We express this relationship with Eq. (9.4):

$$f(r) = \frac{a}{r^s}, \tag{9.4}$$

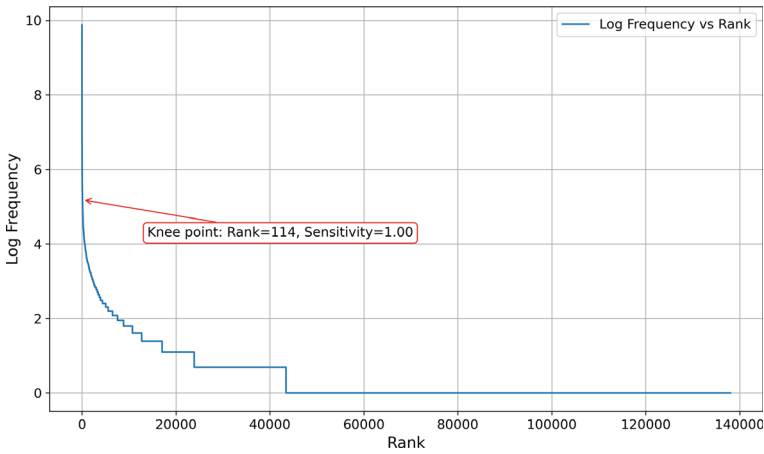


Fig. 9.3 Example of vocabulary optimization for requests from the state of Jalisco during 2020

where a is a constant and s is an exponent usually close to 1. By eliminating infrequent terms and categorizing less common terms as “unknown,” we efficiently condensed the vocabulary size, boosting the model’s performance without sacrificing accuracy or coverage.

To refine our vocabulary further, we applied a knee-point identification method to the frequency distribution. This technique involves steps such as importing data from a CSV file, calculating word frequencies, and applying a logarithmic transformation to these frequencies, as described in Eq. (9.5):

$$\log_f = \log(\text{frequency}), \quad (9.5)$$

and organizing them in descending order by rank. The transformed frequencies (\log_f) facilitate an easier examination of the distribution.

Following this, we illustrated the relationship between the words’ ranks and their logarithmic frequencies in a graph (refer to Fig. 9.3). This visualization is crucial for observing the frequency distribution and locating the “knee point”—the juncture that signifies the ideal vocabulary size. The KneeLocator tool, which uses an algorithm to detect significant changes in the curve’s slope, aids in identifying this point, marking where the frequency of word usage sharply declines.

Identifying the knee point allows us to define an optimal vocabulary size that includes the most impactful words up to the point of significant frequency reduction. This distinction aids in differentiating between common and rare terms, ensuring an effective analysis that retains the relevancy and integrity of the study on public information requests.

Subsequently, we filter the words to include only those at or below the knee point, exporting this refined selection to a text file. Additionally, we create a text representation of this streamlined vocabulary, formatting each word in quotes and separated by commas, readying it for integration into natural language processing (NLP) models.

9.4.4 Topic Modeling with LDA

In this crucial phase of the process, we implemented the Latent Dirichlet Allocation (LDA) model to generate a series of key files essential for the project's advancement. These key files include 'Corpus.mm', representing a structured collection of textual data; 'Dictionary.dict', a file mapping each unique word to a numeric identifier; 'Discarded Words.txt', containing a list of terms filtered out during the data cleaning phase; 'titles.txt', storing the titles or headings associated with each information request; and 'vocabulary.txt', a compilation of all relevant words identified in the corpus.

Equation (9.6) is employed for topic modeling based on LDA and other topic mixture models. Statistical modeling calculates how documents in a collection are generated in terms of a mix of several topics, where a "topic" is a distribution over a fixed vocabulary, and each document is considered a mix of various topics. The purpose of the formula is to calculate the probability of observing a specific word.

$$P(w_i) = \sum_{j=1}^T P(w_i|z_i = j) \cdot P(z_i = j) \quad (9.6)$$

- $P(w_i)$: Represents the marginal probability of encountering the word w_i .
- \sum : Signifies the summation over all categories j , from 1 to T , where T is the total number of topics.
- $P(w_i|z_i = j)$: Indicates the probability of the word w_i given topic j .
- $P(z_i = j)$: Reflects the a priori probability of selecting topic j .

For example, topic modeling in text analysis of public information requests reveals latent topics such as Waste Management, Biodiversity Conservation, and Renewable Energies, enhancing our understanding of key environmental concerns expressed in these requests.

- $P(w_i|z_i = j)$: This probability demonstrates how related a specific word w_i , like "recycling", is to a latent topic $z_i = j$, in this case, Waste Management. A high value indicates a significant association between the word and the topic.
- $P(z_i = j)$: This probability assesses how represented a specific latent topic z_i is within a request. Thus, if a request addresses Renewable Energies themes, the probability of this topic being represented ($z_i = \text{Renewable Energies}$) will be high.

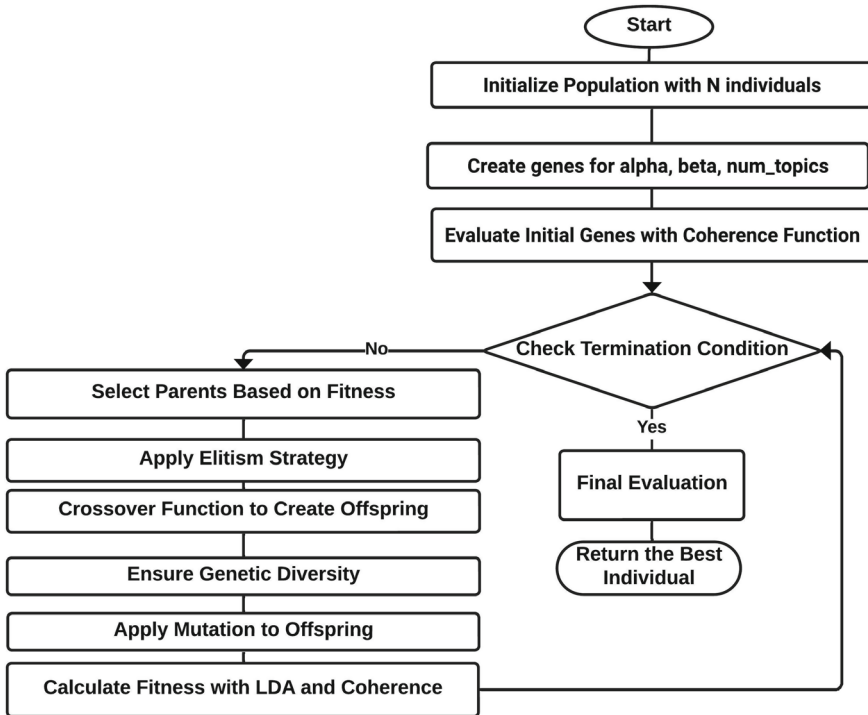


Fig. 9.4 Genetic algorithm flowchart

9.4.5 Deployment and Implementation of a Genetic Algorithm

We leveraged a specific version of the genetic algorithm for the efficient optimization of α and β parameters and the precise selection of the optimal number of topics, with the process evaluated on the strength of the topic coherence score. Figure 9.4 provides a flowchart outlining this optimization. The genetic algorithm harnesses a specific configuration to navigate the search space effectively: a population size of 20 for diversity, 50 generations to ensure ample evolution, selection of 4 parents for a balanced genetic mix, retention of 2 elite solutions to preserve superior genes, and a mutation rate of 0.05 to foster new traits. Furthermore, it explores a gene space that includes α and β parameters within a range of 0.01–1.0, and the number of topics between 2 and 50, ensuring the algorithm meticulously tunes the model parameters and identifies the most coherent topics for analysis. The effectiveness of the algorithm is supported by statistical tests, details of which are provided in the referenced work [13]. This streamlined approach guarantees that the genetic algorithm not only fine-tunes the model with precision but achieves this with optimal efficiency.

Table 9.2 Identifier definitions for the population initialization function

Identifier	Definition
<code>pop_size</code>	The population size to be initialized
<code>parameter_ranges</code>	A dictionary with parameter names as keys and tuples as values, where each tuple contains the minimum and maximum range for the corresponding parameter
<code>population</code>	A list that stores the population, where a list of parameter values represents each individual
<code>individual</code>	A list representing a single individual in the population, generated within the specified parameter ranges. The last parameter is cast to an integer, assuming it's the number of topics ('num_topics')

9.4.5.1 Initializing the Population

The 'initialize_population' function creates an initial population precisely, as detailed in Table 9.2. It utilizes two critical inputs: 'pop_size', which specifies the total number of individuals to be generated, each representing a potential solution within the designated parameter space, and 'parameter_ranges', offering a framework to ensure solutions comply with predefined constraints.

Starting with an empty 'population,' the function iterates until 'pop_size' is reached, each iteration creating an 'individual' by randomly assigning values within 'parameter_ranges' for each parameter using the 'random.uniform' function. This method guarantees that the values fall within the allowed limits. Special focus is placed on the 'num_topics' parameter, which is converted to an integer to match the optimization problem's requirements. This step ensures the solution's integrity, confirming that each individual's parameters meet the model's standards. This approach details the population's generation process and emphasizes adhering to the model's parameters to produce viable solutions.

9.4.5.2 Coherence Fitness Function

The relevance and coherence of topics derived from topic modeling algorithms are crucial for their application in real-world scenarios. The evaluate_coherence function addresses this by offering a measurable indicator of an LDA model's coherence, a concept emphasized in research by [19, 28]. This metric is essential for assessing the utility of topic models.

Table 9.3 Identifiers and their definitions in the evaluate coherence function

Identifier	Definition
alpha	Hyperparameter that controls the mixture of topics within documents. A lower value leads to documents containing fewer topics
beta (eta)	Hyperparameter that controls the distribution of words across topics. A lower value means fewer words represent a topic
num_topics	The number of topics to extract from the corpus
corpus	A collection of documents represented as a list of lists of tuples (document, word frequency)
id2word	A dictionary mapping word IDs to words
texts	The set of original documents, preprocessed and ready for analysis
lda_model	A Latent Dirichlet Allocation (LDA) model built on the corpus
coherence_model_lda	A model for calculating the coherence of the LDA model, assessing how well the inferred topics correspond to the texts

The `evaluate_coherence` function aims to verify the coherence of topics produced by an LDA topic model, ensuring they are both interpretable and cohesive. This verification confirms the topics' relevance. Key parameters that influence the model's structure and its coherence evaluation include `alpha`, `beta (eta)`, and `num_topics`, as detailed in Table 9.3.

The LDA model is created using Gensim's `LdaModel` module. This module focuses on the algorithm's iteration count through the dataset during training, a vital factor for model quality. The `CoherenceModel` class from the Gensim python module is then used to set up a `coherence_model_lda`, which aids in assessing topic coherence.

Moreover, a fitness function is introduced to calculate and return the fitness values for model populations, evaluating each based on its ability to generate coherent and meaningful topics. This process, which begins with creating a fitness list to store each model's coherence scores, underscores the model's success in producing relevant and interpretable topics.

9.4.5.3 Parent Selection

Parent selection is a critical component in genetic algorithms, significantly influencing the evolutionary process by favoring individuals with higher fitness to breed the next generation. This approach promotes the evolution of the population towards better solutions by assuming that fitter individuals will likely transfer their superior traits to their offspring. The implementation leverages `numpy` for efficient array manage-

Table 9.4 Identifier definitions for the parent selection function

Identifier	Definition
<code>population</code>	The complete set of individuals in the current generation, from which parents are selected
<code>fitness</code>	An array or list of fitness values, each corresponding to an individual in the population, used to assess their suitability
<code>num_parents</code>	The specified number of parents to be selected from the population for the next generation
<code>parents</code>	The subset of the population selected as parents for the next generation, chosen based on their high fitness values

ment, where individuals are ranked by fitness using `np.argsort()` on the `fitness` array, and the top `num_parents` are chosen for reproduction. This selection ensures the perpetuation of the most capable traits, aligning with the principles outlined in Table 9.4.

9.4.5.4 Crossover

Genetic crossover, pivotal in genetic algorithms, combines traits from two parents to produce offspring, facilitating solution space exploration and potential fitness improvements in future generations. It employs a midpoint crossover technique within the offspring’s parameter vector, delineated in Table 9.5, to divide genetic contributions from both parents.

Offspring generation involves iterating a loop corresponding to the desired number of offspring, with each cycle creating a new individual. Parental selection for crossover uses indices from the table, ensuring equitable parent contribution and cyclic selection. Offspring results from merging genes before and after the crossover point from respective parents via `np.concatenate`, enabling inheritance of beneficial traits and aiming for ongoing population fitness enhancement.

9.4.5.5 Mutation

To enhance the genetic diversity within a population and mitigate premature convergence to local optima, mutation is a pivotal mechanism in genetic algorithms. It systematically introduces genetic variations by adjusting individual parameters, thereby facilitating the exploration of new solution spaces. The mutation process is governed by a mutation rate, dictating each individual’s mutation probability to ensure a dynamic yet controlled exploration.

Table 9.5 Identifier definitions for the crossover function in a genetic algorithm

Identifier	Definition
parents	An array containing the genetic information of the current generation's parents from which offspring will be derived
offspring_size	A tuple specifying the desired number of offspring (<code>offspring_size[0]</code>) and the dimensionality of each offspring (<code>offspring_size[1]</code>), guiding the size and structure of the generated offspring array
offspring	The resulting array of offspring generated by the crossover operation, each inheriting genetic traits from two parents
crossover_point	A calculated point (half the offspring's second dimension size) that determines the split in genetic information between two parents for each offspring
parent1_idx, parent2_idx	Index variables used to select pairs of parents for genetic material combination, ensuring a diverse genetic mix in the offspring generation

The mutation operation commences with evaluating each individual against the mutation rate, determined by generating a random number between 0 and 1. Should this number fall below the mutation rate threshold, the individual is earmarked for mutation. A parameter for mutation is then randomly selected from the set of mutable parameters, as delineated in the table referred to by '9.6'. For parameters requiring specific conditions—such as 'num_topics', which necessitates an integer value—a new value is randomly assigned within the predefined range. This procedure ensures the introduction of genetic variability and adheres to the constraints of valid parameter values, thereby enriching the genetic algorithm's potential for uncovering novel solutions.

9.4.5.6 Genetic Diversity

Genetic diversity is essential for averting premature convergence to local optima and enhancing the algorithm's capacity to navigate the solution space efficiently. This mechanism assesses each individual's contribution to the population's genetic variance by examining their fitness against a stipulated threshold. When an individual's fitness falls below this threshold, it signifies a potential shortfall in augmenting the population's diversity. Consequently, such individuals undergo a regeneration process, receiving a new array of parameters generated within predefined limits, as detailed in Table 9.7.

Table 9.6 Identifier definitions for the mutation function in a genetic algorithm

Identifier	Definition
offspring	The array of offspring to be mutated, representing individuals of the current generation
parameter_ranges	A dictionary specifying the allowable range for each parameter that can be mutated, used to ensure mutations result in valid values
mutation_rate	The probability of any given individual undergoing mutation, controlling the frequency of mutations within the population
param_to_mutate	The parameter selected for mutation, chosen randomly from the keys of the <code>parameter_ranges</code> dictionary
index	The position of the mutating parameter within an individual's parameter list, used when applying mutations to parameters other than 'num_topics'

Table 9.7 Identifier definitions for the function aimed at controlling population diversity based on fitness in a genetic algorithm

Identifier	Definition
population	The array of individuals constituting the current generation, subject to diversity control based on fitness evaluation
fitness	An array of fitness scores corresponding to each individual in the population, utilized for assessing the need for diversity introduction
threshold	A predetermined fitness score threshold; individuals with fitness below this value are considered for genetic diversity enhancement
parameter_ranges	A dictionary outlining the permissible range for each genetic parameter, ensuring modifications remain within viable bounds

This rejuvenation process entails assigning each parameter a random value from its allowable range, thus ensuring the revamped individual adheres to the established parameter boundaries. By substituting individuals with fitness levels beneath the set threshold with newly generated ones, this approach systematically promotes genetic heterogeneity within the population. It guarantees that the population retains a baseline level of quality while reintroducing variation, facilitating ongoing exploration and mitigating the risks of evolutionary stagnation.

Table 9.8 Identifiers and their definitions in the elitism function

Identifier	Definition
Population	The collection of all individuals in the current generation
Fitness	A list of fitness scores for each individual in the population
num_elites	The number of top performers to select
elites_idx	The indices of the selected elites based on fitness
elites	The selected top-performing individuals

9.4.5.7 Elitism

Elitism in genetic algorithms preserves high-quality solutions across generations by protecting them from being lost during random selection, crossover, or mutation. This method starts by sorting individuals based on their fitness values using `'np.argsort(fitness)'`, from lowest to highest, to identify the most fit individuals or elites. The top individuals, indicated by the `'num_elites'` parameter, are then selected by slicing the sorted array. These elites are extracted from the population and maintained for future generations, as detailed in Table 9.8. This approach ensures the continuous preservation of superior solutions, reducing the risk of their loss due to genetic operations.

9.4.6 Topic Identification by State

The generation of specific topics for each state is underway. The key to this generation lies in the values obtained automatically through the implementation of the genetic algorithm. By adjusting the hyperparameters of alpha, beta, and topic, the corresponding topics for the 32 states of the Mexican Republic is obtained for each year of the analyzed period. The optimization of hyperparameters is paramount, exerting a direct influence on both the granularity and applicability of the derived topics. Ensuring that these topics are not only statistically robust but also of substantive practical relevance is critical. This rigorously tailored approach facilitates an in-depth understanding of regional nuances, empowering policymakers and researchers to discern and interpret trends and transitions in public interests and issues effectively.

9.4.7 Interpretation of Results

In the field of text analysis through Latent Dirichlet Allocation (LDA), we face the significant challenge of assigning accurate titles and formulating appropriate descriptions for identified topics. This challenge arises from the crucial need to interpret and synthesize the thematic complexities and patterns extracted through LDA analysis in a coherent and concise manner. To address this issue, the study referenced in [22] employs the Generative Pre-training Transformer (GPT) model to enhance the interpretation of identified topics. This method proposes a three-phase strategy: initially, topics are identified using LDA; then, a domain expert assigns specific themes to these topics; and, finally, GPT is used to generate clear and comprehensible descriptions.

To optimize the task of interpreting and describing topics, our methodology emphasizes the detailed analysis of keywords and their associated probabilities, aiming for an in-depth understanding of each topic. The automatic generation of titles and descriptions is carried out using GPT models, complemented by a human validation phase to ensure the accuracy and relevance of the results. In this process, word clouds emerge as an indispensable tool, providing an intuitive visual representation that highlights the most significant words of each topic. This approach not only facilitates the creation of more accurate titles and descriptions but also ensures a deeper connection with the analyzed content, thereby improving the clarity, structure, and coherence of the text. Below, we present an example illustrated in Fig. 9.5, where the word cloud reveals a variety of terms associated with environmental themes, informational content, and legal or regulatory action. Words like “resources”, “marine”, “protected”, “environment”, and “natural” suggest a focus on conserving or regulating natural or marine areas. For a more detailed interpretation, keywords and their interpretation are used, which are processed through the GPT-3.5-turbo model, as shown in Table 9.9.

9.4.8 Application of Embedding to Identify Topic Similarity

At this stage, embeddings for the topic descriptions are generated, which are vector representations of the themes’ interpretation using the word embedding model known as “sentence- embeddings-BETO”. The process involves calculating the cosine similarity between each pair of topic descriptions, setting a similarity threshold greater than 0.8. Subsequently, topics are clustered based on their similarity using the K-means algorithm, which allows for organizing and segmenting the themes along with their respective status. Finally, based on the clustered titles, a name that encompasses the topics is assigned to each group. Figure 9.6 shows an example of topic similarity by status.



Fig. 9.5 Example topic: environmental resources in protected areas

Table 9.9 Topic example: title and description generated by GPT-3.5-turbo

Title	Keywords and probability	Description
Management of Water Resources and Rights	Simple (0.090), general (0.034), resources (0.033), yes (0.020), years (0.020), natural areas (0.018), all (0.018), actions (0.017), protected (0.015), federal (0.014), marine (0.013), indicate (0.012), SARS virus (0.011), information (0.010), number (0.008), foreign (0.007), return (0.007), agent (0.007), plan (0.007), environment (0.007), perform (0.007), activities (0.007), etc.	This topic suggests a strategy focused on the conservation of natural resources, highlighting the simplicity and scope of marine and terrestrial protection measures. It focuses on the planning and review of federal policies, possibly influenced by environmental and regulatory factors, in a context marked by digital information management and ecosystem health surveillance

9.4.9 Visualization of Results

Ultimately, our goal is to present the findings of our analysis in an intuitive and captivating way. To achieve this, we will harness various visual tools, each carefully chosen for its unique ability to simplify and effectively convey complex information. Word clouds will spotlight the dominant themes by visually emphasizing key terms based on their frequency, offering a snapshot of the dataset’s core subjects. Geo-

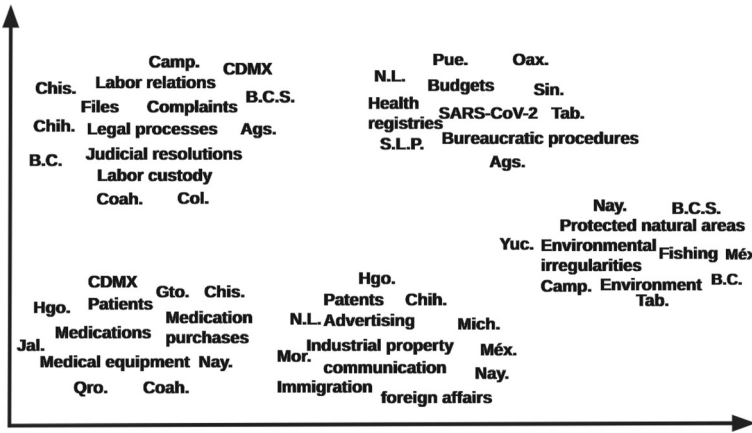


Fig. 9.6 Example of topic similarity

graphic maps will contextualize spatial data, illustrating trends and distributions that reveal regional insights. In contrast, heat maps will visualize data intensity across different dimensions, pinpointing high activity or concentration areas. Additionally, similarity diagrams will map out the relationships and patterns among data entities, uncovering clusters and connections that elucidate the underlying structure of the dataset. This multifaceted visual approach aims to transform intricate datasets into accessible insights, bridging the gap between detailed analysis and practical understanding for a diverse audience, thereby ensuring that our findings’ significance is recognized and deeply comprehended.

9.5 Results

Figure 9.7 illustrates the fitness evolution within a genetic algorithm over 50 generations, applied to a dataset of 2580 instances from the state of Puebla. An upward trend is noticeable in the early stages, suggesting that the algorithm was capable of quickly identifying promising candidates and significantly improving the quality of solutions compared to the initial generations. Following this initial increase, the fitness curve stabilizes, which is typical in genetic algorithms, as the population tends to converge towards a local or global optimum solution.

The best individual that emerged from this process exhibited a parameter configuration with an alpha of 0.18 and a beta of 0.81, handling 8 different topics in its model. This set of parameters reached a coherence value of 0.73, which is a robust indicator that the solution found is significant and well-structured, according to the metrics of the LDA model used to measure the coherence among topics. The coherence value is particularly notable, as it indicates strong relevance and distinction between the top-

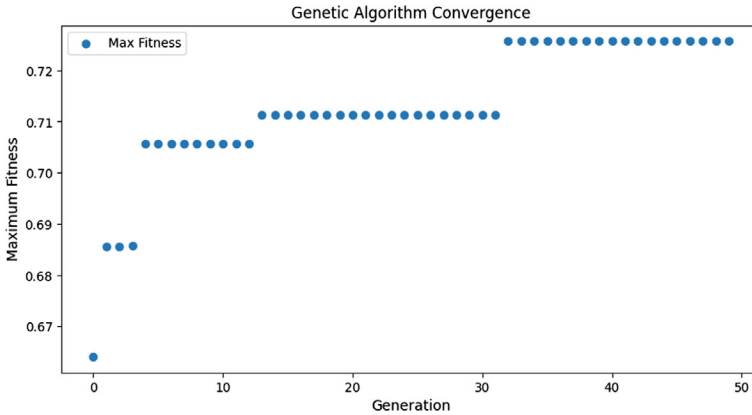


Fig. 9.7 Evolution of performance across generations

ics generated by the LDA model. This means that the topics are mutually exclusive and each captures a distinct set of information from the dataset.

To obtain results and validate the effectiveness of the the proposed methodology, data from the year 2020 were analyzed, which included a total of 97,360 public information requests. In Table 9.10, the values of the hyperparameters alpha, beta, and topic are presented, along with their respective coherence values for each federal entity. The detailed analysis of this table reveals an interesting trend: although in some cases it is observed that a higher number of requests corresponds to an increase in the number of topics, this relationship does not remain uniform across all instances. This finding suggests variability in the distribution of topics that is not directly correlated with the volume of requests in all entities.

In Fig. 9.8, a detailed analysis of the predominant topics in the state of Baja California throughout the year 2020 is presented. Based on the analysis of keywords and their respective probabilities, it is identified that Topic 0 focuses on the management of resources in protected natural areas and environmental issues. This topic encompasses crucial aspects such as biodiversity preservation and marine ecosystem protection.

The various topics identified in the analysis weave together a comprehensive narrative on the intricate dynamics of social, environmental and economic activities. Topic 1 unveils irregularities in social and environmental resources, spotlighting key aspects such as results, projects, programs, and expenditures, setting the stage for a deep dive into governance and resource management. Transitioning smoothly into Topic 2, the focus shifts to the fishing industry, encapsulated through reports and fishing records, indicating a specialized examination of this sector’s operational facets.

As the narrative progresses, Topic 3 brings to light discussions from social cabinet sessions and presidential decisions, with keywords like discussions and agreements underlining the political and decision-making processes that impact both social and

Table 9.10 Hyperparameter results by entity in 2020

Entity	Num. of requests	Alpha	Beta	N Topics	Coherence score
Ags.	533	0.28	0.49	5	0.62
B.C.	2540	0.68	0.22	9	0.84
B.C.S.	483	0.39	0.43	4	0.70
Camp.	205	0.9	0.39	4	0.74
Chis.	370	0.71	0.66	5	0.61
Chih.	1354	0.85	0.87	4	0.52
CDMX	56,091	0.26	0.79	56	0.82
Coah.	651	0.56	0.45	10	0.54
Col.	234	0.67	0.47	6	0.59
Dgo.	346	0.85	0.6	3	0.60
Mex.	9740	0.28	0.75	9	0.66
Gto.	916	0.9	0.17	9	0.50
Gro.	351	0.86	0.54	7	0.65
Hgo.	859	0.77	0.46	4	0.51
Jal.	4109	0.92	0.6	5	0.46
Mich.	386	0.2	0.19	3	0.64
Mor.	904	0.99	0.14	5	0.52
Nay.	633	0.3	0.61	3	0.51
N.L.	1111	0.91	0.95	7	0.45
Oax.	554	0.84	0.35	6	0.49
Pue.	2580	0.18	0.81	8	0.73
Qro.	1503	0.76	0.84	9	0.54
Q. Roo	847	0.18	0.76	4	0.60
S.L.P.	410	0.53	0.93	3	0.64
Sin.	1590	0.51	0.6	7	0.55
Son.	1139	0.91	0.27	6	0.63
Tab.	2150	0.23	0.48	3	0.55
Tamps.	1003	0.72	0.84	7	0.58
Tlax.	480	0.05	0.81	4	0.68
Ver.	1767	0.85	0.1	7	0.48
Yuc.	1339	0.72	0.94	8	0.58
Zac.	182	0.29	0.07	5	0.51

Source Own elaboration with data from INAI (2023)

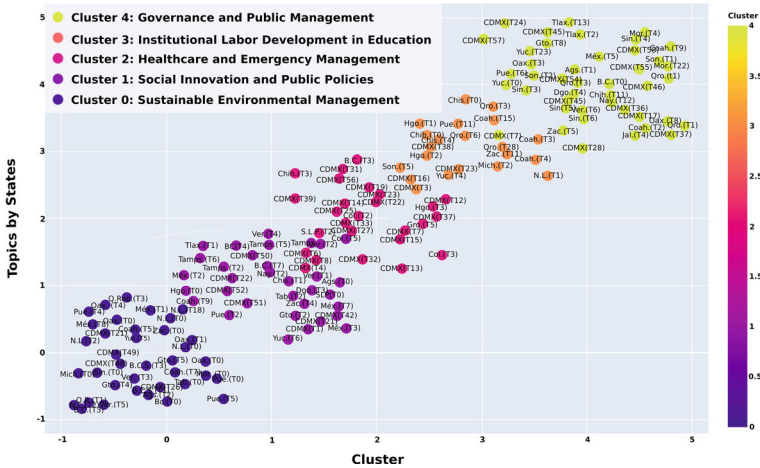


Fig. 9.8 Topics from the state of Baja California during 2020

environmental policies. This the political and administrative lens is complemented by Topic 4, which narrows down on fishing activities during October and September, especially in the coastal areas of Campeche and Veracruz and the challenges posed by the SAR virus, reflecting on the sector’s resilience and adaptability to health crises.

Further deepening the analysis, Topic 5 explores the SAR virus’s impact on the fishing sector and the health of workers in August, revealing the virus’s broader implications on occupational health and sector stability. Topic 6 then takes a different turn, focusing on the development project on Salsipuedes Island, where aspects such as transportation plans, budgets and execution details highlight the logistical and financial planning essential for environmental and infrastructural projects.

In addressing the accountability and labor aspects, Topic 7 delves into project responsibility and labor demands, emphasizing the need to identify responsible parties and manage labor and costs efficiently, showcasing the human resource and financial aspects critical to the project success. The narrative culminates with Topic 8, which zooms in on the details about employees, including names, salaries, and expenses, providing a granular view of the workforce component in these endeavors. Together, these topics paint a multifaceted picture of the challenges and considerations involved in managing social, environmental, and economic activities within a complex and interconnected framework.

By applying embedding techniques to the topic descriptions and their subsequent grouping using the K-Means algorithm based on similarities, as illustrated in Fig. 9.9, we have achieved significant results, especially in thematic categorization by geographical location.

The first cluster, related to Environmental Management and Sustainable Development encompasses a wide variety of topics focusing on environmental management, conservation of natural resources, sustainable development, and regulations in var-

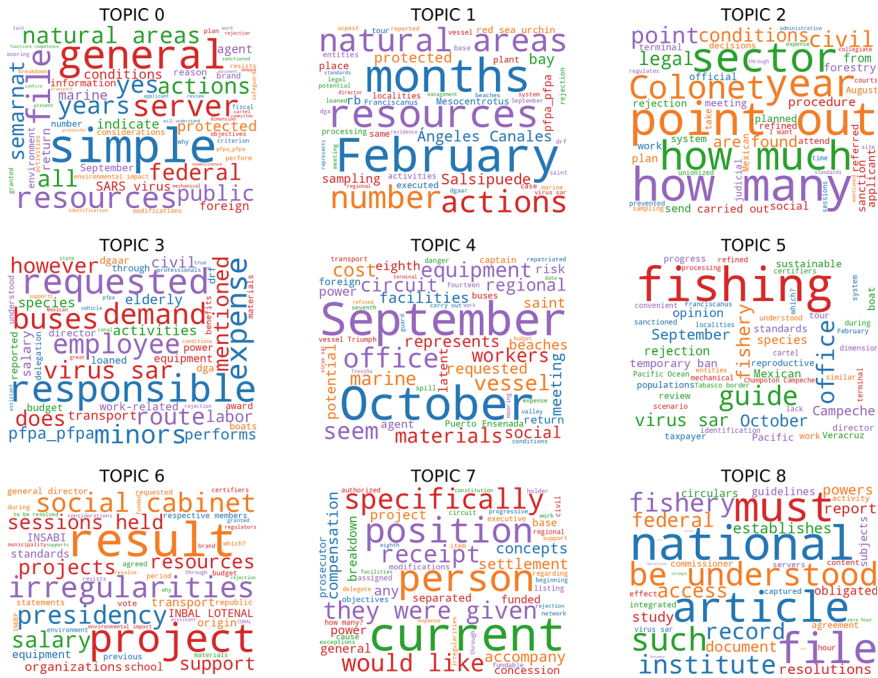


Fig. 9.9 Similar topics by state during 2020

ious sectors, such as energy, agriculture, tourism, and water management. Topics range from the authorization and management of activities in official bodies to the environmental impact assessment, development of national projects, water resource management, species conservation, and environmental compensation strategies in the gas industry. It also includes aspects related to sustainability in fishing, regional development in Oaxaca, handling of clandestine operations, and environmental protection in urban and rural projects and developments.

The second cluster, Social Innovation and Public Policies in Mexico, addresses a broad spectrum of topics related to implementing and evaluating social programs, social security management, labor and migration dynamics, and other key aspects of public policy and administration in Mexico.

The third cluster, Healthcare Management and Emergency Response, focuses on a series of topics primarily linked to public health management, sanitary and commercial regulations, the impact of the COVID- 19 pandemic on various sectors, and the distribution and oversight of medications. It also covers aspects related to public transport, import and export regulations, and protective measures and management during health emergencies. The topics reflect a concern for ensuring the population’s well-being, adapting policies and procedures to emerging challenges, and maintaining effectiveness in the management of essential services in times of crisis.

The fourth cluster, Institutional and Labor Development, focuses on a wide range of topics related to institutional and labor management, electoral processes, labor regulations, professional and educational development, and the impact of community and educational activities on various societal sectors. Topics span from public administration and general management in specific regions like Michoacán and Coahuila to the professional integration of students and interns through to specific challenges in higher secondary education and the management of educational projects.

The fifth cluster centers on topics related to public information management, copyright, document management in public entities, sanitary regulations, transparency, auditing, and control in various public sector areas, including education, health, and judicial. It also addresses resource and contract management, anti-money laundering efforts, import and export regulations, labor rights, and the management of files and official documentation, among others.

9.6 Discussion

The genetic algorithm has shown remarkable effectiveness in optimizing the LDA model's Alpha, Beta, and Topic parameters, significantly boosting its ability to analyze data. In parallel, embedding techniques have revealed thematic connections, offering great potential for developing proactive transparency strategies and improving open data access. This innovative approach, particularly applied to data from the National Transparency Platform, has unveiled hidden patterns and trends, marking a substantial step forward in enhancing transparency and the accessibility of public information in Mexico.

This method has identified 4131 topics across various federal entities from 2003 to 2020 by optimizing key LDA hyperparameters. These findings are crucial for meeting citizen needs and promoting transparency and open data availability, showcasing remarkable progress in processing and analyzing large volumes of government data. Furthermore, these advancements empower the government to analyze and utilize data more effectively and bolster public trust by making governmental processes more transparent and accessible. This approach serves as a benchmark for refining data management practices. It catalyzes increased public engagement in governance, strengthening the foundations of democracy and accountability in Mexico.

9.7 Conclusion and Future Research Directions

In conclusion, this study has successfully demonstrated the feasibility of automating the selection of hyperparameters (alpha, beta, and the number of topics) in the Latent Dirichlet Allocation (LDA) model through the implementation of a genetic algorithm, using topic coherence as the fitness criterion. This methodology has efficiently identified 4131 relevant topics annually and by state throughout the analyzed

period. Moreover, the integration of GPT models for generating titles and descriptions of topics has facilitated their validation by transparency experts, achieving significant time savings. Although the analysis encountered limitations in states with few requests or little thematic variability, this work lays the groundwork for applying it to any unstructured data set requiring topic identification.

The methods and insights from this study could be applied across a wide range of research fields, such as social media analysis, market research, and biomedical literature review, where understanding large volumes of text data is crucial. The potential real-world applications of our work are vast, including enhancing search engine algorithms, improving recommendation systems, and aiding in curating content for personalized newsfeeds, significantly impacting how information is organized and retrieved.

One limitation of our study is the reliance on available digital data, which may not fully represent the spectrum of topics of interest, particularly in areas with lower digital footprints.

Future research could focus on adapting the genetic algorithm to address the issue of redundant topics in LDA models by incorporating mechanisms that penalize or eliminate overlapping topic clusters. This adjustment could significantly improve the distinctiveness and relevance of the identified topics. Additionally, efforts could be directed toward fine-tuning the algorithm to minimize the inclusion of topics characterized by words with low probability, thereby enhancing the overall coherence and interpretability of the topics. Exploring these modifications would refine the precision of topic identification and contribute to a deeper understanding of topic distribution and separation in large datasets.

Drawing from these findings, future comparisons of this methodology with other topic modeling techniques, such as BERTopic, represent a promising direction to expand these findings. Additional research could also explore the application of these methodologies to different datasets or within varied geopolitical landscapes, which could unveil distinct patterns of public interest and information solicitation on a global scale.

The topics of public interest identified in this study, including environmental management, public policy initiatives, and responses to health emergencies, underscore a significant opportunity for shaping public policy formulation. Future endeavors could investigate avenues through which these insights can be systematically woven into the fabric of policy-making processes, ensuring that governmental actions are in tight congruence with the populace's concerns, thereby fostering policies that are both responsive and reflective of the citizens' needs.

Data Availability Statement

All data and code pertaining to this study are published by INAI. To access the datasets, please visit <https://www.plataformadetransparencia.org.mx/>. For the associated code, visit <https://github.com/hermecp/AutoTopicGen-INAi> (last accessed on May 29, 2024).

References

1. Aguilar Miranda, A., Ramírez González, K.: La plataforma nacional de transparencia en México y la gestión municipal. In: XXIV Congreso Internacional del CLAD sobre la reforma del Estado y de la Administración Pública. Buenos Aires, Argentina (2019)
2. Bagozzi, B.E., Berliner, D., Almqvist, Z.W.: Predicting government (non) responsiveness to freedom of information requests with supervised latent dirichlet allocation (2016). <https://api.semanticscholar.org/CorpusID:43093099>
3. Bagozzi, B.E., Berliner, D., Almqvist, Z.W.: When does open government shut? predicting government responses to citizen information requests. *Regulation and Governance* (2019). <https://doi.org/10.1111/rego.12282>
4. Bautista-Farías, J.: La nueva ley general de transparencia: alcances y retos. *DFH - Revista Análisis Plural* (2015)
5. Berliner, D., Bagozzi, B.E., Palmer-Rubin, B.: What information do citizens want? evidence from one million information requests in Mexico. *World Development* (2018). <https://doi.org/10.1016/j.worlddev.2018.04.016>
6. Berliner, D., Bagozzi, B.E., Palmer-Rubin, B., Erlich, A.: The political logic of government disclosure: Evidence from information requests in Mexico. *The Journal of Politics* **83**, 229 – 245 (2020) <https://doi.org/10.1086/709148>
7. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Journal of Machine Learning Research* **3**, 993–1022 (2003)
8. Caballero, J.A., Carbonell, M., Fix-Fierro, H., López Ayllón, S., Roldán Xopa, J., Salazar Ugarte, P.: El futuro del instituto federal de acceso a la información pública y protección de datos personales. consideraciones sobre su autonomía constitucional. México, UNAM (2012)
9. Cárdenas Sánchez, J., Gillo de la Cruz, M.G.: Eficacia institucional de los organismos independientes pro-rendición de cuentas: el caso del instituto nacional de transparencia, acceso a la información y protección de datos personales (inai). *Estudios En Derecho a La Información* **1**(17), 35–59 (2023). <https://doi.org/10.22201/ij.25940082e.2024.17.18781>
10. De Diputados, C.: Constitución política de los estados unidos mexicanos. Cámara de diputados, México (2012)
11. DOF: Ley general de transparencia y acceso a la información pública. Recuperada de <http://www.diputados.gob.mx/LeyesBiblio/pdf/LGTAIP.pdf> (2015)
12. Goldberg, D.E.: Optimization, and machine learning. *Genetic algorithms in Search* (1989)
13. Kuri-Morales, A.F., Aldana-Bobadilla, E., López-Peña, I.: The best genetic algorithm ii: A comparative study of structurally different genetic algorithms. In: Mexican International Conference on Artificial Intelligence, pp. 16–29. Springer (2013)
14. López Ayllón, S.: El derecho a la información como derecho fundamental. *Derecho a la información y derechos humanos* (2000)
15. López-Ayllón, S.: La creación de la ley de acceso a la información en México: una perspectiva desde el ejecutivo federal. Hugo A. Concha Cantú, Sergio López Ayllón y Lucy Tacher Epstein,(Coords.), *Transparentar al Estado: La experiencia Mexicana de Acceso a la Información*, Instituto de Investigaciones Jurídicas, UNAM (2005)
16. Mitchell, M.: An introduction to genetic algorithms. MIT press (1998)
17. Nash Rojas, C., Chacón Fregoso, G., Rodríguez Atero, M.: Estudio comparado sobre el impacto que tienen las instituciones que resguardan el acceso a la información pública en Chile y México sobre los derechos humanos en la ciudadanía. Available at <https://repositorio.uchile.cl/handle/2250/142492> (2016)
18. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108 (2010)
19. Omar, M., On, B.W., Lee, I., Choi, G.S.: Lda topics: Representation and evaluation. *Journal of Information Science* **41**(5), 662–675 (2015). <https://doi.org/10.1177/0165551515558783>

20. Pathik, N., Shukla, P.: Simulated Annealing Based Algorithm for Tuning LDA Hyper Parameters, pp. 515–521 (2020) https://doi.org/10.1007/978-981-15-4032-5_47
21. Piantadosi, S.T.: Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **21**(5), 1112–1130 (2014) <https://doi.org/10.3758/s13423-014-0585-6>
22. Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., Kaymak, U.: Towards interpreting topic models with chatgpt. In: *The 20th World Congress of the International Fuzzy Systems Association* (2023)
23. Rivera, G., Florencia, R., García, V., Ruiz, A., Sánchez-Solís, J.P.: News classification for identifying traffic incident points in a spanish-speaking country: A real-world case study of class imbalance learning. *Applied Sciences* **10**(18), 6253 (2020). <https://doi.org/10.3390/app10186253>
24. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: *Proceedings of the eighth ACM international conference on Web search and data mining*, pp. 399–408 (2015) <https://doi.org/10.1145/2684822.2685324>
25. Ruelas, A.: La transparencia en México: un trabajo colectivo. Consultada en www.library/fes.de/pdf-files/bueros/mexiko/12451.pdf (2016)
26. Salas Suárez, J.: El papel de los órganos garantes del acceso a la información pública en el contexto del Estado abierto. No. 44751 in *Libros de la CEPAL. Naciones Unidas Comisión Económica para América Latina y el Caribe (CEPAL)* (2017). <https://ideas.repec.org/b/ectr/col015/44751.html>
27. Sandoval Ballesteros, I.E.: *Leyes de acceso a la información en el mundo*. Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (2008)
28. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 952–961 (2012)
29. Tekin, Y.: Optimization of lda parameters. 2020 28th Signal Processing and Communications Applications Conference (SIU) pp. 1–4 (2020) <https://doi.org/10.1109/SIU49456.2020.9302034>
30. Instituto Nacional de Transparencia, A.a.I.I.y.P.d.D.P.I.: Plataforma infomex (2007). <https://www.infomex.org.mx/gobiernofederal/homeOpenData.action>. Plataforma utilizada para solicitar información pública en México
31. Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI): Plataforma Nacional de Transparencia (2023). https://www.plataformadetransparencia.org.mx/web/guest/datos_abiertos. Portal para el acceso a información pública y datos abiertos del Gobierno de México
32. Zbigniew, M.: Genetic algorithms + data structures = evolution programs. *Comput. Stat.*, pp. 372–373 (1996)

Chapter 10

Analysis of Accuracy on Data Visualization Techniques for Multi-objective Algorithm Performance Based on Convergence and Diversity Towards the Pareto Frontier



Manuel Paz-Robles , Claudia Gomez-Santillan , Nelson Rangel-Valdez ,
Ma. Lucila Morales-Rodriguez , and Georgina Castillo-Valdez 

Abstract Understanding the behavior of strategies that generate solutions for optimization problems with three or more objectives, such as evolutionary algorithms, is crucial for researchers. One practical approach to achieving this understanding is by the comparison of visualizations of the set of solutions that approximate the set of optimal solutions that address optimization problems with three or more objectives. Visualizations support the depiction of the approximate Pareto front shape, the establishment of relationships between objectives, the distribution of solutions, and the representation of convergence and diversity levels of algorithms. This chapter reviews the visual representations of algorithm performance with data of up to four dimensions, mainly analyzing convergence and diversity using Scatter Plot Matrix (ScPM), Heatmaps (HM), Parallel Coordinates Plots (PCP), and Radar Plots (RP) with state-of-the-art visualization techniques. As a result, this work contributes with identified shortcomings for a good visualization based on features such as accuracy,

M. Paz-Robles (✉) · C. Gomez-Santillan · N. Rangel-Valdez · Ma. L. Morales-Rodriguez
Tecnológico Nacional de México, Instituto Tecnológico de Ciudad Madero, Ciudad Madero,
Tamaulipas, México

e-mail: G97071322@cdmadero.tecnm.mx

C. Gomez-Santillan

e-mail: claudia.gs@cdmadero.tecnm.mx

N. Rangel-Valdez

e-mail: nelson.rv@cdmadero.tecnm.mx

Ma. L. Morales-Rodriguez

e-mail: lucila.mr@cdmadero.tecnm.mx

G. Castillo-Valdez

Universidad Politécnica de Altamira, Ingeniería en Tecnologías de la Información, Altamira,
Tamaulipas, México

e-mail: georgina.castillo@upalt.edu.mx

clearness, empowerment, and conciseness; also, it guides on how to use the aforementioned visual representations to reveal characteristics of the approximate Pareto sets and the distributions of the solutions.

Keywords Visualization techniques · Multi-objective optimization · Evolutionary algorithms · Pareto front · Good visualization

10.1 Introduction

Due to the nature of some applications, exact methods may not be feasible for some optimization problems, making it computationally expensive and infeasible to generate all optimal solutions [1]. The collection of all vectors of optimal values for the decision variables is known as the Pareto set (PS), and the collection of the corresponding vectors of objective values is known as the Pareto Front (PF). Meta-heuristic algorithms have proved to be effective in finding an approximation to the PS [2]. Another intrinsic difficulty with such problems is visualizing the solutions directly when the objectives that are present in an optimization problem that must be optimized simultaneously are greater than three. Finding techniques and visualizations that make it possible to present large amounts of data is crucial to researchers and participants to observe the shape of the approximate Pareto front, the relationships between objectives, the distribution of solutions, and the convergence and diversity of solutions.

Tušar and Filipič in [3] made a taxonomy of visualization techniques for PF approximations; in this taxonomy, it was found that there are methods with which the original values of the solutions of problems where must be optimized more than three objectives can be shown: they are ScPM [4], PCP [5], HM [6, 7], and RP [8]; Grinstein et al., in [9], classified these methods as high-dimensional data visualization techniques that are capable of presenting a large number of dimensions.

The literature discusses the limitations of visualization techniques, and some works that address this issue are analyzed here. Zhen et al. [10] stated RP can help identify high or low-scoring variables, making it a valuable tool for performance evaluation. However, in some optimization problems, creating hundreds of polygons in a single radar chart can make it cluttered and difficult to read with overlapping polygons. In their study, Zhen et al. proposed a method of dimensionality reduction that transforms a set of solution vectors with many objectives into a set of vectors with a smaller number of objectives. This method retains the original distribution and the Pareto dominance relations that existed among the solutions before the dimensionality reduction. However, Tušar and Filipič [3] pointed out that ScPM leads to a loss of information due to dimensionality reduction. In [11], PCP and HMP were presented as tools to depict the spreading of the solutions, diversity, and the compromises among solutions of multi-dimensional objectives, but interpreting them is often difficult due to solutions' superimposition or arbitrary ordering of the solutions. Gao et al. [12] stated that ScMP may result in cluttered information, which may hinder

decision-makers from exploring the characteristics of the approximation set, such as Pareto dominance relation, PF shape, and knee region.

Based on the analysis, trade-offs exist not only between solutions' objectives but also among different visualization techniques. Therefore, it is worth investigating the specific attributes of each visualization technique.

This work aims to assess whether ScPM, PCP, HM, and RP share the four characteristics of a good visualization known as ACES (Accurate, Clear, Empowering, and Succinct) [13]. According to [13], visualization in an accurate way should represent the trends of the data, the understanding should not be difficult (clear), the visualization, should empower the reader to make decisions (empowerment), and the message should be understood quickly (succinct).

The chapter is structured as follows: The theoretical framework on multi-objective optimization is presented in Sect. 10.2. The proposed methodology is explained in Sect. 10.3. Section 10.4 displays the results of an experimental study on the approximations of solutions to the DTLZ1 problem [14] with three and four objectives. The results were obtained using the NSGA-II [15] and MOEA/D [16] algorithms, which are implementations of the Pymoo optimization framework [17]. The visualization techniques under review include ScPM, PCP, and RP visualization techniques of this same framework, while the Python matplotlib library [18] was used for HM. Lastly, Sect. 10.5 provides the conclusions of the review conducted.

10.2 Theoretical Framework

This section briefly mentions and describes the principal words related to this research; later, the visualization techniques for approximation sets are reviewed.

10.2.1 *High Dimensional-Data Visualizations*

For Grinstein et al. [9], visualization is a visual representation of data; high-dimensional data visualizations can visualize a large amount of data or parameters. Also, in [9], the authors commented that the data is converted into numerical form and then translated into a graphical representation, e.g., the translation of high-dimensional data visualizations from tabular structure to graphical representation using scatter plot matrix [4], parallel coordinate plot [5], and the heatmap visualization [6, 7], among others. Tušar and Filipič [3] included in their taxonomy the high-dimensional data visualizations mentioned and the radar plot [8] as visualization techniques for representing Pareto front approximations that show the original values of the individual solutions. These four visualization techniques are presented in this section.

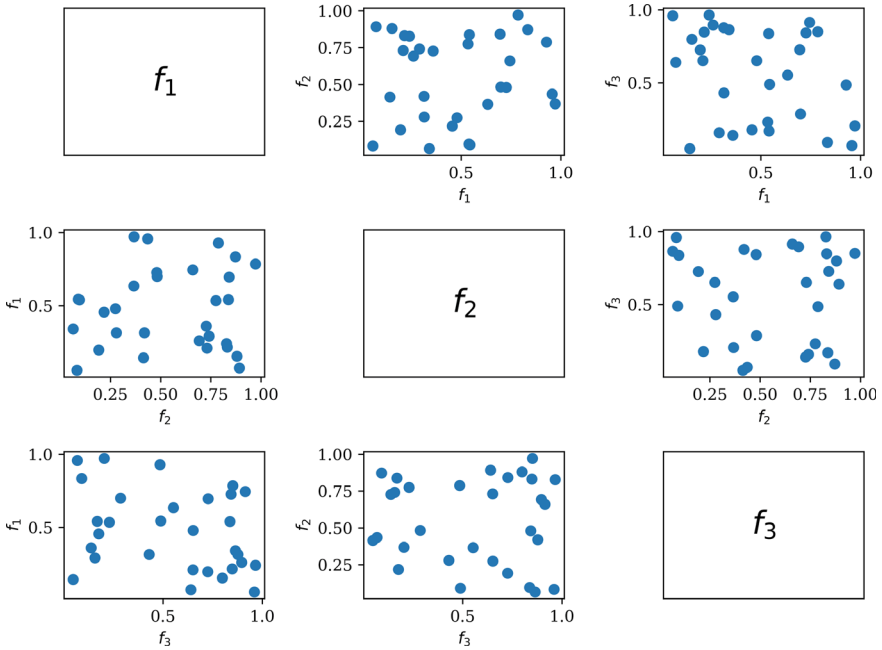


Fig. 10.1 The scatter plot matrix visually represents multivariate data involving p variables (f_1, f_2, f_3). It consists of an ordered arrangement of scatter plots. According to Carr et al. [4], there are a total of $p(p - 1)$ plots (for this example, $p = 3$), resulting in six scatter pots. Each plot represents the relationship between two variables among the set of variables

10.2.1.1 Scatter Plot Matrix

Carr et al., In [4], described the scatter plot matrix as follows: A scatter plot matrix depicting p -variable data consists of an ordered arrangement of $p(p - 1)$ scatter plots. Making aside the layout, the data structure behind the scenes comprises an $N \times p$ matrix, that makes a user able to choose a graphical subset from any plot.

The scatter plot matrix has proven helpful in multi-objective optimization. According to Tušar and Filipič in their publication [3], it makes a projection of points from the set of objectives vectors onto a chosen lower-dimensional space by taking off all other dimensions. This results in a visualization of all possible combinations of the lower-dimensional spaces, enabling comparisons of pairwise solutions for each objective pair. Figure 10.1 depicts this.

10.2.1.2 Parallel Coordinates Plot

Inselberg and Dimsdale In [5] described the Parallel Coordinates Plot as a graph with x and y axes. From the y -axis, N number of lines are evenly distributed and

perpendicular to the x -axis. These lines are labeled x_1, x_2 , and so on up to x_i , acting as axes for the parallel coordinate system. This setup is depicted in Fig. 10.2.

Figure 10.3 shows an example of four solutions for a hypothetical five-objective problem.

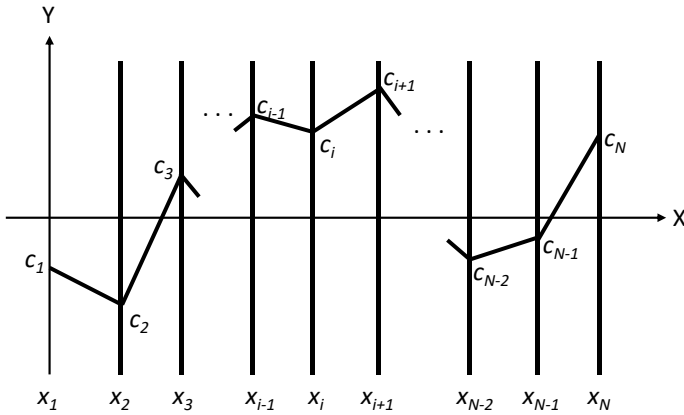


Fig. 10.2 Parallel axes for R^N (Reproduced from [5]). Let's say there is a point C with coordinates (c_1, c_2, \dots, c_N) . In the plot, this point is represented by straight lines. Each line connects to a point on the x -axis, with coordinates $(i - 1, c_i)$, for $i = 1$ to N

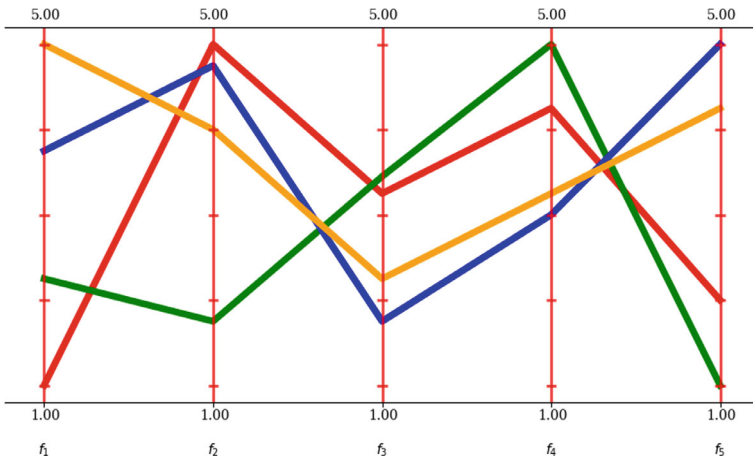


Fig. 10.3 The parallel coordinates plot depicts four five-dimensional points. The red line represents the point at $(1, 5, 3.25, 4.25, 2)$, the green one at $(2.25, 1.75, 3.45, 5, 1)$, the blue one at $(3.75, 4.75, 1.75, 3, 5)$ and the orange one at $(5, 4, 2.25, 3.25, 4.25)$

10.2.1.3 Heatmap Visualization

This data visualization technique involves an array of cells representing values using a gradient of colors [6]. In [7], heatmap visualization was employed to display solutions to multi-objective problems, where for a single solution a row existed, and for a parameter or objective a column existed. Figure 10.4 is a heatmap used to plot a data set from Table 10.1.

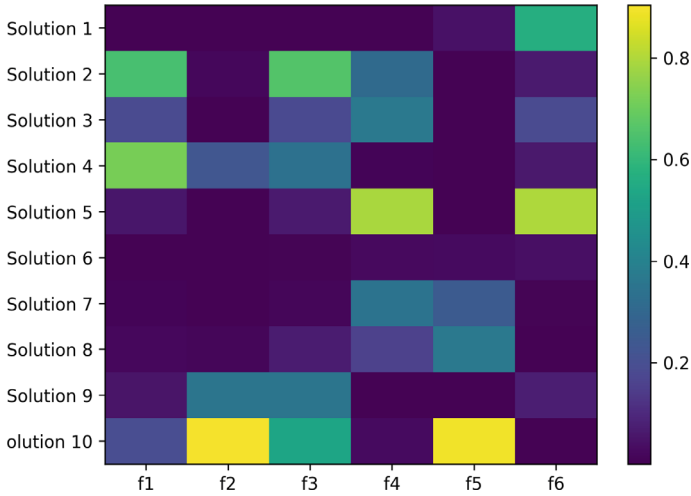


Fig. 10.4 In a heatmap visualization, each row represents a solution, while each column represents an objective. For instance, in this example, there are ten solutions from a subset of an approximate Pareto set obtained from DTLZ1 with six objectives. Each heatmap cell represents the value by color, stronger colors indicating low values and brighter colors indicating high values

Table 10.1 Ten random solutions of an approximation Pareto set from a DTLZ1 problem with normalized data

	f ₁	f ₂	f ₃	f ₄	f ₅	f ₆
Solution 1	0.000676	8.14E-05	0.000677	0.000859	0.044627	0.563722
Solution 2	0.636518	0.018267	0.659257	0.311171	5.75E-05	0.065702
Solution 3	0.18662	0.003464	0.183126	0.367244	0.001718	0.184042
Solution 4	0.718593	0.230891	0.33749	0.009468	4.45E-05	0.062075
Solution 5	0.053267	0.00098	0.063733	0.789741	2.2E-05	0.795575
Solution 6	0.000429	5.56E-05	0.005095	0.027306	0.030298	0.03996
Solution 7	0.007271	0.001122	0.017264	0.343742	0.251595	0.004052
Solution 8	0.02439	0.014888	0.069563	0.155858	0.367061	0.002799
Solution 9	0.051042	0.356647	0.354438	0.002658	2.49E-05	0.07494
Solution 10	0.193781	0.904943	0.532065	0.031384	0.893898	2.33E-05

10.2.1.4 Radar Plot

A Radar plot is a visualization technique representing multidimensional data in a two-dimensional space, as explained in [8]. It consists of multiple axes inclined at equal angles, representing one dimension or variable. A data point is taken at a time, and the value for each variable dimension is marked on these axes, forming a glyph-like structure. Singh and Tewari [9] asserted that radar plots can be constructed in two general forms: polygonic layout and circular layout. In the Polygonic layout, grid lines form concentric polygons depending on the number of represented variables. Each grid line represents a value for all the variables on that grid. Pymoo, the polygonic layout is implemented, and Fig. 10.5 depicts a visual representation of this implementation for five data points with six elements—see Table 10.2.

10.2.2 Multi-objective Optimization

Multi-objective Problems imply optimizing k objective functions $f(x)$ simultaneously, where k is greater than one. These objective functions evaluate a set of decision variables X . These problems are commonly encountered in real-world applications. Due to their importance in various fields, finding effective solutions to these problems is crucial.

For such problems, not only one solution exists but a set of solutions. Strategies such as evolutionary algorithms aim to generate values for decision variables, and

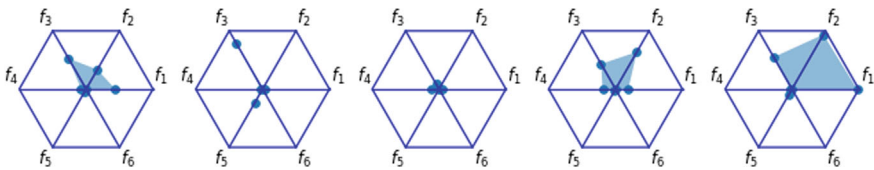


Fig. 10.5 Radar plots. Each solution from Table 10.2, arranged from top to bottom, is represented in a radar plot from left to right

Table 10.2 Five random solutions of an approximation Pareto set from a DTLZ1 problem with normalized data

f_1	f_2	f_3	f_4	f_5	f_6
0.13077135	0.10065709	0.15740833	0.02252036	0.01086121	0.00000007
0.00761107	0.00109214	0.23453671	0.00771430	0.07064618	0.00000014
0.01595633	0.00159859	0.02591002	0.03790499	0.00018110	0.00282709
0.05703629	0.19168182	0.12943249	0.05827036	0.00780017	0.00000570
0.29233723	0.28124773	0.16464740	0.00660085	0.03069174	0.00000132
0.13077135	0.10065709	0.15740833	0.02252036	0.01086121	0.00000007

when they are evaluated by objective functions, the components of the objective vector cannot be improved simultaneously. The set of decision variable value vectors is known as the Pareto optimal set, and the corresponding set of objective value vectors is known as the *Pareto Front*.

This section contains the formal definitions of the concepts mentioned earlier, as defined by Coello in [19].

10.2.2.1 Multi-Objective Optimization Problem

A Multi-Objective Optimization Problem (MOP) is defined by maximizing or minimizing a vector function of the form $F(x) = (f_1(x), \dots, f_k(x))$ subject to $g_i(x) \leq 0$, $i = \{1, \dots, m\}$, and $h_j(x) = 0$, $j = \{1, \dots, p\}$, $x \in \Omega$. The solution of a MOP minimizes or maximizes the components of a vector $F(x)$ where x denotes a vector of n -dimensional decision variables $x = (x_1, \dots, x_n)$ of some Ω universe, where $g_i(x) \leq 0$, and $h_j(x) = 0$ represent constraints imposed on the problem. Ω contains all the values of x that satisfy an evaluation of $F(x)$ in Ω .

10.2.2.2 Pareto Dominance

A vector $u = (u_1, \dots, u_k)$ it is said to dominate another vector $v = (v_1, \dots, v_k)$ (denoted by $u \preceq v$) if and only if u is partially less than v , i.e., $\forall i \in \{1, \dots, k\}, u_i \leq v_i \wedge \exists i \in \{1, \dots, k\} : u_i < v_i$.

The interpretation is that one vector, u , is better than v in at least one component and is not worse than the others.

10.2.2.3 Pareto Optimal Set

For a given MOP, $F(x)$, the Pareto Optimal Set (POS), P^* , is defined as:

$$P^* := \{x \in \Omega \mid \neg \exists x' \in \Omega, F(x') \prec F(x)\}$$

Pareto-optimal solutions are those within the search space (decision space) whose corresponding components of the objective vector cannot all be improved simultaneously [19].

10.2.2.4 Pareto Front

According to Coello in [19], the set of optimal Pareto solutions in objective space that are non-dominated (the components of the objective vector are not simultaneously improved) is called the Pareto front and is defined as:

$$PF^* := \{u = F(x) | x \in P^*\}$$

10.2.3 Evolutionary Algorithms

Zitzler et al., in [1], argued that obtaining the PS for some optimization problem due to the nature of the application can be computationally complex and infeasible, so techniques such as evolutionary algorithms are used to obtain an approximation of the PS. Zitzler et al., in [1], gave the process of an evolutionary algorithm as follows: A set of candidate solutions is maintained, which are evaluated based on their quality and subsequently selected based on their evaluation, once selected, variations are made with recombination and alteration operators, from these variations are chosen based on their quality to form a set called generation. This process is repeated in iterations called generations until a completion criterion is reached, such as a certain number of generations or when the quality of the solutions stagnates. This process is depicted in Fig. 10.6. According to Zitzler et al., evolutionary algorithms follow natural evolution to find optimal or close solutions to a given problem.

It is worth noting that NSGA-II [15] and MOEA/D [16] implement the evolutionary algorithm procedure, and they form part of the area of multi-objective evolutionary algorithms (MOEAs).

10.2.3.1 NSGA-II Algorithm

The NSGA-II algorithm described in [15] starts by randomly creating an initial P_0 population of size N , ordering the population with the criterion of non-dominance.

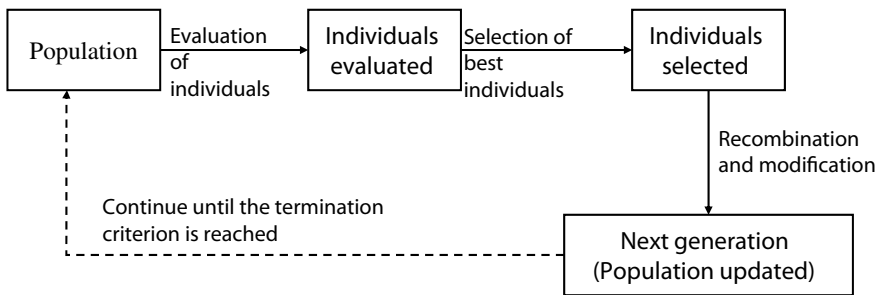


Fig. 10.6 Evolutionary algorithm process. An evolutionary algorithm starts with a *Population* of solutions randomly created. Objective functions evaluate those solutions. From this group of *Individuals evaluated*, a set of best solutions are selected. This set of *individuals selected* undergoes a transformation by recombination and modification, and then the original *Population* is updated with those solutions (this is called the *next generation*). This process continues until a completion criterion is reached

Each solution is assigned a rank based on its level of non-dominance. An initial number of solutions to N -sized Q_0 offspring population is created using evolutionary operators. From these populations, the steps of the n th generation are as follows:

1. The parent population (P_t) and offspring population (Q_t) are combined, leading to a R_t combined population ($R_t = P_t \cup Q_t$). The size of R_t is $2N$.
2. The combined population R_t is sorted according to the criterion of non-dominance, which creates a set of fronts $F = (F_1, F_2, \dots)$. In F_1 , it will find the best solutions from R_t . Suppose the size of F_1 is less than N . In that case, the solutions of F_1 for the new population P_{t+1} will be chosen, and the remaining members of the latest population P_{t+1} will be selected from F_2, F_3 , etc. until N solutions are completed.
3. Suppose the F_T is the last to be included in the new population P_{t+1} . Then to include exactly N members in the population, the solutions of the previous F_T front are sorted using the crowded operator \prec_n .

Figure 10.7 shows the procedure described above.

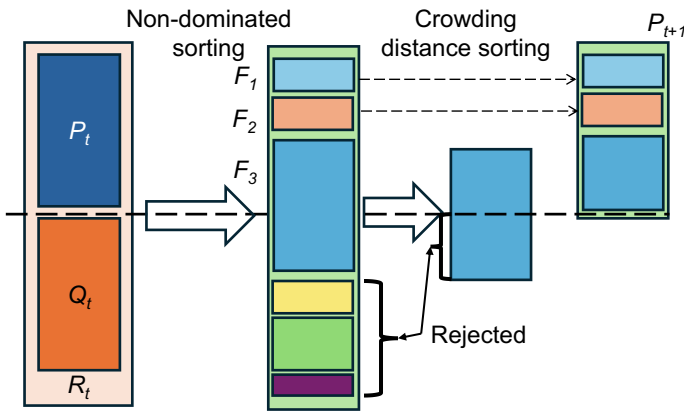


Fig. 10.7 Scheme of NSGA-II procedure. Reproduced from [15]. The NSGA-II algorithm begins by creating an initial population (P_t) with N solutions. An offspring population (Q_t) of the same size is generated from P_t . P_t and Q_t are combined to create a set (R_T) of size $2N$. The population R_T is sorted based on non-dominance criteria, creating a set of fronts (F_1, F_2, \dots), each containing a specific number of solutions from R_T . If the size of F_1 is less than N , the solutions of F_1 for the new population (P_{t+1}) will be chosen, and the remaining fronts (F_2, F_3, \dots) will be used to complete the remaining N solutions. Suppose the F_T front is the last to be included in the new population P_{t+1} . Then to include exactly N members in the population, the solutions of the previous F_T front are sorted using the crowded operator. This procedure is repeated over a number of generations

10.2.3.2 MOEA/D Algorithm

The Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) was introduced in [16], focusing on maintaining a set of scalar optimization subproblems to approximate the optimum of a MOP. The general framework of MOEA/D provided by Zhang & Li in [16], can be summarized as follows:

Let $\lambda^1, \dots, \lambda^N$ be a set of uniformly distributed weight vectors and z^* a reference point. The PF approximation problem of an MOP can be decomposed into N scalar optimization subproblems using the Tchebycheff approach [20]. The objective function of the j -th problem is:

$$\text{minimize } g^{te}(\lambda, z^*) = \{\lambda_i | f_i(x) - z_i^* | \}$$

$$\text{subject to } x \in \Omega$$

where $\lambda^j = (\lambda_1^j, \dots, \lambda_m^j)^T$.

MOEA/D aims to minimize all these objective functions simultaneously in a single run.

$$z^* = (z_1^*, \dots, z_m^*)^T \text{ is the point of reference, this is}$$

$$z_i^* = \max\{f_i(x) | x \in \Omega\} \text{ for each } i = 1, \dots, m$$

In MOEA/D, a vector neighborhood λ^i is defined as a set of several closest vectors belonging to $\{\lambda^1, \dots, \lambda^N\}$. The neighborhood of the i -th subproblem consists of all subproblems with the weight vectors from the neighborhood of λ^i .

In each t generation, MOEA/D with Tchebycheff's approach is maintained:

- A population of N points $x^1, \dots, x^N \in \Omega$ where x^i is the current solution to an i -th problem;
- FV^1, \dots, FV^N , where FV^i is the F -value of x^i , this is $FV^i = F(x^i)$ for each $i = 1, \dots, N$;
- $z = (z_1, \dots, z_m)^T$, where z_i represents the best value for the objective f_i ;
- An external population (EP), which stores solutions non-dominated during the search.

The MOEA/D algorithm outlined by the authors in [16] works as follows:

Input:

- A MOP;
- N : number of sub-problems considered in MOEA/D;
- An N number of uniformly distributed vectors: $\lambda^1, \dots, \lambda^N$;
- T : the number of weight vectors about each weight vector.

Output: *EP*.

Step 1 Initialization:

Step 1.1: Set $EP = 0$.

Step 1.2: Compute the Euclidean distances between any two weight vectors and then work out the closest weight vectors. For each i , set $B(i) = \{i_1, \dots, i_T\}$ where $\lambda^{i_1}, \dots, \lambda^{i_T}$ are the T closest weight vectors to λ^i .

Step 1.3: Generate an Initial Population x^1, \dots, x^N randomly or by a problem-specific method. Set $FV^i = F(x^i)$.

Step 1.4: Initialize $z = (z_1, \dots, z_m)^T$ by a problem-specific method

Step 2 Update: For $i = 1$ to N do

Step 2.1 Reproduction: Randomly select two indices k, l from $B(i)$ and then generate a new solution y from x^k y x^l by using genetic operators

Step 2.2 Improvement: Apply problem-specific repair/improvement heuristic on y to obtain y'

Step 2.3 Update of z: For each j from 1 to m , if $z_j < f_j(y')$, then set $z_j = f_j(y')$.

Step 2.4 Update of Neighboring Solutions: For each index $j \in B(i)$ If $g^{te}(\lambda^j, z) \leq g^{te}(\lambda^j, z)$, set $x^j = y'$, and $FV^j = F(y^i)$.

Step 2.5 Update of EP: Remove from EP all vectors dominated by $F(y')$
Add $F(y')$ to EP if vectors in EP dominated by $F(y')$

Step 3 Stopping criteria: If stopping criteria are satisfied, stop and output EP . Otherwise, go to step 2.

10.2.4 DTLZ1 Test Problem

DTLZ1 problem form part of a test suite problems developed in [14] and is defined by:

M : Number of objectives.

n : Number of variables.

X : Decision variable vector.

$f_i(X)$: i -th objective function.

$g(X_M)$: Functional.

k : number of variables for $g(X_M)$.

The formal definition of the DTLZ1 problem provided by Farina et al. in [14] is as follows:

$$\text{Minimize } f_1(X) = \frac{1}{2}x_1x_2 \dots x_{M-1}(1 + g(X_M)),$$

$$\begin{aligned}
 \text{Minimize } f_2(X) &= \frac{1}{2}x_1x_2 \dots (1 - x_{M-1})(1 + g(X_M)), \\
 \text{Minimize } f_{M-1}(X) &= \frac{1}{2}x_1(1 - x_2)(1 + g(X_M)), \\
 &\vdots \\
 \text{subject to } &0 \leq x_i \leq 1, \text{ for } i = 1, 2, \dots, n.
 \end{aligned}$$

The functional $g(X_M)$ requires $|X_M| = k$ variables and must take any function with $g > 0$ Farina et al., in [11] suggested

$$g(X_M) = 100 \left(|X_M| + \sum_{x_i} (x_i - 0.5)^2 - \cos(20\pi(x_i - 0.5)) \right)$$

For $x_i^* = 0.5 (x_i^* \in X_M)$ conform the Pareto-optimal solution and the objective function values lie on the linear hyperplane define by:

$$\sum_{m=1}^M f_m^* = 0.5$$

Farina et al., suggested $k = 5$, and the number of variables is calculated by:

$$n = M + k - 1$$

Farina et al. stated that the DTLZ1 problem possesses the following properties: (1) The difficulty in converging to the hyperplane, (2) The search space contains $(11^k - 1)$ local Pareto-optimal fronts, (3) his Pareto Front shape is linear.

10.3 Proposed Methodology

The methodology depicted in Fig. 10.8 is an interactive process between the DM and various elements that are incorporated into it. These elements include a MOP, which the DM interacts with by instantiating the problem to be solved. There is also a MOEA, which DM defines and parametrizes to generate a PF. The observable properties of the PF are defined by the DM, which selects an indicator to assess these properties, e.g., the Pareto front shape, the relationships between objectives, and the degree of convergence and diversity. The best PF (F_0^*) is then represented using a set of visualization techniques chosen by the DM. The DM also defines indicators that make observable the ACES features. Finally, the DM visually inspects the graphs and characterizes them, making notations of the strengths and weaknesses of each graph in terms of the expected ACES features.

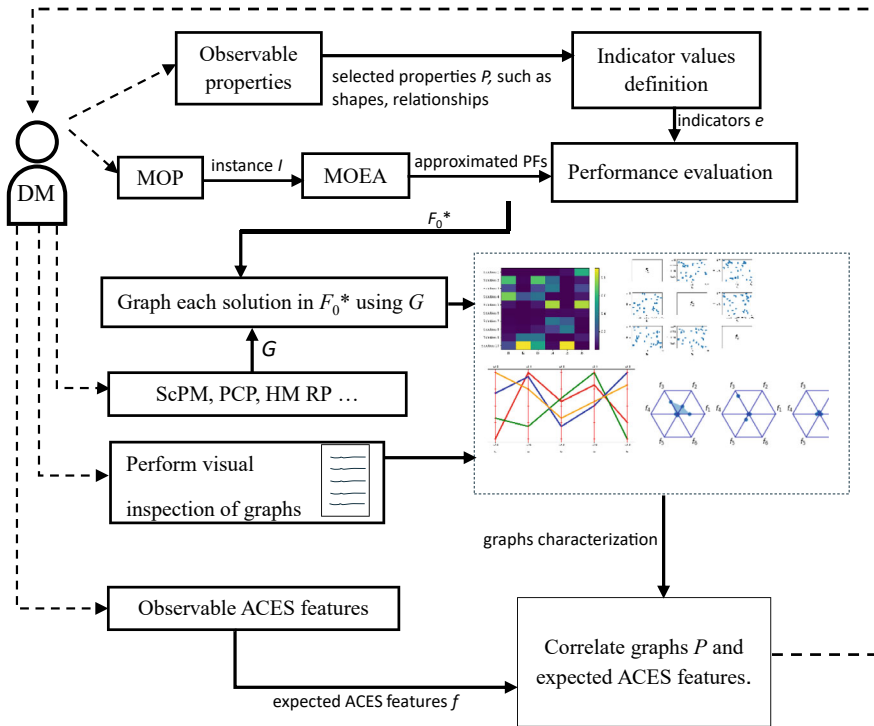


Fig. 10.8 The methodology proposed to assess visualization techniques with ACES

Based on elements present in each visualization technique, an indicator was proposed to assess the accuracy of observing the convergence and diversity of the solutions. Three cases were considered to establish the indicators: in case 1, the real Pareto front is provided; in case 2, two algorithms are compared; and in case 3, an algorithm’s performance is evaluated without a reference point.

The methodology outputs a correlation between a set of expected ACES features, and graph characterization, i.e., a summary of how well the graphs (e.g., ScPM, PCP, HM, RP, etc.) comply with certain desirable characteristics. As a result, DM has enough information to decide whether the data visualization techniques are effective under the predefined context, i.e., the larger the number of ACES features covered, the better the data visualization is. Also, in any case, the methodology allows feedback on data visualization weaknesses, which in turn makes a new iteration to the methodology.

This methodology aims to facilitate the assessment of visualization techniques using the ACES features. By providing a standardized approach, researchers and practitioners can make informed decisions about which techniques to use in their work, ultimately leading to more effective visualizations.

10.4 Experimentation and Results

The experiment will try to answer the question: Is it possible to validate data visualization techniques through the correlation between visual aspects observed in them and the ACES characteristics that define a good visualization technique?

The experiment design consists of the observable properties and the measurement that evaluates these properties, the definition of MOP and the MOEA that obtain approximate Fronts, the evaluation of the performance of approximate Fronts to output the best PF, the establishment of the visualization techniques to represent PF sets and the expected observable ACES features, PF set visualizations, the correlation between observable properties and ACES features, and lastly, an analysis of these correlations.

The experiments were carried out in a Google Colab notebook [21] using the system resources shown in Table 10.3. The programming language was Python 3. Codes are available at <https://colab.research.google.com/drive/1X11mQPM9-jqyOJDt0b2yYQLtoLthyg8o?usp=sharing> for the DTLZ1 with three objectives and for four objectives at <https://colab.research.google.com/drive/1eiMjGEX7LDLkTw1rKL4o1BLnzKYifSS?usp=sharing>.

10.4.1 Observable Properties and Indicator

Observable properties are the concrete aspects desirable for a particular DM to be present in graphs. This work considers the expected properties for the study: the shape of the approximate Pareto front, the relationships between objectives, distribution of solutions, and the degree of convergence and diversity achieved by solutions.

The essence of the previous properties can be captured by means of hypervolume. The hypervolume indicator implemented in Pymoo was used to assess the quality of the approximation sets generated by both algorithms. As explained in [22], the hypervolume indicator offers a single value that can simultaneously assess the convergence and diversity of a solution set. For the analysis, the reference point (1, 1, 1) was selected for three objectives and (1, 1, 1, 1) for the four objectives case.

Table 10.3 Resources of the system where the experiments were carried out

Resource	Valor
CPU	AMD EPYC 7B12
Speed (MHz)	2249.998
RAM	13.61 GB

10.4.2 MOP and MOEA

The MOP selected to solve was DTLZ1; its instances are shown in Table 10.4.

The MOEAs used to obtain approximates of PF were NSGA-II and MOEA/D; the corresponding parametrizations are listed in Tables 10.5 and 10.6.

Table 10.4 Instances of DTLZ1 problem

Test problem	Number of objectives (M)	Number of variables ($n = M + k - 1$) where $k = 5$
DTLZ1	3	7
DTLZ1	4	8

Table 10.5 Parameter values for the NSGA-II algorithm

Parameter	Value	
Population size	105 (adjusted from 300) objectives = 3	120 (adjusted from 300) objectives = 4
Generations	250	
Crossover operator	SBX	
Crossover probability	1.0	
Crossover distribution index	20	
Mutation operator	Polynomial	
Mutation probability	1/n	
Mutation distribution index	20	

Table 10.6 Parameter values for the MOEA/D algorithm

Parameter	Value	
Population size	105 (adjusted from 300) objectives = 3	120 (adjusted from 300) objectives = 4
Generations	250	
Crossover operator	SBX	
Crossover probability	1.0	
Crossover distribution index	20	
Mutation operator	Polynomial	
Mutation distribution index	20	
T	20	
Generation of vectors	Das-Dennis method [23]	
Decomposition approach	Tchebycheff	

Table 10.7 The maximum hypervolume values corresponding to each approximation generated for each algorithm to DTLZ1 with three and four objectives

NSGA-II		MOEA/D	
Max value			
Three objectives	Four objectives	Three objectives	Four objectives
0.9719250300085946	0.9910496983827358	0.973430021035516	0.9939190389042989

10.4.3 Performance Evaluation

A higher hypervolume value indicates a high degree of convergence and diversity of the solutions in PF. Each algorithm was executed thirty times, and each set’s hypervolume was calculated. The approximate PF generated with each algorithm with the highest hypervolume was selected to be represented in the visualization techniques established forward.

Table 10.7 presents the maximum hypervolume values for the thirty approximation sets of DTLZ1, with three and four objectives for each algorithm. These results serve as a basis for analyzing and comparing the two algorithms’ performance.

Table 10.7 shows that the maximum values of hypervolume are obtained from the approximations generated by MOEA/D. This indicates that the solution sets generated by MOEA/D have better convergence and solution diversity than the approximations generated by NSGA-II.

10.4.4 Visualizations Techniques and Observable ACES Features

Visualizations Techniques

The visualization techniques used to represent the set of solutions are scatter plot matrix (ScPM), Parallel Coordinate Plots (PCP), Heat Map (HM), and Radar Plot (RP).

Observable ACES Features

The evaluation of visualization techniques was centered around the observable ACES features, namely the accuracy with which a graph depicts the convergence and diversity of the solutions, the clarity of the relationship between the objectives of the solutions, the extent to which the graph empowers decision-making, and the ability of the graph to succinctly convey the shape of the Pareto front.

10.4.5 Visualizations of the Approximates PF

Scatter Plot Matrix

Figures 9a and 10a represent the scatter plot matrix for DTLZ1 with three objectives and four objectives, respectively, obtained by NSGA-II. On the other hand, Figs. 9b

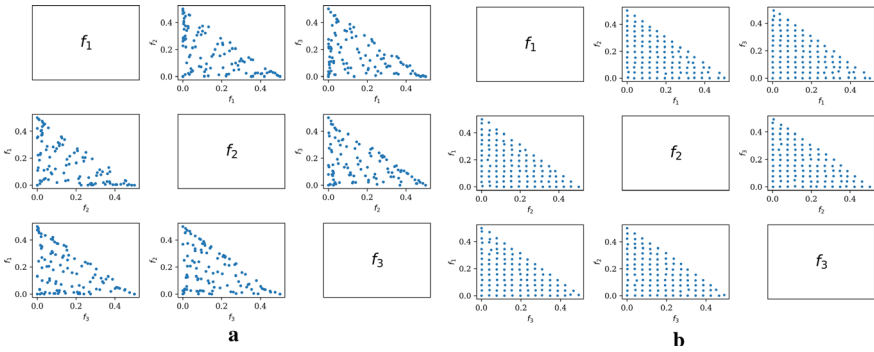


Fig. 10.9 Scatter plots for the best approximations for DTLZ1 with three objectives: **a** NSGA-II, **b** MOEA/D

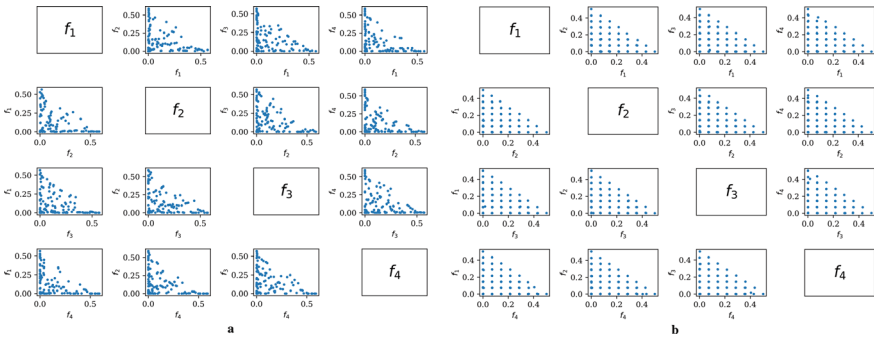


Fig. 10.10 Scatter plots for the best approximations for DTLZ1 with four objectives: **a** NSGA-II, **b** MOEA/D

and 10b depict the scatter plot matrix for DTLZ1 with three and four objectives obtained by MOEA/D.

Figures 9a and 10a show approximations of NSGA-II. The linear shape of DTLZ1 is distinguishable, but with MOEA/D, it is most succinct, as shown in Figs. 9b and 10b. This graph also allows us to observe the relationships between the objectives clearly. The diversity and convergence of solutions are conveyed succinctly. The drawbacks are that it is not possible to select a solution, avoiding decision-making, and it becomes more difficult to manage the scatter plots if the number of objectives increases.

Heatmap

Figures 11a and 12a show the heatmap of the approximations for DTLZ1 with three and four objectives obtained by NSGA-II, respectively; Figs. 11b and 12b show the heatmap of the approximations for DTLZ1 with three and four objectives generated by MOEA/D in each case.

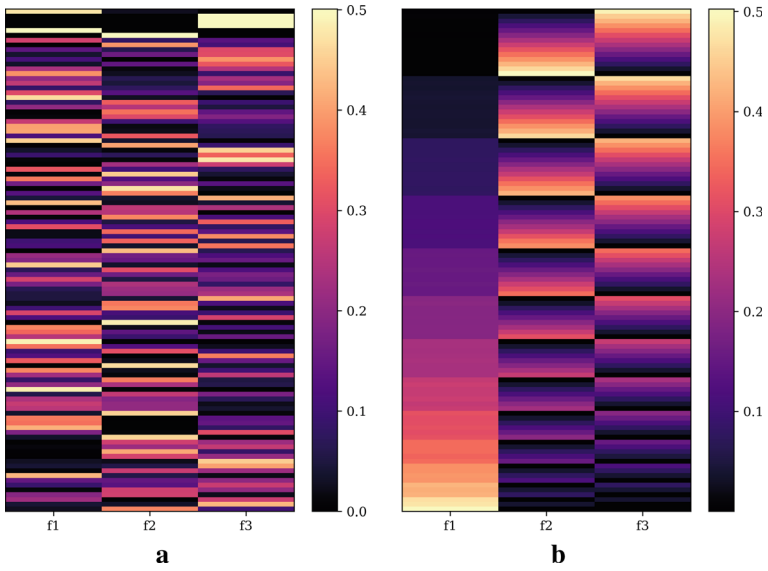


Fig. 10.11 Heatmaps for the best approximations for DTLZ1 with three objectives: **a** NSGA-II, **b** MOEA/D

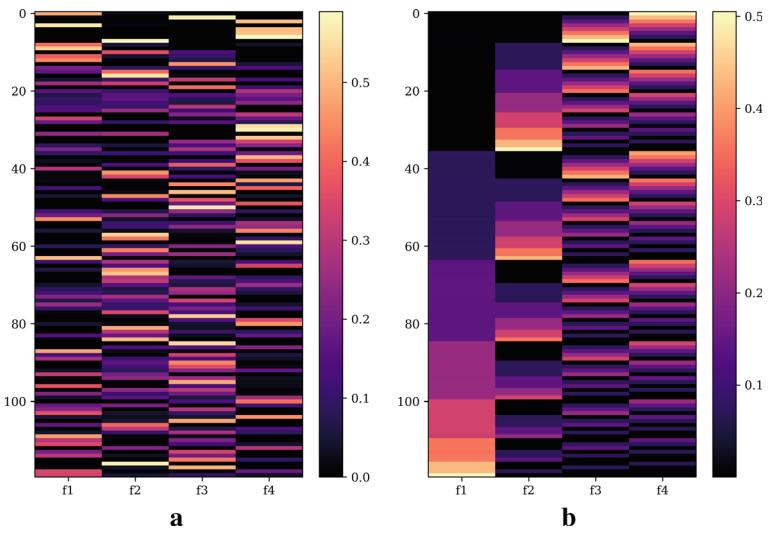


Fig. 10.12 Heatmaps for the best approximations for DTLZ1 with four objectives: **a** NSGA-II, **b** MOEA/D

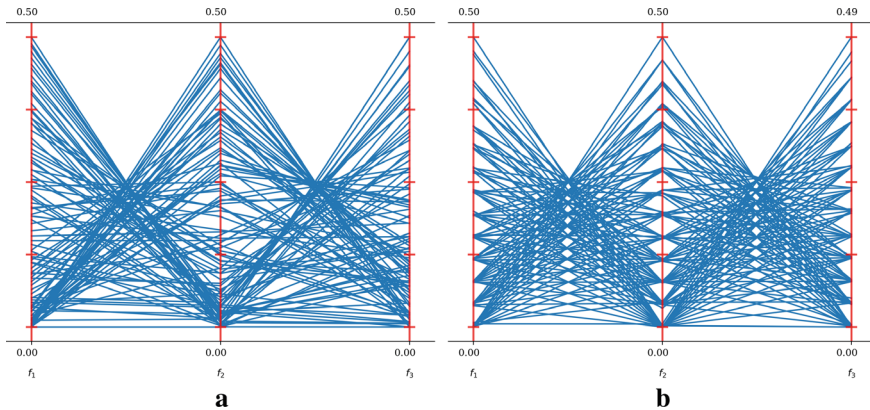


Fig. 10.13 Parallel coordinates plots for the best approximations for DTLZ1 with three objectives: **a** NSGA-II, **b** MOEA/D

Figures 11a and 12a display the solutions for DTLZ1 with three and four objectives respectively, obtained using NSGA-II. These figures provide an accurate visualization of the diversity of solutions. On the other hand, Figs. 11b and 12b, which correspond to solution sets obtained with MOEA/D, demonstrate a high diversity of solutions. Additionally, the differences in colors between the objectives allow for a clear observation of the relationships between them. However, this visualization has a drawback - it does not reveal the shape of the Pareto front, and decision-making is difficult due to the large quantity of solutions.

Parallel Coordinates Plot

Figures 13a and 14a show the parallel coordinates plot of the approximations for DTLZ1 with three and four objectives obtained by NSGA-II respectively, and Figs. 13b and 14b show the parallel coordinates plot of the approximations for DTLZ1 with three and four objectives generated with MOEA/D.

Figures 10.13 and 10.14 clearly illustrate the relationship between the objectives, they accurately display the convergence and diversity of solutions. However, visualization has some limitations, including the inability to observe the shape of the Pareto Front and the difficulty of decision-making due to the superposition of solutions.

Radar Plots

Figures 15a and 16a show the radar plots of the approximations for DTLZ1 with three and four objectives obtained by NSGA-II, respectively, and Figs. 15b and 16b show the radar plots of the approximations for DTLZ1 with three and four objectives obtained with MOEA/D.

From Figs. 10.15 and 10.16, by examining a solution at a given point in time, the relationship between objectives can be clearly observed. However, this visualization has some limitations. For example, Pareto Front's shape cannot be observed, and the large number of solutions can make it challenging decision-making. Additionally, it is not easy to determine the convergence and diversity of the solutions.

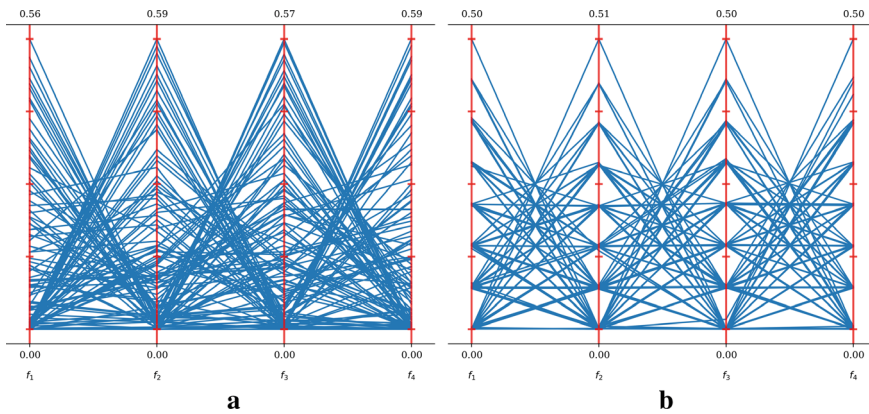


Fig. 10.14 Parallel coordinates plots for the best approximations for DTLZ1 with four objectives: **a** NSGA-II, **b** MOEA/D

10.4.6 Correlations Between Observable Properties and Indicators to Observe ACES Features

ACES stands for *accuracy, clearness, empowerment, succinct* characteristics that according to the state-of-the-art are desirable on visualization techniques. The evaluation of the quality of a particular visualization technique on a given context depends mostly on the chosen indicator and measurements rather than the visualization technique itself.

Let's consider the case of study on this manuscript. The convergence and divergence of metaheuristics must be analyzed. For this purpose, four visualization techniques were carefully chosen. The experiment conducted derived on the generation of an approximation of the Pareto fronts (as described previously), and what remains is a qualitative analysis of the results. A visualization technique will be considered effective if it complies with the ACES characteristics; however, to provide evidence on such a topic, an indicator must be taken into consideration, an indicator that could be measured in the observable context of a visualization technique in the sense of acknowledging how well the studied proper is cover. For this purpose, let's consider the *accuracy feature* and propose indicators for each visualization technique that support the notion of pointing out whether the achieved approximated front has complete convergence and/or divergence.

The problem here lies in identifying a proper indicator derived from observable attributes on a graph. Table 10.8 shows a set of such indicators for the graph and their properties evaluated in the case of study.

Three cases were considered to establish the indicators: in case 1, the real Pareto front is provided; in case 2, two algorithms are compared; and in case 3, an algorithm's performance is evaluated without a reference point.

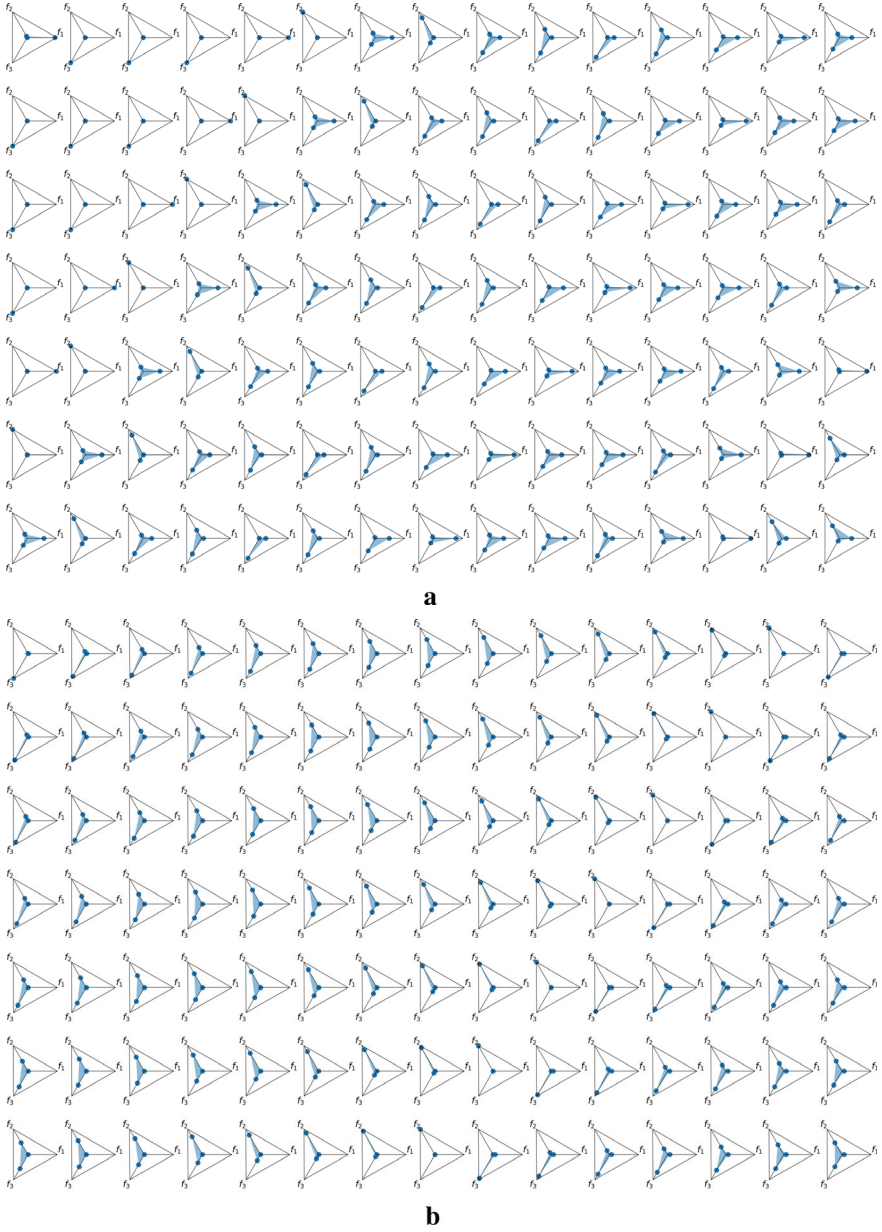


Fig. 10.15 Radar plots for the best approximations for DTLZ1 with three objectives: **a** NSGA-II, **b** MOEA/D

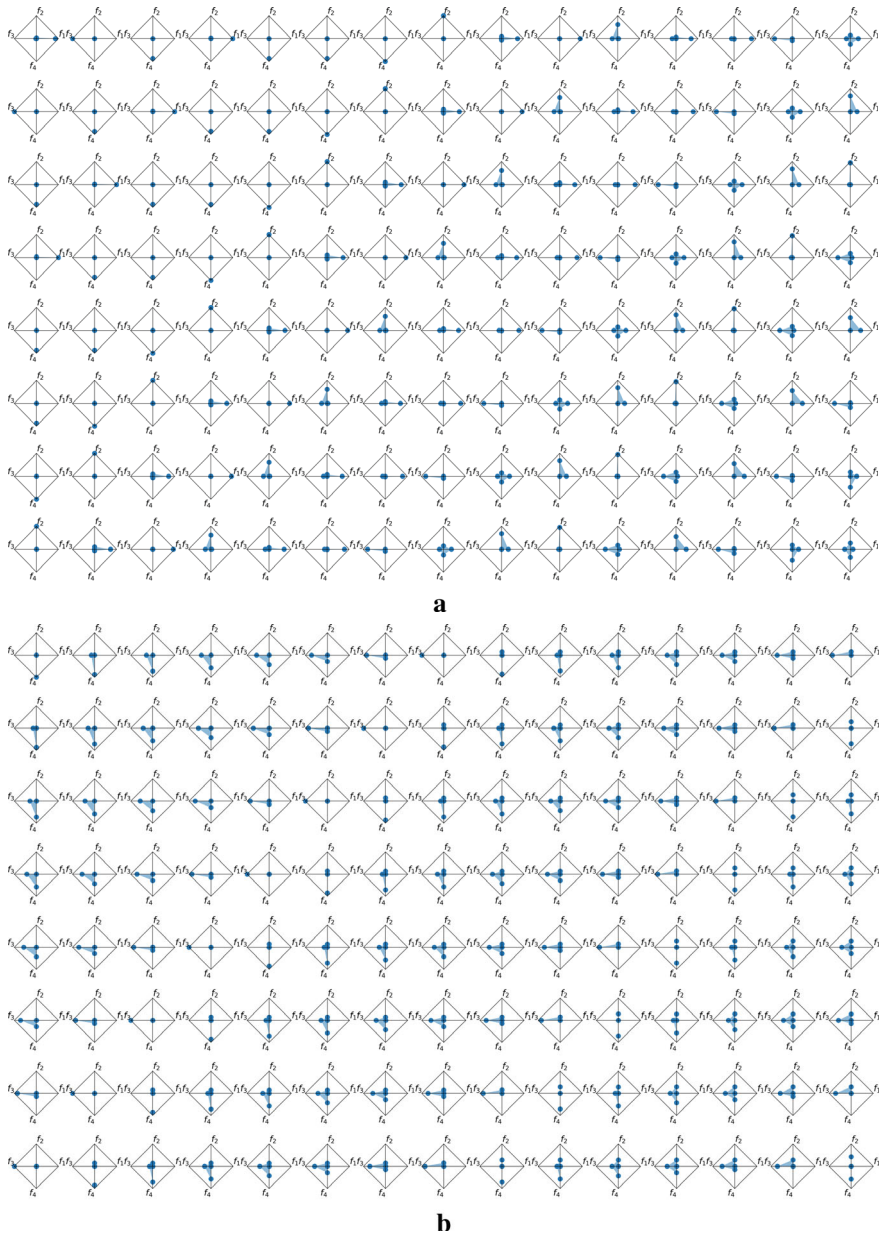


Fig. 10.16 Radar plots for the best approximations for DTLZ1 with four objectives: **a** NSGA-II, **b** MOEA/D

Table 10.8 Summary of the level of accuracy for each visualization technique in all of the three cases: In case 1 the real Pareto front is provided; in case 2 two algorithms are compared, and in case 3 the performance of an algorithm is evaluated and is not provided with a reference point

Visualization Technique	Convergence Case 1	Diversity Case 1	Convergence Case 2	Diversity Case 2	Convergence Case 3	Diversity Case 3
Scatter plot Matrix	High	High	Medium	High	High	High
Heatmap	Medium	Medium	Low	Low	Medium	Medium
Parallel coordinates plot	High	High	High	High	High	High
Radar plot	High	Medium	High	Medium	High	Medium

Scatter plot matrix

The Real Pareto Front is Known

The indicator for accuracy in convergence, where the real Pareto front is known, could be given by the number of graphs associated with pairs of objectives; in which dispersion of the approximate front data points fall within the region of the distribution of the points associated with pairs of objectives of the real front. The accuracy of this technique has a *high* accuracy for observing convergence.

An indicator of accuracy when observing diversity in a scatter plot matrix would be the number of associated plots in which the point dispersion of approximate front objectives pairs coincides with the distribution of the points in the graphs associated with real front objectives pairs. With the above, the scatter plot is *highly* accurate for observing diversity.

Two Algorithms Are Compared

When two algorithms are compared, a region of interest can be established, whose pairwise of objectives scatter plots of the solutions from this region of interest would be a reference point; the indicator would be given by the number of scatter plots in which the points of objectives pairs of the solutions from the approximate fronts approach the points of the solutions of pairs of objectives of the region of interest. This technique has a *medium* accuracy since each of the scatter plots associated with the pairs of objectives of each of the approximate fronts must be observed with the distribution of the points associated with the pairs of objectives of the solutions in the region of interest.

The indicator for diversity would be given by the number of scatter plots in which the distribution of the points of the objectives pairs of the approximate front coincides with the distribution of the points of the pairs of objectives of the reference point. There is a *high* accuracy in observing diversity.

There is no Reference Point, and a Single Algorithm is Analyzed

The convergence of a single algorithm is analyzed, and there is no reference point; convergence can be assessed by looking at changes in the scatter plots of successively

generated fronts. Convergence can be declared by observing stability in the scatter plots associated with the target pairs of the last generated fronts. With the above, there is a *high* precision to observe convergence.

The accuracy for observing diversity is given by the distribution of the points in the scatter plots of the pairs of objectives. There is a *high* accuracy in observing diversity.

Heat Map

The Real Pareto Front is Known

An indicator for accuracy to show convergence on a heat map when the real Pareto front is available could be set by the pattern generated on a heat map of the actual front, if the pattern generated on a heat map of the approximate front matches that of the actual front it can be declared that there was convergence. In this case, the accuracy can be affected by the comparison of colors or differences in the order of the solutions of the two fronts, so the accuracy of observing convergence in a heat map is up for discussion. A *medium* accuracy can be declared to observe convergence in this case.

The indicator to show diversity in a heat map when the real Pareto front is known would be that the distribution of the colors in the approximate front heat map matches the color distribution in the heat map of the actual front. This can be affected if the order of the solutions on the two fronts is different, so the accuracy of a heat map to observe diversity is up for discussion. Therefore, it is possible to declare a *medium* accuracy in observing diversity with this visualization technique.

Two Algorithms Are Compared

An indicator for accuracy of convergence when comparing two algorithms can be defined with a region of interest whose color pattern of the heatmap of the solutions of this region of interest would be the reference point. When representing the solutions of the approximations of the algorithms that are being compared, the convergence would be given by the existence of color patterns like the reference point, however, this becomes complicated by the difficulty of visually comparing colors, a possible way to alleviate this problem would be with ordering of the solutions both in the region of interest and the solutions in the approximate fronts. This visualization technique makes it difficult to accurately observe convergence. So, the accuracy for observing convergence is *low*.

When comparing two algorithms, diversity can be observed by the distribution of colors in the heat maps of each of the approximate fronts. This can be affected by the order of the solutions, so the accuracy of observing diversity in a heat map is debated. So, it can be declared a *low* accuracy for observing diversity on a heat map.

There is no Reference Point, and a Single Algorithm is Analyzed

An indicator of convergence accuracy in a heat map when working with an independent algorithm will be the change of pattern of the heat maps of the successive approximate fronts; the accuracy of the convergence would be given by observing

stability in the color distribution of the last heat maps of the fronts obtained. It can be stated that a *medium* accuracy to observe convergence in this case.

The accuracy of a heat map for observing diversity is given by the distribution of colors in the heat maps of successive approximate fronts: if there is a change in the distribution of colors, diversity can be declared. Accuracy can be affected by the difficulty of comparing colors. However, some techniques can be employed to improve distribution. The accuracy is declared *medium* for this technique for observing diversity.

Parallel Coordinates Plot

The Real Pareto Front is Known

An indicator for accuracy in showing convergence knowing the real Pareto front on a parallel coordinate plot can be given by the number of polylines of the approximated front that fall within the region of the frame formed with the polylines of the real Pareto front. This technique can be declared with *high* accuracy to observe convergence.

An indicator of accuracy in showing diversity by knowing the real Pareto front in a parallel coordinate plot is the visual matching of the frame that forms the polylines of the approximate front with the frame that forms the polylines of the real front. Accuracy is *high* for observing diversity.

Two Algorithms Are Compared

When comparing two algorithms, a region of interest can be established: the algorithms' convergence would be observed if the frame formed with the polylines of solutions from the approximate fronts visually coincides with the frame formed by the polylines of the solutions in the region of interest. The accuracy for observing convergence is *high*.

An indicator of the accuracy of the diversity of the solutions would be given by a uniform distribution of the intersections in the vertical lines of each of the objectives. The accuracy for observing diversity is *high*.

There is no Reference Point, and a Single Algorithm is Analyzed

An indicator of the accuracy of observing convergence in a PCP when there is no reference point and work with an individual algorithm would be the visual change in the plot formed by the polylines of the successive approximate fronts. Convergence would be declared when there is stability in the plot formed by the polylines of the last represented fronts. The accuracy is *high* for observing convergence.

The distribution of the intersections in the vertical lines associated with each of the objectives would give the accuracy to observe diversity; this would lead to declaring a *high* accuracy.

Radar Plot

The Real Pareto Front is Known

The indicator for accuracy in showing convergence knowing the real Pareto front on a radar plot can be given by the number of radar plots of the approximate front that fall within the real Pareto front radar plot. The accuracy is *high* for observing convergence.

An indicator for accuracy when showing diversity knowing the real front on a radar graph is by the visual coincidence of the plot formed by the solutions of the approximate front with the plot formed by the solutions of the real front however if the number of solutions is high, thus the accuracy suffers a detriment due to the superposition of figures, this could be alleviated by plotting each of the solutions on a separate radar plot, which would lead to looking at each of the graphs to declare diversity, due to aforementioned shortcomings a *medium* accuracy for observing diversity is declared.

Two Algorithms Are Compared

When comparing two algorithms, a region of interest can be established. The algorithms' convergence would be observed if the solution frames of the approximate fronts visually matched the radar plot formed by the solutions in the region of interest. The accuracy for observing convergence is *high*.

An indicator of the accuracy of the diversity of the solutions would be given by a uniform distribution of figures of the radar graphs of each of the solutions, this can be affected by the overlapping of the figures, and a way to alleviate it would be to graph each of the solutions individually being able to better observe the diversity, however, if the number of solutions is very large, the accuracy to observe diversity is affected so a *medium* accuracy is declared.

There is no Reference Point, and a Single Algorithm is Analyzed

An indicator of accuracy to observe convergence when you do not have a reference point and work with an individual algorithm would be the visual change in the plot formed by the solutions of the successive approximate fronts. Convergence would be declared when there is stability in the plot formed by the solutions of the last represented fronts. Accuracy is declared as *high* for observing convergence.

The accuracy to observe diversity would be given by a uniform distribution of the figures of the solutions, this can be affected if there are a large number of solutions since there would be overlapping of the figures. The above could be improved if each of the solutions is plotted on an independent radar graph; however, seeing the difference in each of the figures affects the accuracy to observe diversity, hence a *medium* accuracy is stated to observe diversity.

Table 10.8 summarizes the scoring criteria for the *accuracy* property considering elements present on each visualization technique to observe convergence and diversity.

Table 10.9 summarizes the observable ACES features covered by each visualization technique.

Table 10.9 Each visualization technique is marked with a “✓” if the observable ACES feature is covered and with a “✗” if it is not. A mark “?” indicates if the feature is partially covered

Visualization technique	Observable ACES features				
	Accurately depicts the convergence and diversity of the solutions	Accurately depicts the diversity of the solutions	Clarity of the relationship between the objectives	Empowers decision-making	Succinctly convey the shape of the Pareto Front
Scatter plot matrix	?	✓	✓	?	✓
Heatmap	?	?	✓	?	✗
Parallel coordinates plot	✓	✓	✓	?	✗
Radar plot	?	?	✓	?	✗

10.4.7 Analysis

Table 10.8 summarizes the level of accuracy of each visualization to display convergence and/or diversity. It can be concluded that the parallel coordinates plot has high accuracy in displaying both convergence and diversity since their elements allow it. The scatter plot matrix has a medium accuracy related to case 2 because each of the scatter plots associated with the pairs of objectives of each of the approximate fronts must be observed with the distribution of the points associated with the pairs of objectives of the solutions in the region of interest; this method has a high accuracy to observe diversity. The radar plot has medium accuracy in displaying both convergence and diversity due to the overlapping of the figures of each solution. This is highly affected when there are many solutions in the approximated front. However, this can be alleviated by plotting each solution in a single radar plot, it is hard to observe each of the individual’s radar plots. Heatmap is a technique with low accuracy to observe both convergence and diversity due to the difficulty of visually comparing colors. Although this can be alleviated in order to reach a similar pattern, the difficulty of comparing colors remains.

Table 10.9 shows that the Scatter Plot matrix is effective in showing the relationship between objectives and presenting diversity and convergence of solutions and is the only one that can show the Pareto front shape. On the other hand, Heatmap provides a clear visualization of relationships between objectives and solution diversity, but it fails to reveal the shape of the Pareto front. The Parallel Coordinates technique is good at depicting the relationship between objectives and the convergence and diversity of the solutions but lacks visibility in the shape of the Pareto front. Similarly, Radar Plots offer a clear observation relationship between objectives but lack

visibility in the shape of the Pareto front. From this table the positive aspect of visualization to empower decision-making is perceived with limitations in all the techniques; the reason for this situation relies on the way the graphs were constructed. To improve the ability to empower decision-making, the DM should establish an aspect to observe, such as whether the graph is able to convey an algorithm's convergence.

10.5 Conclusions

Is it possible to validate data visualization techniques through the correlation between visual aspects observed in them and the ACES characteristics that define a good visualization technique? Yes, it is possible to validate visualization techniques through such correlation.

It was found that it is possible to define indicators to evaluate a property such as the accuracy of a visualization technique to display a desirable aspect like the convergence and diversity levels of algorithms, considering elements present in them.

The construction of a visualization technique affects how they are interpreted either because of an arbitrary order of the solutions such as in a heatmap or for the superposition of the polylines in a parallel coordinates plot.

Here, the accuracy was mainly evaluated. However, the rest of the features could be evaluated similarly, with visual aspects present in the visualization technique.

Future work to be considered is improving visualization techniques. One idea could be applying a ranking system to the solutions that transform a set of solutions with an arbitrary order into a set of solutions ordered, leading to improvement in, for instance, the display of convergence and diversity of the solutions.

Acknowledgements The authors want to thank Laboratorio Nacional de Tecnologías de la Información and the support of: (a) Cátedras CONAHCYT Program Number 3058, (b) the TecNM project 21336.24-P, (c) the support granted through the Scholarship for Postgraduate Studies with CVU 1260161.

Declaration of Conflicting Interests The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Zitzler, E., Laumanns, M., Bleuler, S.: A tutorial on evolutionary multiobjective optimization. In: *Metaheuristics for Multiobjective Optimisation*, pp. 3–37 (2004). https://doi.org/10.1007/978-3-642-17144-4_1
2. Rivera, G., Porras, R., Sanchez-Solis, J.P., Florencia, R., García, V.: Outranking-based multi-objective PSO for scheduling unrelated parallel machines with a freight industry-oriented application. *Eng. Appl. Artif. Intell.* **108**, 104556 (2022). <https://doi.org/10.1016/j.engappai.2021.104556>

3. Tušar, T., Filipič, B.: Visualization of Pareto front approximations in evolutionary multiobjective optimization: a critical review and the projection method. *IEEE Trans. Evol. Comput.* **19**(2), 225–245 (2014). <https://doi.org/10.1109/TEVC.2014.2313407>
4. Carr, D.B., Littlefield, R.J., Nicholson, W.L., Littlefield, J.S.: Scatterplot matrix techniques for large N. *J. Am. Stat. Assoc.* **82**(398), 424–436 (1987). <https://doi.org/10.1080/01621459.1987.10478445>
5. Li, M., Zhen, L., Yao, X.: How to read many-objective solution sets in parallel coordinates [educational forum]. *IEEE Comput. Intell. Mag.* **12**(4), 88–100 (2017). <https://doi.org/10.1109/MCI.2017.2742869>
6. Metsalu, T., Vilo, J.: ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucl. Acids Res.* **43**(W1), W566–W570 (2015). <https://doi.org/10.1093/nar/gkv468>
7. Pryke, A., Mostaghim, S., Nazemi, A.: Heatmap visualization of population based multi objective algorithms. In *Evolutionary Multi-Criterion Optimization: 4th International Conference, EMO 2007, Matsushima, Japan, 5–8 Mar 2007. Proceedings 4*, pp. 361–375. Springer, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-70928-2_29
8. Singh, V.K., Tewari, V.: Design of improved 3D Radar charts for multidimensional data visualization. *Int. J.* **10**(12) (2022). <https://doi.org/10.30534/ijeter/2022/0110122022>
9. Grinstein, G., Trutschl, M., Cvek, U.: High-dimensional visualizations. In: *Proceedings of the Visual Data Mining Workshop, KDD*, vol. 2, p. 120 (2001)
10. Zhen, L., Li, M., Peng, D., Yao, X.: Objective reduction for visualising many-objective solution sets. *Inf. Sci.* **512**, 278–294 (2020). <https://doi.org/10.1016/j.ins.2019.04.014>
11. Ibrahim, A., Rahnamayan, S.: 3D-RadVis: visualization of pareto front in many-objective optimization. In: *IEEE Congress on Evolutionary Computation (CEC)*, pp. 736–745 (2016). <https://doi.org/10.1109/CEC.2016.7743865>
12. Gao, H., Nie, H., Li, K.: Visualisation of pareto front approximation: a short survey and empirical comparisons. In *2019 IEEE congress on evolutionary computation (CEC)*, pp. 1750–1757. IEEE (2019). <https://doi.org/10.1109/CEC.2019.8790298>
13. Gupta, P., & Bagchi, A.: Data visualization with Python. In: *Essentials of Python for Artificial Intelligence and Machine Learning. Synthesis Lectures on Engineering, Science, and Technology*. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-43725-0_7
14. Farina, M., Deb, K., Amato, P.: Dynamic multiobjective optimization problems: test cases, approximations, and applications. *IEEE Trans. Evol. Comput.* **8**(5), 425–442 (2004). <https://doi.org/10.1109/TEVC.2004.831456>
15. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.A.M.T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**(2), 182–197 (2002). <https://doi.org/10.1109/4235.996017>
16. Zhang, Q., Li, H.: MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Trans. Evol. Comput.* **11**(6), 712–731 (2007). <https://doi.org/10.1109/TEVC.2007.892759>
17. Blank, J., Deb, K.: Pymoo: Multi-objective optimization in Python. *IEEE Access* **8**, 89497–89509 (2020). <https://doi.org/10.1109/ACCESS.2020.2990567>
18. Matplotlib. (n/d.). Annotated Heatmap. Retrieved from https://matplotlib.org/stable/gallery/images_contours_and_fields/image_annotated_heatmap.html
19. Coello Coello, C.A.: *Evolutionary algorithms for solving multi-objective problems*. Springer (2007). <https://doi.org/10.1007/978-0-387-36797-2>
20. Miettinen, K.: *Nonlinear Multiobjective Optimization*, vol. 12. Springer (1999). <https://doi.org/10.1007/978-1-4615-5563-6>
21. Google. (n/d). Google Colaboratory. Retrieved from <https://colab.research.google.com/>
22. Shang, K., Ishibuchi, H., He, L., Pang, L.M.: A survey on the hypervolume indicator in evolutionary multiobjective optimization. *IEEE Trans. Evol. Comput.* **25**(1), 1–20 (2020). <https://doi.org/10.1109/TEVC.2020.3013290>
23. Das, I., Dennis, J.E.: Normal-boundary intersection: a new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM J. Optim.* **8**(3), 631–657 (1998). <https://doi.org/10.1137/S1052623496307510>

Chapter 11

Enhancing Supply Chain Management: A Hybrid Approach for Smart Decisions and Performance



Sandra Rodríguez-Figueroa , Liliana Ramos-Guerrero ,
Efraín Solares-Lachica , and Alberto Aguila-Tovar

Abstract Supply chains (SC) are essential for the operation of companies, evolving with changes in the market and the parties involved. Effective Supply Chain Management (SCM) is key to success and competitive advantage, yet SCM faces challenges such as collaboration, adoption of emerging technologies (including Artificial Intelligence, AI), and best practices. The Supply Chain Operations Reference (SCOR) model is a powerful tool for evaluating and comparing SC activities that aims to reduce the challenges in achieving an effective SCM. This chapter explores the integration of AI in SCM using the SCOR model. A diagnostic methodology based on the SCOR model is designed, and a hybrid model is developed to improve decision making and the performance of SCM. Resulting in a treatment with mathematical tools and optimization methods that seeks a uniform, efficient and effective process. An important feature of the proposed methodology is its applicability to various industries. The methodology is expected to automate decisions, standardize processes, eliminate unnecessary activities, and offer future adaptability. The methodology is reproducible and applicable in virtually any SC, providing intelligent management that optimizes the effectiveness of SCM, positively impacting performance indicators, increasing benefits, and providing flexible responses to changes.

Keywords Business management · Supply chain · SCOR model · Decision making · Evolutionary algorithms

S. Rodríguez-Figueroa (✉) · L. Ramos-Guerrero · E. Solares-Lachica · A. Aguila-Tovar
Faculty of Accounting and Administration, Universidad Autónoma de Coahuila, Torreón,
Coahuila, México
e-mail: rodriguez.sandra@uadec.edu.mx

E. Solares-Lachica
e-mail: efrain.solares@uadec.edu.mx

A. Aguila-Tovar
e-mail: alberto.aguilera@uadec.edu.mx

11.1 Introduction

Organizations have faced significant changes, primarily in the field of Information Technology and Communication, which have needed changes in their relationships with other actors in their Supply Chains (SC), focusing on suppliers and customers [1]. Supply Chain Management (SCM) is a critical concept in the business sphere that synchronizes and coordinates critical logistics processes through information and product flows, facilitating collaboration and integration within the SC [2]. However, the challenges of SCM include developing collaboration and trust among SC partners, identifying best practices to align and integrate the SC, and successfully implementing collaborative information, such as recent Internet systems and technologies that drive efficiency and quality throughout the SC [3]. Modern competitive standards and effective SCM and logistics are indispensable for all businesses, whether small or large, serving domestic or export markets [4]. Integration of the SC is a broad concept that encompasses various aspects like coordination, collaboration, cooperation, interaction, and partnership among SC agents [5].

In today's dynamic environment, companies can no longer afford to compete as individual entities but must compete as complete value chains, relying on an integrated approach [6]. The evolving global market presents new challenges for companies. The lack of professionalization, resistance to innovative practices, and a lack of conviction to adopt new organizational structures and administrative practices hinder profitability and competitiveness, particularly in the marble subsector [7]. To address these challenges, integrating technologies and fostering a culture of innovation and adaptability are necessary [8]. Despite advancements in understanding human decision-making in SCs, there is a need to construct an instrument that generalizes findings from qualitative studies [9]. As CS increasingly integrates different technologies, there is a need for a comprehensive tool to capture demand patterns and customer behaviors more effectively [10].

The Supply Chain Operations Reference (SCOR) model is a product of the Association for Operations Management (APICS), resulting from the merger between the Supply Chain Council and APICS in 2014. Established in 1996, the SCOR model is periodically updated to adapt to changes in business practices within the supply chain SC. This model captures a consensus view of SC management and provides a unique framework linking business processes, metrics, best practices, and technology in a unified structure to support communication among SC partners and improve supply management effectiveness and chain improvement-related activities [11].

The SCOR model exhibits versatility by addressing multi-sectoral SCs and covering all processes and indicators present in each. However, the SCOR model lacks mathematical descriptions and heuristics for precise recommendations from the interpretation of its results. To address these shortcomings, this research aims to integrate decision-making methods and models into the SCOR model to create a more comprehensive and adaptable approach for modern businesses [12].

This research initiative was instigated by a marble trading company's imperative need to enhance delivery processes, mitigate losses attributed to planning

uncertainties, and align with market demands for enhanced quality, reduced lead times, and cost efficiencies. Operating as a private family-owned entity with over a decade of experience in marble commercialization and distribution within and beyond the Comarca Lagunera, Mexico, this company has witnessed substantial growth in product diversification and customer base expansion under its second-generation management.

Subsequently, two additional companies joined this collaborative effort: a metal-mechanic sector enterprise established in March 2014, specializing in machining, metalworking, and laser cutting and engraving processes. Renowned for its diverse clientele including industry giants like Caterpillar, Grupo Modelo, Lala, and Iberdrola, its primary focus lies in manufacturing specialized tools and components for the food, metalworking, and mining sectors. Despite its strategic positioning and advanced equipment utilization, it acknowledges the imperative need for enhancing planning processes to ensure timely supply and response due to its extensive service range.

The third company, operational since 1972 in Mexico, specializes in managing poultry-related inputs for industrial canteens, restaurants, and governmental institutions. Its commitment to personalized customer service is evident in its ability to cater to specific presentation and specification requirements. Partnering with leading suppliers like Tyson, San Antonio, and Pilgrim's Pride ensures product freshness and quality. However, operational challenges, notably delivery delays stemming from product conservation and handling issues, underscore the necessity for supply chain optimization through modern machinery implementation and strategic alignment with the SCOR model.

The overarching goal of this study is to furnish these companies with an essential tool for continuous improvement, specifically addressing delivery time optimization, post-sale services enhancement, and performance indicator refinement. Additionally, the model developed herein holds promise for regional development in Mexico, particularly contributing to the advancement of the Manufacturing and Automotive Industry Cluster of Laguna (CIMAL), a strategically significant sector for the region [13].

The chapter is structured as follows. Section 11.2 details the materials and methods used to address the problem. Section 11.3 provides the steps of the proposed methodology. Section 11.4 describes and discusses the results of the application. Finally, Sect. 11.5 concludes the chapter.

11.2 Materials and Methods

As a tool for SC, the SCOR model proposes improvements in logistical and financial contexts, facilitating an understanding of the business process that identifies consumer satisfaction-driving characteristics. It is divided into four sections: process, practices, people, and performance [14]. Unlike other models, SCOR provides a unified structure to support communication among SC participants, linking business

processes, best practices, metrics, and technology, thereby improving effectiveness and promoting improvement activities in SC management [15]. With an operations-focused approach on product and information flows, SCOR does not consider functions of finance, marketing, or human resources, stemming from a strategic vision of SC [13].

The SCOR Model is used as a diagnostic tool to assess the state of the supply chain and its relevant processes. In this proposal, the outcomes of this diagnostic phase are intended to complement the qualitative data by applying Genetic Algorithms (GAs). In the proposed methodology, explained in Sect. 11.3, a GA is employed to determine the most amenable elements to corrective action, thereby achieving the greatest potential impact. Let us first start by describing the SCOR model.

11.2.1 SCOR Model

The framework of the SCOR model is divided into four levels. Level one identifies key SC processes (plan, source, make, deliver, and return), assisting companies in establishing SC management objectives. Level two explains the main process categories existing in real and created SC within a company (inventory, make-to-order, and make-to-stock products). The third level includes information for supply chain management (SCM), for sourcing planning, and setting strategic SC management objectives (comprising definitions, benchmarking, and software). Finally, the fourth level focuses on implementation, which varies for each company and is not explained in the SCOR [15].

It is worth noting that the SCOR Model addresses only the first three levels, considering them neutral in scope, leaving the definition of the final level beyond its scope. The responsibility for and strategy of the fourth level activities are vested in the companies and sectors involved.

Within the SCOR Model, key elements pertaining to supply chain performance, processes, practices, and human resources are described.

- **Performance:** This element standardizes a set of metrics for evaluating supply chain process performance and establishing strategic objectives. It comprises two types of elements: (1) Performance Attributes and (2) Metrics, as summarized in Table 11.1.

The SCOR Model defines “Processes” as standardized descriptions of the activities that constitute the functioning of supply chains. These processes are structured into several hierarchical levels. At the highest level (Level 1), five processes are defined based on their scope:

- **Plan (Planning):** Involves planning activities necessary for the operation of the supply chain.
- **Source:** Concerns orders from suppliers.

Table 11.1 Level of indicators of the metrics of the five performance attributes

Performance attribute	Level-1 strategic metrics
Reliability	Perfect order fulfillment (RL.1.1)
Responsiveness	Order fulfillment cycle time (RS.1.1)
Agility	Upside supply chain adaptability (AG.1.1) Downside supply chain adaptability (AG.1.2) Overall value at risk (AG.1.3)
Cost	Total supply chain management cost (CO.1.1) Cost of goods sold (COGS) (CO.1.2)
Asset management efficiency	Cash-to-cash cycle time (AM.1.1) Return on supply chain fixed assets (AM.1.2) Return on working capital (AM.1.3)

- **Make (Production):** Encompasses the transformation of raw materials or semi-finished products into finished products. This extends beyond traditional manufacturing to include processes such as repair, recycling, and product reconditioning.
- **Deliver (Distribution):** Encompasses the management, preparation, and delivery of customer orders.
- **Return:** Addresses reverse logistics, encompassing both returns from customers and returns to suppliers.
- **Enable:** Pertains to aspects related to supply chain management, covering information management, risk management, regulatory compliance, and more.

Levels 2 and 3 processes refine the capabilities within Level 1 processes, with Level 3 processes representing specific process steps that, when executed in sequence, plan supply chain activities, source materials, manufacture products, deliver goods and services, and manage product returns.

The “Good Practices” component of the SCOR Model comprises a set of proven, efficiency-enhancing practices that significantly improve the performance of supply chain processes. These encompass a wide array of common supply chain practices, including inventory management, maintenance tasks, order management, reverse logistics, warehousing, application of the Six Sigma methodology, traceability via Radio-Frequency Identification (RFID), and various inventory reduction strategies like Just-in-Time.

Introduced in SCOR 10, the “Working Capital” section of the SCOR Model provides standards for describing the skills necessary for task execution and process management. It defines talent management standards specific to the supply chain. Some of these skills may be applicable beyond the supply chain domain. Skills are defined regarding experiences, level of training, and competence required to efficiently execute each supply chain task and manage associated processes, aligning with the overall metrics and best practices of the model.

11.2.2 Genetic Algorithms

The Genetic Algorithm (GA) employed in our study represents an adaptive heuristic search method grounded in population genetics. It is a probabilistic search algorithm that simulates the mechanics of natural genetic variation and selection. The GA begins with a set of solutions, referred to as a population, with each solution represented as a chromosome. The population size remains constant throughout each generation. In each generation, the fitness of each chromosome is assessed, and the chromosomes for the subsequent generation are probabilistically selected based on their fitness values. Some selected chromosomes then undergo random pairing to generate offspring. During this process, random crossover and mutation events occur. Chromosomes with higher fitness values have a greater likelihood of selection, potentially resulting in improved average fitness values in subsequent generations.

In general terms, GAs stand as a compelling method for solving complex optimization problems through a process inspired by natural evolution. They encode potential solutions as chains of characters known as chromosomes, comprised of genes, each representing variables x_i within a solution set $x = (x_1, x_2, \dots, x_n)$. In contexts like non-linear 0–1 programming, genes are represented by binary codes (0 or 1), a concept easily extendable to portfolio optimization problems where binary values indicate the absence or full support of investment objects, with fractional values indicating partial support.

The success of GAs in navigating towards optimal solutions is contingent upon the implementation of a fitness measure. This measure evaluates the quality of solutions, enabling the algorithm to differentiate between more and less desirable outcomes. The fitness measure plays a pivotal role in guiding the evolutionary process, ensuring that selections made during the algorithm's execution favor superior solutions. The algorithm's effectiveness is also influenced by the size of the population, which the user typically specifies. A smaller population size might lead to premature convergence and suboptimal results due to a lack of diversity. Conversely, overly large populations can result in increased computational demands without proportional gains in solution quality. The evolutionary process in GAs unfolds through several key steps, beginning with the generation of an initial population. This population, typically formed randomly, serves as the starting point for the evolutionary search. To enhance solution diversity and quality, the algorithm employs operations like crossing (or crossover) and mutation. The crossover operation mixes genes from parent chromosomes to produce offspring with potentially superior traits. This operation can be executed in various ways, such as one-point, two-point, or multiple-point crossover, each method differing in how parental genes are combined.

Mutation, on the other hand, introduces random changes to individual solutions, ensuring the algorithm explores a broader section of the search space and avoids premature convergence. This operation typically occurs with a set probability, altering one or more genes to generate a new chromosome. Selection is the process through which the next generation of solutions is chosen, based on their fitness. This

operation biases the selection towards fitter solutions, thereby embodying the “survival of the fittest” principle within the algorithm. Methods like the roulette wheel approach are commonly used, allocating selection chances in proportion to fitness levels. Finally, the newly formed population, created through crossover, mutation, and selection, replaces the original population. This cycle repeats, with the algorithm iteratively refining its population of solutions towards optimality. Various replacement techniques, including elitist, intelligent generation, and steady-state methods, are deployed to maintain or enhance the quality of solutions across generations. GAs have proved to be adequate algorithms to address complex real-world optimization problems (e.g., [16, 17]).

11.3 Methodology

The proposed methodology uses a systematic procedure to develop the hybrid model. It focuses on diagnosing the SC through the SCOR Model, both for individual processes and the entire chain. This involves assessing the existence, functionality, and interrelationships of the SCOR model elements. Subsequently, an optimization is performed through a GA, subject to specific constraints and objectives defined by the participating companies, that allows the methodology to provide recommendations. The algorithm analyzes the relationships and evaluations of the elements provided by the SCOR model and identifies those with the most significant impact on improving SC operations, thereby establishing priorities for improvement. The key steps in this methodology are outlined below.

To create a hybrid model that allows for qualitative diagnosis of the SC, complementing deficiencies with quantitative information through the application of a GA, the following steps detailed as shown in Fig. 11.1 are taken.

We can divide the process into four stages. In the Preparation Stage all the information about the SCOR model and Diagnosis of the SC.

The checklist based on the SCOR model is applied to the process or processes of interest where it is said that attributes (Good Practices, Indicators and Human Capital skills) exist and are evaluated according to their current functioning (0 = does not exist, 1 = exists, but works poorly.... 10 = exists and is in its best possible working order).

Meanwhile, the mapping of the ideal relationship tree according to the SCOR model for processes of interest is elaborated. The preparation of the hierarchical relationship tree involves each of the main processes considered in the SC operation, including Plan, Source, Make, Deliver, Return, and Enable. The resulting relationship tree aims to reflect the ideal structure described by the SCOR model, encompassing all its elements perfectly. This product provides a comprehensive overview of operations and the complete influences of Processes, Subprocesses, Indicators, Good Practices, and Human Capital Skills. The purpose is to facilitate comparison with the individual relationship tree of each participating company based on their self-assessment data obtained through the existence and evaluation checklist of each element.

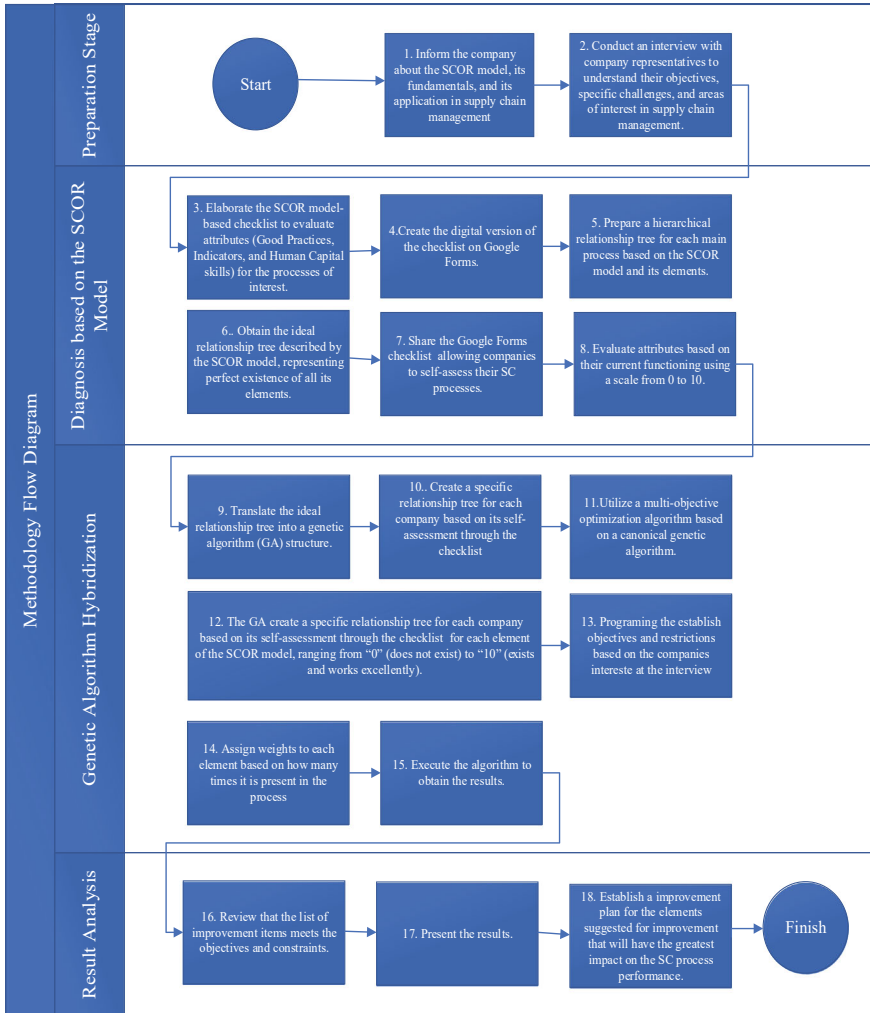
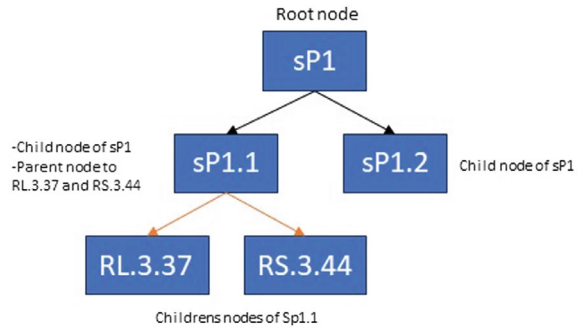


Fig. 11.1 Methodology flow diagram of our hybrid model

Modeling the relationships and their weights involves translating the ideal relationship tree into the GA. This translation is based for example on the premise that node sP1 is the beginning of the process (root node) and serves as the parent of nodes sP1.1, sP1.2, and so forth, encompassing all nodes with a direct relationship beneath it. Conversely, sP1.1 is considered a child node of sP1 and a parent of nodes RL.3.37 and RS.3.44, as illustrated in Fig. 11.2.

For more information on the code for reading the process trees of the companies by the GA, access to the following links is available:

Fig. 11.2 Descriptive tree of the relationships for the Algorithm



```

Relationship Tree Code Example:
sP1[root]={sP1.1,sP1.2}
sP1.1 [sP1]={RL.3.37,RS.3.44}
    
```

- Planning process: https://drive.google.com/file/d/1ZatLL6kWcjqIXLzMqK-ftnq_xohC_o0z/view?usp=drive_link
- Make process: https://drive.google.com/file/d/1AsQTYIiQTP1fibHLwNWqYRVxG1Qb-xaH/view?usp=drive_link
- Delivery process: https://drive.google.com/file/d/1wQVglYOTNgyrqoD4XgJlPyOPvr0tdiZK/view?usp=drive_link

In the third stage, the hybridization of the model with the GA is applied. A specific relationship tree is generated for each company within the SC process. The algorithm constructs this tree based on the self-assessment information obtained through the checklist. For each element of the SCOR model, the company provides a self-evaluation ranging from “0” (non-existent) to “10” (fully operational). This data is recorded in a document titled "Assessment," from which the GA extracts information. The algorithm calculates the weight of each element according to its influence within the SCOR model.

In the SCOR model, Indicators, Good Practices, and Human Capital Skills are repeatedly evaluated across various processes. It is imperative to recognize that certain elements exert a greater impact than others; therefore, any malfunctioning could detrimentally affect multiple elements, ultimately impeding performance. To address this, the algorithm systematically tallies the occurrences of each element within the specific relationship tree, thereby encompassing the entire SC.

The implemented GA is a multi-objective optimization algorithm based on a canonical genetic algorithm. In this algorithm, each chromosome represents an integer value indicating the recommended improvement in a specific activity. Consequently, each individual comprises a set of genes, where the number of genes corresponds to the activities subject to the algorithm’s recommendations, as shown in Table 11.2.

Table 11.2 Individual representing a solution to the problem

v_1	v_2	...	v_n
-------	-------	-----	-------

Table 11.3 $n - 1$ cut-off points

1	2		$n - 1$
v_1	v_2	...	v_n

Where $v_i \in \mathbb{Z}^+$ and $v_i \in [0, \delta_i]$, with δ_i representing the maximum number of units that the algorithm can recommend to the i th activity. There are $n - 1$ crossover points, as depicted in Table 11.3.

Algorithm 1 outlines the procedure, starting with the random creation of an initial population of L individuals, P_0 . Subsequently, for each generation, from the current population P_g the algorithm creates an offspring H_g , also comprising L individuals, using a combination of Selection, Crossover and Mutation operators.

Selection of parents in each generation occurs through binary tournaments, with the winners of two independent tournaments undergoing single-point crossover, resulting in a descendant individual as indicated in Table 11.2.

Mutation involves the random selection and modification of a gene. If gene i is selected for mutation, then the value of the i th gene is randomly generated to satisfy $v_i \in \mathbb{Z}^+$ and $v_i \in [0, \delta_i]$. The probability of selecting an individual for mutation is 1%.

The subsequent step involves combining parents and offspring into a set from which the algorithm selects the individuals with the best fitness, evaluated based on the objectives. The top-fitness individuals within this set from the next parental generation, P_{g+1} . This procedure repeated for G generations. Subsequently, the algorithm returns the individual that represents that represents the best solutions with the highest fitness values in the final population. To mitigate randomness in the procedure, we generate the best individual L times. To leverage these individuals, we employ them as a “seed population” for the algorithm’s last run. The individual generated in this final run is recommended as the best solution to the problem.

The values of the parameters L (population size) and G (number of generations) must be determined by the decision maker based on the specific problem context.

Algorithm 1. Genetic Algorithm proposed to solve the problem

1	$i \leftarrow 1$
2	for $i \leq L$ do
3	$g \leftarrow 0$
4	$P_g \leftarrow \text{create initial population } ()$
5	while $g \leq G$ do
6	$H_g \leftarrow \text{createOffspring}(P_g, \text{selection}, \text{crossover}, \text{mutation})$

(continued)

(continued)

7	$P_{-}(g + 1) \leftarrow \text{better fitness Individuals } (P_{-}g \cup H_{-}g)$
8	$g \leftarrow g + 1$
9	end while
10	$\rho_{-}i \leftarrow \text{findBestIndividual } (P_{-}g)$
11	$i \leftarrow i + 1$
12	end for
13	$g \leftarrow 0$
14	$P_{-}g \leftarrow \{\rho_{-}1, \rho_{-}2, \dots, \rho_{-}L\}$
15	while $g \leq G$ do
16	$H_{-}g \leftarrow \text{createOffspring}(P_{-}g, \text{selection}, \text{crossover}, \text{mutation})$
17	$P_{-}(g + 1) \leftarrow \text{findBestIndividual } (P_{-}g \cup H_{-}g)$
18	$g \leftarrow g + 1$
19	end while
20	$\rho_{-}\text{final} \leftarrow \text{findBestIndividual } (P_{-}G)$

Input: L, population size; G, number of generations.

Output: final, representing the population with the best fitness value.

As part of the third stage, objectives and constraints are established based on the companies' needs.

Multi-objective optimization problems aim to maximize or minimize a set of variables to achieve a solution closest to optimal. During this stage, objectives are determined to guide the GA in its operations. These objectives, set by the research team, are flexible and can be tailored to suit the specific interests of the company implementing the model.

To begin the analysis of the problem, three objectives are set:

1. Maximize the total value of the evaluation process selected. This involves summing all evaluations (both given by the SCOR model and recommended by the GA) from elements in the specified tree.
2. Minimize the total effort, quantified as the number of points that need improvement across all activities; that is, the recommendations of the GA.
3. Maximize the number of activities impacting the selected processes in the ideal SCOR model (opposed to the specific tree that the organizations is working on).

Formally, these objectives are calculated as follows.

Let Φ be the set of activities performed by the organization within the selected process (that is, Indicators, Best Practices, and Human Capital Skills according to the SCOR model); similarly, let Ω be the set of activities within the selected process but according to the ideal framework provided by the SCOR model; let $no(\phi_i)$ (respectively, $no(\omega_i)$) be the number of times that activity $\phi_i \in \Phi$ (resp., $\omega_i \in \Omega$), $i = 1, \dots, n$, appears in such a process; let $eval_{scor}(\phi_i)$ be the score of the i th activity as

evaluated by the SCOR model; and, as stated above, let $\mathbf{x} = \{v_1, \dots, v_n\}$ be a chromosome of the GA, $v_i \in \mathbb{Z}^+$ and $v_i \in [0, \delta_i]$ be the units of score that the algorithm recommend for the i th activity to be improved, and δ_i represent the maximum number of units that the algorithm can recommend for that specific activity. Objective 1 is then formalized as follows:

$$\text{Maximize } f_1(\mathbf{x}) = \sum_{i=1}^n \{no(\phi_i) \cdot (eval_{scor}(\phi_i) + v_i)\}$$

Similarly, Objective 2 is defined as follows:

$$\text{Minimize } f_2(\mathbf{x}) = \sum_{i=1}^n v_i$$

And Objective 3 is:

$$\text{Maximize } f_3(\mathbf{x}) = \sum_{i=1}^n \{no(\omega_i) \cdot v_i\}$$

To achieve the objectives, constraints are incorporated into the GA to optimize its performance and identify the solution closest to the desired optimal outcome. This step can be adjusted based on each company's interests during the application process, providing enhanced flexibility and acceptance in problem description and resolution.

Initially, the following constraints are established for programming and initial model runs:

1. The evaluations of all elements can have a maximum value of 10
2. The number of iterations of the GA is 150,000
3. Of the existing elements in the specific relationship tree, it is desired that in the solutions the performance of all of them is at least a value of 6.
4. Only nodes that no longer have children can be solutions. This follows the logical thinking of the SCOR model that to impact a Level 1 indicator it is necessary to take actions on its direct descendant either Level 2 or Level 3.
5. The number of elements proposed for improvement is 15.

The restriction on the number of activities in which the GA should recommend improvements is not a strict constraint, but a weak one. That is, it will try to comply exactly with that amount, but if the GA considers that it is better to recommend improvements in more or fewer activities, it will do so. Then Constraint violations can be positive, if it recommends making improvements in more activities than those indicated by the user, or negative, if it recommends less.

The fourth stage involves analyzing the results to ensure they meet the objectives and constraints. These results are presented to the companies along with a proposed improvement plan for discussion and implementation.

11.4 Results and Discussion

To demonstrate the application of the proposed methodology, three companies with different processes will be considered, that is, a metalworking company to analyze its planning process, a marble company to analyze the delivery process and a food company to analyze its production process. “sP” is designated as the root node of the metalworking company, “sD” is designated as the root node of the marble company and “sM” is the root node of the food company so that the algorithm recognizes them and the program is executed.

The work equipment to develop the experiments is a LENOVO brand laptop equipped with an AMD Ryzen 7 5700U processor with Radeon graphics at 1.80 GHz, with 16 GB of RAM installed. All this under the Windows 11 Home 64-bit operating system.

For each of the processes, five runs of the GA were carried out, adhering to the parameters of the constraints to select the best solution. The GA provides outputs such as a list of elements to be improved and the recommended extent of recommended improvement. It also yields results for “Cost”, “Assessment”, and “Constraint violations”, as evidenced in Fig. 5.14. “Cost” represents the score units required to invest in each element for improvement. Meanwhile, the “Assessment” outcome evaluates the solution provided by the GA, and the “Constraint violations” result indicates suggestions exceeding (+) or falling short (–) of the constraints set by the GA. The solution showed here is based on the one offering the best Assessment among the five.

The set of solutions for each process resulting from data processing by the Algorithm is shown in Table 11.4, where sP corresponds to the Planning process, sD to the Delivery process, and sM to the Production process.

In the Planning process, the set of solutions provided entails 15 suggestions, whereas the GA deems 16 necessities, resulting in a Constraint violation of 1. The cost of these solutions amounts to 51 with an assessment score of 4971. The process obtained a total evaluation of 1207 points (sum of the evaluations of its elements) out of 1780 possible (1840 of a perfect total) which represents a performance of 67.8% at the time of diagnosis, indicating its functioning prior to improvement efforts. The result of the algorithm projects that with this arrangement the overall evaluation of the process would amount to 1251, which means 72.3% of the process performance, which in percentage measurement is an improvement of 4.5%.

In the case of the Delivery process out of the 15 suggestions, the GA also deems 16 necessities, resulting in a Constraint violation of 1. The cost of implementing these solutions would be 47 with an evaluation score of 3679. The starting pattern of the study begins with a total evaluation of 584 points out of 990 possible with the elements that exist in the marble company representing a performance of 58.9%. This performance is less than ideal in any business environment. The arrangement resulting from the analysis of the Algorithm makes the general evaluation of the process rise to 631, which means 61.86% of the performance of the process, which

Table 11.4 Algorithm results for each process

Arrangement information	List of elements to improve		
Results of the GA for the planning process (sP)			
Cost = -51 Assessment = 49,871 Constrain violation = 1	HS.0120 = 3	HS.0058:1	HS.0070:3
	HS.0139:1	BP.013:7	BP.159:8
	BP.118:1	BP.135:8	BP.035:4
	HS.0037:1	BP.090:3	BP.024:4
	BP.115:1	BP.026:2	
	BP.887:1	BP.145:1	
Results of the GA for the production process (sM)			
Cost = -47 Assessment = 3679 Constrain violation = 1	HS.0092:3	BP.012:1	RL.2.2:4
	RL.3.33:1	CO.3.14.4	RL.2.4:1
	RL.3.35:2	CO.3.15:3	BP.176:1
	RL.3.34:3	RS.3.117:6	HS.0075:5
	RL.3.41:5	HS.0069:1	
	HS.0028:1	AM.3.45:6	
Results of the GA for the delivery process (sD)			
Cost = -44 Assessment = 1678 Constrain violation = 1	HS.0058:3	AG.2.2:7	AG.2.2:7
	CO.3.14:1	CO.3.11:2	CO.3.11:2
	RL.1.1:1	CO.3.12:3	CO.3.12:3
	AM.3.17:2	AM.1.2:1	AM.1.2:1
	RL.3.36:4	RS.2.2:3	RS.2.2:3
	AM.3.19:4	BP.119:3	BP.119:3

in percentage measurement is an improvement of almost 3% that exceeds the previous operation of the process.

Finally, the GA solution to the Production process out of the 15 suggestions targeted for resolution, the GA deems 16 necessities, resulting in a Constraint violation of 1. These solutions come with a cost of 44 and an evaluation score of 1670. The result of the evaluation of the Production process has a score of 531 out of a possible 860 if the existing elements were evaluated with the maximum score, indicating a performance of 61.74%. In this case, the arrangement proposed by the algorithm offers a percentage improvement of more than 5% in the operation of the process, taking the general evaluation to 575, which means 66.86% of the performance of the process.

The group of elements on which it is proposed to take improvement actions for each case would be as follows (See Table 11.5). These proposed elements for improvement include Indicators, Best Practices, and Human Capital Skills.

For each of these Indicators, Human Capital Skills and Good Practices, a set of improvement measures is proposed, which is prepared through a meeting between the work team and the managers of each of the companies in question. The proposals

Table 11.5 Elements proposed to be improved by the Algorithm

Planning process metal-mechanical company	Delivery process marble company	Production process food company
HS.0120 return plan aggregation	RL.3.33 delivery item accuracy	RL.1.1 perfect order fulfillment
HS.0139 supplier relationship management (SRM)	RL.3.35 delivery quantity accuracy	RL.3.36 fill rate
HS.0037 demand management	RL.3.34 delivery location accuracy	RL.3.58 performance
HS.0058 inventory management	RL.3.41 orders delivered damage-free	RS.2.2 production cycle time
HS.0070 logistics network modeling	RL.2.2 on-time delivery customer commit date	RS.3.21 current manufacturing order cycle time
BP.118 transportation management outsourcing	RL.2.4 perfect condition	CO.3.11 direct material cost
BP.115 transportation management system	RS.3.117 route shipment cycle time	CO.3.12 production-related indirect cost
BP.087 ABC inventory classification	CO.3.14 order management costs	CO.3.14 order management costs
BP.013 item rationalization	CO.3.15 order delivery and/or installation costs	AG.2.2 brand upscale adaptability
BP.135 return authorization	AM.3.45 supply inventory days	AG.3.38 current brand volume
BP.090 MRP-based sourcing proposal management days	HS.0028 customer order management	AM.3.19 packaging as % of total material
BP.026 process improvement SOP	HS.0092 pricing management	AM.3.17 supply inventory days—WIP
BP.145 supplier collaboration	HS.0069 logistics management	AM.1.2 supply chain fixed asset profitability
BP.024 supply chain optimization (SCO)	HS.0075 material handling equipment usage	HS.0058 inventory management
BP.159 electronic data interchange (EDI)	BP.176 omnichannel	BP.152 automatic data capture (ADC)
BP.035 business rules review	BP.012 batch tracking	BP.119 dynamic bill of materials generation

are also based on a search for references in the scientific literature that have been successful in similar elements to guarantee an updated and effective strategy in the company that is not only based on the experience accumulated by those involved.

11.5 Conclusions

This research offers a novel approach to improve SC performance by combining qualitative diagnosis with quantitative optimization. By integrating the SCOR model with a GA, the proposed model provides a structured framework for identifying areas of improvement and optimizing SC processes. This proposal contributes to the field of SC management by addressing the need for more precise analysis and decision-making tools. It offers a comprehensive solution for companies looking to enhance their SC performance. This research sets a precedent for future studies in SC management, emphasizing the importance of integrating AI techniques and mathematical optimization into SC diagnosis and improvement processes.

Three cases were considered to demonstrate and assess the capacities of the proposal. The findings show improvements in the performance of key SCM processes (Planning, Delivery, and Production) in around 5%. This demonstrates the efficacy of the hybrid model in optimizing decision-making within the considered processes, paving the way for more efficient and effective supply chain operations.

It is very important to highlight that the parameters used in the methodology have not been optimized and the constraints considered during the search for best paths of recommendation are somewhat demanding for the algorithm. Therefore, further research should focus on refining the algorithmic parameters used in the proposal. Fine-tuning these parameters may lead to even more precise decision-making and performance optimization, ensuring the adaptability of the model across various SCM scenarios. Investigating the model's ability to dynamically adapt to rapid changes in the market and external factors could also enhance its practical utility; implementing real-time data feeds and feedback loops would enable the SCM system to respond promptly to unforeseen challenges and opportunities.

References

1. Addo-Tenkorang, R., Helo, P.T.: Big data applications in operations/supply-chain management: a literature review. *Comput. Ind. Eng.* **101**, 528–543 (2016). <https://doi.org/10.1016/j.cie.2016.09.023>
2. Correa Espinal, A., Alvarez Lopez, C.E., Gómez Montoya, R.A.: Identification systems that use radiofrequency and barcodes and their relation with supply chain management. *Estudios Gerenciales* **26**(116), 115–141 (2010). [https://doi.org/10.1016/S0123-5923\(10\)70126-1](https://doi.org/10.1016/S0123-5923(10)70126-1)
3. Robinson, C.J., Malhotra, M.K.: Defining the concept of supply chain quality management and its relevance to academic and industrial practice. *Int. J. Product. Econ.* **96**(3), 315–337 (2005). <https://doi.org/10.1016/j.ijpe.2004.06.055>
4. Olivos, P.C., Carrasco, F.O., Flores, J.L.M., Moreno, Y.M., Nava, G.L.: Modelo de gestión logística para pequeñas y medianas empresas en México. *Contaduría y Administración* **60**(1), 181–203 (2015). [https://doi.org/10.1016/S0186-1042\(15\)72151-0](https://doi.org/10.1016/S0186-1042(15)72151-0)
5. Martínez Jurado, P.J., Moyano Fuentes, J.: Lean production y gestión de la cadena de suministro en la industria aeronáutica. *Investigaciones Europeas de Dirección y Economía de La Empresa* **17**(1), 137–157 (2011). [https://doi.org/10.1016/S1135-2523\(12\)60048-3](https://doi.org/10.1016/S1135-2523(12)60048-3)

6. Premm, M., Kirn, S.: A multiagent systems perspective on industry 4.0 supply networks. *Lect. Notes Comput. Sci.* **9433**, 101–118 (2015). https://doi.org/10.1007/978-3-319-27343-3_6
7. Aguilar, C.: Caracterización de la cadena productiva del mármol-travertino en el Estado de Puebla, México. In: *Internacional, Red Congreso, Competitividad X I I*, p. 23 (2020)
8. Guzmán, E., Poler, R., Andrés, B.: Un análisis de revisiones de modelos y algoritmos para la optimización de planes de aprovisionamiento, producción y distribución de la cadena de suministro. *Dirección y Organización*, **70**(70), 28–52 (2020). <https://doi.org/10.37610/DYO.VOI70.567>
9. López Vera, J., González Soriano, F.: De la SCM tradicional a SCM social, inteligente y verde: Una revisión de la literatura (2017)
10. Agrawal, P., Narain, R.: Digital supply chain management: an overview. *IOP Conf. Ser. Mater. Sci. Eng.* **455**, 12074 (2018). <https://doi.org/10.1088/1757-899X/455/1/012074>
11. APICS Supply Chain Council: Supply Chain Operations Reference Model. SCOR Version 12.0. In *APICS* (Vol. 12, Issue 2) (2017)
12. Akkawuttiwanich, P., Yenradee, P.: Fuzzy QFD approach for managing SCOR performance indicators. *Comput. Ind. Eng.* **122**, 189–201 (2018). <https://doi.org/10.1016/j.cie.2018.05.044>
13. Calderón-Lama, J.-L., Lario-Esteban, F.-C.: Análisis del modelo SCOR para la Gestión de la Cadena de Suministro. *IX Congreso de Ingeniería de Organización* **4**, 41–50 (2005)
14. Jassir, E., Domínguez, M., Paternina, C., Henríquez, G.: Impacto de los indicadores del modelo scor para el mejoramiento de la cadena de suministro de una siderúrgica, basados en el ciclo cash to cash. *Innovar* **70**(28), 147–161 (2018). <https://doi.org/10.15446/innovar.v28n70.74454>
15. Icarte Ahumada, G.A.: Aplicaciones de inteligencia artificial en procesos de cadenas de suministros: una revisión sistemática. *Ingeniare Revista Chilena de Ingeniería* **24**(4), 663–679 (2016). <https://doi.org/10.4067/s0718-33052016000400011>
16. Rivera, G., Cisneros, L., Sánchez-Solís, P., Rangel-Valdez, N., Rodas-Osollo, J.: Genetic algorithm for scheduling optimization considering hetero-geneous containers: a real-world case study. *Axioms* **9**(1), 27 (2020). <https://doi.org/10.3390/axioms9010027>
17. Fernandez, E., Rangel-Valdez, N., Cruz-Reyes, L., Gomez-Santillan, C., Rivera-Zarate, G., Sanchez-Solis, P.: Inferring parameters of a relational system of preferences from assignment examples using an evolutionary algorithm. *Technol. Econ. Develop. Econ.* **25**(4), 693–715 (2019). <https://doi.org/10.3846/tede.2019.9475>

Chapter 12

Attribute Weighting Model for Breast Cancer Prediction with the Harmony Search Algorithm



Clara Antonio-Hernández, Jesús D. Terán-Villanueva, José A. Castán-Rocha, Mirna P. Ponce-Flores, and Zurisadai Ponce-Flores

Abstract Breast cancer is a disease that affects many women worldwide. Identifying risk factors is important for prevention and early treatment. Although models such as Gail, Tyrer-Cuzick, and BOADICEA can predict breast cancer risk at five to ten years based on risk factors, the Gail model has been shown to have poor accuracy. Moreover, accurately assessing the influence of risk factors remains a challenge. Hence, accurate models are needed for early detection. In this paper, we used the harmony search algorithm to assign weights to each risk factor value to produce an accurate predictive model. Our model achieved a 96% precision and an 81% accuracy, outperforming Gail's results, which obtained a 67% precision and a 60% accuracy. Furthermore, the simplicity of our model makes it a valuable tool for both patients and medical professionals.

Keywords Breast cancer · Risk factors · Attribute weighting · Harmony search algorithm

12.1 Introduction

Breast cancer is a highly significant disease, and early detection is crucial for effective treatment. Currently, there are awareness campaigns about the risk of developing breast cancer and the importance of leading a healthy lifestyle, as well as the timely

C. Antonio-Hernández · J. D. Terán-Villanueva · J. A. Castán-Rocha (✉) · M. P. Ponce-Flores · Z. Ponce-Flores
Universidad Autónoma de Tamaulipas (UAT), Centro Universitario Tampico Madero, P.C. 89109
Tampico, Tamaulipas, Mexico
e-mail: jacastan@docentes.uat.edu.mx

C. Antonio-Hernández
e-mail: clarahdz92@gmail.com

J. D. Terán-Villanueva
e-mail: jdteran@docentes.uat.edu.mx

detection of the disease. Despite efforts to reduce breast cancer, the number of women dying from this cause in Mexico is increasing. According to the National Institute of Geography, the institute registered 7888, 7973, and 7880 in 2022, 2021, and 2020, respectively [21–23]. According to the World Health Organization, globally, 2.3 million women were diagnosed with breast cancer, resulting in 685,000 deaths, making it the leading cause of death among women [37].

Screening tests, which are essential to identify cancer, are divided into two categories: invasive examinations (such as tissue removal or blood studies) and noninvasive methods, such as self-examination (e.g., palpation of an area of the body or assessment of risk factors) [18]. An advantage of noninvasive testing is that it eliminates the need to remove tissue from the skin for subsequent analysis, which reduces the risks of infections and minimizes patient discomfort. Additionally, these tests are less painful and more cost-effective compared to invasive procedures [30]. As noninvasive methods, probabilistic models like the Gail [12, 28], BOADICEA [32], and Tyrer-Cuzick [43] models play a crucial role in predicting breast cancer within five to ten years. Many countries, especially the United States, use such models regularly, which combine biological and reproductive history factors. Their particular difference is that BOADICEA and Tyrer-Cuzick use laboratory test results.

Mammography is currently the most effective tool available to physicians for detecting cancer in healthy women, as it has demonstrated a significant reduction in fatalities from this disease in many cases [26]. However, undergoing repeated mammograms can increase the risk due to continuous radiation exposure. If an initial mammogram shows no issues, follow-up mammograms can further raise this risk. Additionally, several factors can make it difficult to accurately detect tumors on mammograms, often resulting in *false positives*, a concept used by medical professionals [11, 16].

Meta-heuristic algorithms are approximate methods designed to address complex combinatorial optimization problems where conventional heuristics demonstrate limited effectiveness and efficiency [36, 42]. In the scientific literature, various proposals for meta-heuristic algorithms can be identified. Meta-heuristic algorithms are constantly being used and extended to solve currently challenging problems (e.g., [8, 38, 44]). Among the meta-heuristic algorithms, we propose using the Harmony Search Algorithm (HSA) [1, 48] to weigh the elements of each breast cancer risk factor. The HSA is adaptable to various optimization problems without intensive parameter tuning [45]. It has proven effective in complex, nonlinear problems where the structure of the searching space can be difficult to understand. Additionally, it avoids complex calculations, making the processing significantly fast, and time becomes a distinguishing factor compared to other meta-heuristic algorithms. The HSA bases its functionality on the musical improvisation process [7]. The importance of improvising harmony with parameters such as random selection, Harmony Memory Consideration Rate (HMCR), and Pitch Adjustment Rate (PAR) allows us to effectively explore the search space [14].

Another crucial aspect of HSA is the objective value or function. The objective value also plays a crucial role in guiding the algorithm towards efficient solutions and is a vital component in any optimization algorithm, including HSA. This function

takes a candidate solution (harmony) as input and computes a numerical value that reflects the solution's quality, referred to as fitness or the objective function value. Therefore, we proposed to use correlation as the objective value in HSA. Correlation is a widely used statistical measure that indicates the relationship or dependence between two variables [41, 47]. Correlation is used in various studies and disciplines for its ability to measure and quantify the relationship between two variables [19, 24]. In this paper, we used it as the objective function in HSA, assessing the quality of a solution based on correlation.

In the literature, we find papers addressing issues related to attribute weighting, breast cancer, and the Gail model. Doppala et al. [10], in 2023, proposed a hybrid feature selection methodology using a radial basis function genetic algorithm (GA-RBF) to detect coronary diseases. They compared it with six models: Naive Bayes, Decision Tree, Logistic Regression, Support Vector Machine, Random Forest, and KNN. Their results showed an accuracy of 85.40% using 14 attributes. The accuracy increased to 94.20% using 9 features after feature selection, demonstrating superior performance compared to other models.

Khozama et al. [27] in 2021 proposed a machine learning tool based on decision trees for early breast cancer. They used 280,660 records from Breast Cancer Surveillance Consortium (BCSC), a medical questionnaire, and international reports to identify crucial risk factors. They normalized and balanced the dataset and assigned weights to risk factors based on the Degree Of Importance (DOI). The results indicate that weighting significantly enhanced the model performance, achieving 95% accuracy on different balanced datasets. Finally, the authors suggested that this weighting strategy can be instrumental in improving breast cancer risk estimation.

Saleh et al. [39] in 2021 proposed the Gail model to estimate 5-year breast cancer risk and evaluate new risk factors among Egyptian women. They used 7009 records from women in urban and rural areas of Egypt. They performed Chi-square tests to determine the association among characteristics and a binary logistic regression algorithm to examine correlations. The results revealed that 8.75% of the sample had an elevated risk of developing cancer. The authors suggest that the Gail model is useful in the Egyptian population, although it cannot accurately predict whether a woman will develop breast cancer. However, they believe it is valuable in selecting appropriate prevention strategies based on each individual's level of risk.

Nouira et al. [34] in 2020 examined various feature selection methods using the BCSC dataset. They utilized Chi-square, T-score, F-score, and Gini index as attribute selection techniques along with two classification algorithms: Random Forest and ANN. The authors emphasize that employing diverse feature selection techniques enhances data understanding, refines classifier models, and aids in pinpointing irrelevant variables. Additionally, they note that selecting the best feature subset improves performance in terms of speed, cost, and domain insights. In this paper, feature reduction in this study bolstered performance, increasing accuracy from 99.88% to 99.90% for ANN and from 99.94% to 99.95% for Random Forest. Finally, they concluded that feature reduction did not have a negative impact but rather improved performance.

Ali and Ahmed [3] in 2019 proposed a hybrid intelligent prediction methodology for phishing websites using Deep Neural Networks (DNN) and Genetic Algorithm (GA) as a feature weighting and selection method. They used the phishing website dataset from UCI. They performed experiments with and without feature selection and weighting. They compared their results without applying DNN feature weighting and selection with five machine learning techniques: Backpropagation Neural Network, Support Vector Machine, K-Nearest Neighbor, C4.5, and Naïve Bayesian, achieving better accuracy with 88.77%. By implementing feature selection and weighting the authors achieved 90.39% and 91.30% accuracy, respectively. Finally, they demonstrated that DNN with GA-based feature selection and weighting contributed to enhancing classification performance using fewer features.

In this paper, we propose a new model using a Harmony Search algorithm to improve the precision in classifying breast cancer using risk factors. This study presents a simple and practical approach that differs from methods such as neural networks or random forests. Our proposal aims to be accessible to both medical professionals and patients; the patients are our priority, and we aim to provide a simple-to-use method useful to people who have not taken any laboratory test. It is important to note that the classification is according to the values proposed by the HSA weighting the elements of each attribute. The rest of the paper is organized as follows: Sect. 12.2 describes the dataset used, providing details on data preparation for this study. Section 12.3 describes the theoretical foundations of the Gail model and Pearson correlation. Section 12.4 presents a detailed description of the proposed methodology, including algorithm steps and features. Section 12.5 showcases the results obtained by the model. Section 12.6 provides a discussion of the results. Finally, Sect. 12.7 presents the conclusions and future work, summarizing the paper's contributions and suggesting potential future research.

12.2 Dataset Description and Preprocessing

In this research, we are using a publicly available dataset from the Breast Cancer Surveillance Consortium (BCSC) <http://bcsc-research.org/> from the United States that gives detailed information on breast cancer-related risk factors from the years 2000 to 2009. The dataset comprises 1,144,564 records, including 11 risk factors and the labeled class *Diagnosis*, which indicates the presence (value of 1) or absence (value of 0) of breast cancer, as detailed in Table 12.1.

12.2.1 Data Preparation

Data preprocessing is an important step to ensure the quality and suitability of the information collected for our analytical purposes. This process involves cleaning, transforming, and reducing data in order to eliminate any incomplete, redundant, or incoherent information [17]. This process is detailed below.

Table 12.1 Breast cancer risk factors description

No.	Risk factor	Description (values)
1	Year	2000–2009
2	Age group 5 years	Group1 = 18–29; Group2 = 30–34; Group3 = 35–39; Group4 = 40–44; Group5 = 45–49; Group6 = 50–54; Group7 = 55–59; Group8 = 60–64; Group9 = 65–69; Group10 = 70–74; Group11 = 75–79; Group12 = 80–84; Group13 = more 85
3	Race eth	1 = Non-hispanic white; 2 = Non-hispanic black; 3 = Asian/Pacific Islander; 4 = Native American; 5 = Hispanic; 6 = Other/Mixed; 9 = Unknown
4	Family history (First degree)	0 = No; 1 = Yes; 9 = Unknown
5	Age menarche	0 = Age more equal 14; 1 = Age 12–13; 2 = Age under 12; 9 = Unknown
6	Age first birth	0 = Age under 20; 1 = Age 20–24; 2 = Age 25–29; 3 = Age more 30; 4 = Nulliparous; 9 = Unknown
7	BIRADS breast density	1 = Almost entirely fat; 2 = Scattered fibroglandular densities; 3 = Heterogeneously dense; 4 = Extremely dense; 9 = Unknown or different measurement system
8	Hormone treatment (Current)	0 = No; 1 = Yes; 9 = Unknown
9	Menopaus	1 = Pre or peri menopausal; 2 = Post menopausal; 3 = Surgical menopause; 9 = Unknown
10	BMI group	1 = 10–24.99; 2 = 25–29.99; 3 = 30–34.99; 4 = 35 or more; 9 = Unknown
11	Biopsy	0 = No; 1 = Yes; 9 = Unknown
12	Diagnosis	0 = No; 1 = Yes; 9 = Unknown

12.2.2 Data Cleaning and Ordinal Encoding

Data cleaning consisted of removing all records with missing values. Specifically, it was decided not to perform data imputation to preserve the original dataset integrity and the method behavior. Due to the crucial nature of the information related to a deadly disease, we decided to preserve the integrity of the original data, avoiding any artificial modification or imputation. Fortunately, having sufficient test cases allowed us to eliminate the records. This decision is crucial, as correct classification plays a vital role, with consequences from life to death. Thus, we prioritize data integrity and reliability. We removed all missing data and some attributes that were not necessary for our analysis, leaving us with 153,821 cases out of 1,144,564.

For categorical variables, we used ordinal encoding, which assigns numerical values to the categories to improve their applicability in the analysis and ensures a consistent scale across all variables to avoid influencing the results. We assigned numerical values to use as identifiers within the algorithm. It is important to note that these numerical values are not used directly in the classification process. This procedure is crucial for averting potential biases in the analysis arising from variations among variables, ensuring a fair and accurate comparison [6]. It is worth noting that the data set is unbalanced, with an approximate 14/86% of positive and negative cases, respectively. We artificially increased the number of positive cases by replicating them six times, resulting in 132,042 positive and 131,814 negative records, giving a total of 263,856 records for our analysis. Our main objective is to focus on the positive cases that pose a health threat. Thus, we proposed a simple sum that is accessible to anyone without the need for specialized tests or the use of advanced artificial intelligence techniques, such as decision trees or other more complex methods, focusing on the probability of positive cases rather than the overall accuracy.

12.3 Theoretical Foundations

12.3.1 Gail Model Evaluation

The Gail model is widely employed as a statistical tool for predicting the five-year risk of breast cancer. It relies on specific risk factors and provides valuable clinical and research information for guiding medical decisions and identifying women at risk of developing breast cancer. The model considers the patient’s age, family history, number of biopsies, atypical hyperplasia, race, menarche age, and pregnancy age, as shown in Fig. 12.1.

<i>Gail Model</i>				<i>Prediction</i>
<i>Age/ Race</i>	<i>Age menarche</i>	<i>...</i>	<i>Age pregnancy/Family History</i>	<i>Gail score</i>
0.050	1.10	...	2.83	1.22
0.134	1.21	...	2.68	2.08
1.006	1.10	...	5.78	0.67
...
...
0.049	1.00	...	1.55	1.88

Increased risk of breast cancer > 1.66

Precision of 67%

Fig. 12.1 Gail model

Table 12.2 Description of risk factors and weighting of the Gail model

Risk factors considered by the Gail model and their weighting					
#	Age/Race non-hispanic black	Weight	#	Num biops/Age orientation > 50	Weight
1	18–29	0.050	28	No biopsy	1.00
2	30–34	0.120	29	1 biopsy	1.70
3	35–39	0.224	30	≥ 2 biopsies	2.88
4	40–44	0.310		Num biops/Age orientation ≥ 50	Weight
5	45–49	0.355	31	No biopsy	1.00
6	50–54	0.416	32	1 biopsy	1.27
7	55–59	0.511	33	≥ 2 biopsies	1.62
8	60–64	0.562		Age pregnancy/Family history > 20	Weight
9	65–69	0.586	34	No family history	1.00
10	70–74	0.646	35	1 family history	2.61
11	75–79	0.713	36	≥ 2 family history	6.8
12	80–84	0.659		Age pregnancy/Family history 20–24	Weight
	Age/Race non-hispanic white	Weight	37	No family history	1.24
13	18–29	0.049	38	1 family history	2.68
14	30–34	0.134	39	≥ 2 family history	5.78
15	35–39	0.278		Age pregnancy/Family history 25–29	Weight
16	40–44	0.450	40	No family history	1.55
17	45–49	0.584	41	1 family history	2.76
18	50–54	0.703	42	≥ 2 family history	4.91
19	55–59	0.859		Age pregnancy/Family history ≥ 30	Weight
20	60–64	1.018	43	No family history	1.93
21	65–69	1.116	44	1 family history	2.83
22	70–74	1.157	45	≥ 2 family history	4.17
23	75–79	1.140		Hiper.atip	Weight
24	80–84	1.006	46	No biopsy	1.00
	Age menarche	Weight	47	1 biopsy and no atypical hyperplasia	0.93
25	Age ≥ 14	1.00	48	No atypical hyperplasia was found	1.00
26	Age 12–13	1.10	49	Atypical hyperplasia was found	1.82
27	Age > 12	1.21			

The model indicates an increased risk of developing breast cancer in the next five years if the score is 1.66 or higher. Table 12.2 shows the weights assigned to each value of the risk factors according to Gail’s model. The score is calculated as follows:

$$\text{Relative risk} = \text{AgeMenarche} \times \text{NumBiopsAgeOrientation} \times \text{AgePregnancyFH} \times \text{Hiper.atip.} \tag{12.1}$$

$$\text{FiveYearRisk} = \text{Relative risk} \times \text{AgeRace} \tag{12.2}$$

where:

- AgeMenarche: The age of the first occurrence of menstruation.
- NumBiopsAgeOrientation: Number of biopsies and first age at which you received breast screening information.
- AgePregnancyFH: Age at first pregnancy and whether you have a family history (FH) of any family member with first-degree breast cancer.
- Hyper.atyp: If the patient presents atypical hyperplasia.
- AgeRace: The patient's current age and race.

We evaluated the predictive performance of the Gail model. The initial analysis revealed that this model has a precision of 67% and an accuracy of 60%. These results show the Gail model's poor performance and the need to improve prediction precision.

12.3.2 Pearson Correlation as Objective Value

Pearson correlation, also known as the correlation coefficient, is a statistical measure that describes the relationship between two sets of data or variables [5]. In other words, it indicates if two variables move together or are related. The most commonly used statistical measure in the literature is Pearson's correlation. The mathematical formula for Pearson's correlation between two variables, X and Y, is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \cdot \sum (Y_i - \bar{Y})^2}} \quad (12.3)$$

where:

X_i, Y_i are the individual values of the variables X and Y.

\bar{X}, \bar{Y} are the means of X and Y, respectively.

r is the Pearson estimation coefficient.

Pearson's correlation produces an r value that varies between -1 and 1 , where:

- $r = 1$: Perfect positive correlation.
- $r = -1$: Perfect negative correlation.
- $r = 0$: No linear correlation.

12.4 Harmony Search Algorithm to Produce New Weightings

The Harmony Search algorithm was first introduced by Geem [15] in 2001. It is a meta-heuristic optimization algorithm inspired by musical process improvisation, where musicians gradually adjust their notes to achieve a pleasing composition called harmony. Harmony Search uses a set of promising candidate solutions from the Harmony Memory (HM), to create new solutions. If a new solution performs better than the worst in HM, it must replace it. In the next sections, we will explain how to evaluate harmony and describe the Harmony Search Heuristic in detail. In contrast to other meta-heuristic algorithms, this algorithm does not perform complex mathematical computations, allowing fast processing. Additionally, it is adaptable to discrete and continuous problems. As mentioned above, we propose a new weighting model using a Harmony Search algorithm to achieve higher precision in predicting breast cancer using risk factors. Figure 12.2 shows the steps of the method.

Algorithm 1 presents the pseudocode of harmony search. We define the algorithm parameters in line 1. The algorithm begins by creating an initial collection of solutions called *harmonies*. Each harmony represents a 1-dimensional vector containing the values for optimization. The HM is a matrix that stores the best-found harmonies.

We define a memory size of eleven harmonies, where each harmony stores initial random values, and one of these is used to replace the worst-quality harmony. The weights of the value associated with each risk factor are produced randomly from 0 to 20, differing from Gail’s model, which relies on decimal numbers. Overall, we considered 39 weights, one for each value of the risk factors.

The HMCR denotes the probability of selecting an individual from the population, set at 0.7, falling within the typical range proposed in the literature, 0.7–0.95. The PAR determines the probability of modifying a selected individual by slightly

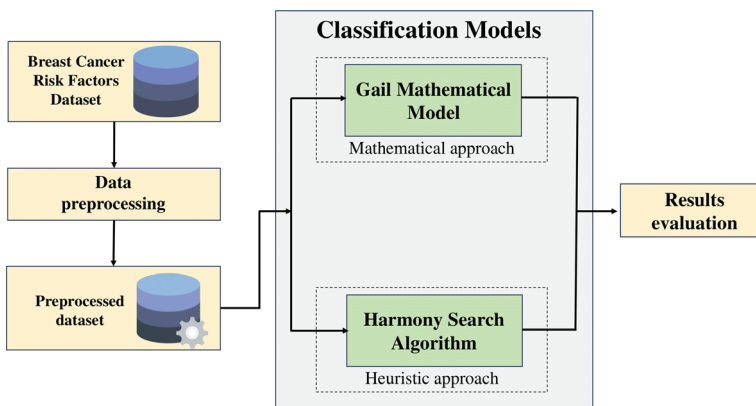


Fig. 12.2 Proposed method

Algorithm 1 Harmony Search Algorithm (HSA)

```

1: Input: Pitchnum, Pitchbounds, HMemorysize, HMCR, PAR, Improvisationmax
2: Out: Harmonybest
3: Harmonies ← InitializeHarmonyMemory(Pitchnum, Pitchbounds, HMemorysize);
4: EvaluateHarmonies(Harmonies);
5: for  $i$  to Improvisationmax do
6:   Harmony ←  $\emptyset$ 
7:   for all Pitch $i$  ∈ Pitchnum do
8:     if Rand() ≤ HMCR then
9:        $X_{new, j} = X_{k, j}$ 
10:    else
11:       $X_{new, j} = (Max - Min) \cdot rand(0, 1) + Min$ 
12:    end if
13:    if Rand() ≤ PAR then
14:       $X_{new, j} = (Max - Min) \cdot rand(-1, 1) \cdot 0.3 + x$ 
15:    end if
16:  end for
17:  if f(Harmony $i$ ) ≤ f(Worst(Harmonies)) then
18:    Worst(Harmonies) ← Harmony $i$ ;
19:  end if
20: end for
21: return Best(Harmonies);

```

adjusting its values, established at 0.3, within the commonly used range of 0.1–0.5. Finally, we set the stopping criterion based on the number of required improvisations.

The HSA uses an objective value to assess and continually improve the solutions generated throughout the search process, aiming to converge towards the best possible solution for the given problem. We used Pearson’s correlation as the objective value to measure the relationship between the Calculated Risk Value (CRV) and the current class. It does not matter if the values differ; the important part is that high CRV corresponds to a 1 (positive case), while low CRVs should be associated with a 0 (negative case). Thus the correlation seems useful for associating high and low CRVs with positive and negative cases, respectively.

12.4.1 Harmony Improvisation

In the improvisation process, the initial step is to find the worst solution among HM. We generate a new harmony vector $X_{new} = [X_{new,1}, X_{new,2}, \dots, X_{new,d}]$ rewriting that worst solution. Table 12.3 shows items for each attribute representing harmony; e.g., age ranges from 18–29 to 80–84 years, while age at menarche is less than 12 years, 12–13, and greater than or equal to 14. Calculating the objective value is essential for each harmony, and we calculate it using the correlation. Once all harmonies have been generated, we will evaluate their performance to identify the one with the lowest objective value. This will be replaced by the next harmony generated by the algorithm.

Table 12.3 Harmony memory

	Age			Age menarche			...	Biopsy		Objective value
	18–29	...	80–84	>12	12 a 13	≥ 14		No	Yes	
Harmony 1	1	...	8	10	6	7	...	2	9	0.2392
Harmony 2	14	...	2	6	9	4	...	3	5	0.0444
Harmony 3	9	...	13	12	5	5	...	0	19	0.6681
...
...
...
Harmony N	8	...	2	1	3	6	...	13	2	0.5304

Algorithm 1 of lines 8–15 improvise X_{new} according to the parameters HMCR and PAR. The HMCR determines the probability of using a stored pitch in HM or producing X_{new} randomly. Algorithm 1 in lines 8–12 iterates over $X_{new,j} \forall j \in \text{Pitch}$ until completing the new harmony. We yield a random number with a uniform distribution between [0,1]. Where $X_{new,j}$ is a new harmony for the elements of each decision variable, Max and Min is a value maximum or minimum of pitch between 0–20 of the risk factors elements.

PAR decides to seek a neighboring value, helping to avoid being trapped in local optima. PAR controls the frequency of adjusting selected pitches from the HM; smaller values of PAR indicate fewer adjustments, while larger values support more pitch adjustments. We yield a random number with a uniform distribution in $[-1, 1]$ shown in Algorithm 1 in lines 13–15. Where $X_{new,j}$ is a new harmony for the elements of each decision variable, Max and Min is a value maximum or minimum between 0–20, and x is a value within the max-min range. We evaluate the newly created harmonies using the objective value, where in each iteration, we choose whether or not to modify the worst harmony in HM.

12.4.2 Solution Interpretation

In Sect. 12.4.1, we discussed the improvisation of HSA and the selection of the least favorable harmony. In this section, we will approach the interpretation of the obtained solutions. Table 12.4 shows detailed information for each pitch, showing for the variable age twelve elements, two for race, three for age at menarche, and thus for each attribute. In total, there are 39 pitches for weighting.

Table 12.5 shows the solution with the weights assigned by HSA to each risk factor element. The values assigned in the *Weight* column indicate the relative importance of the elements in each risk factor, providing a general idea of the relevant elements within the data set.

Table 12.5 Solutions representation

Pitch	Elements of each attribute	Weight
Pitch 1	Age 18–29	8
Pitch 2	Age 30–34	8
Pitch 3	Age 35–39	8
Pitch 4	Age 40–44	9
Pitch 5	Age 45–49	10
...
...
...
Pitch 38	Bioph No	0
Pitch 39	Bioph Yes	19

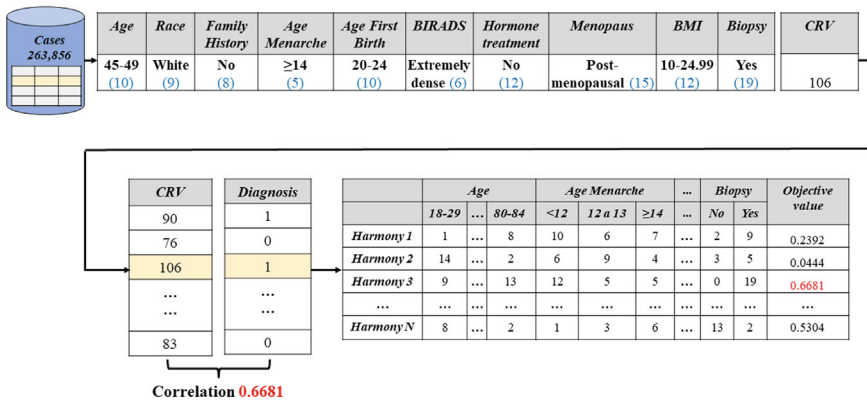


Fig. 12.3 Example to understand the HSA weighting procedure and evaluation

Once the algorithm has proposed a set of weights, we apply them to our data set. Figure 12.3 illustrates an example case of an individual with specific features, showing the corresponding values for each pitch, see Table 12.5. After assigning values to each element of the attribute, we sum her specific weights (specific pitches) in the CRV structure. For this example, the sum is 106. The process is performed for all cases in the dataset using the same procedure. Finally, we calculate the correlation between the column *Diagnosis* and CRV. Where the *Diagnosis* column is the actual classification as a binary category. A significant positive correlation suggests that as the CRV increases, the probability of having cancer, according to the actual diagnosis, also increases. Conversely, a significant negative correlation indicates the opposite. We aim to achieve a high correlation between breast cancer classification and the derived values from CVR. To achieve this, the heuristic algorithm will adjust the weights of the risk factor values to improve their correlation with the dataset class.

12.4.3 Update and Stopping Criteria

The update process is crucial, as it involves the selection of the least suitable harmony in HM during each iteration. When creating a *new harmony*, we evaluate its performance based on the corresponding value of the objective value. The algorithm returns the best harmony stored in HM. This selection, improvisation, and updating process is iterated until the algorithm reaches ten thousand consecutive iterations without improvement. It is necessary to mention that when an iteration starts with the worst correlation, it is unlikely to outperform the best one. Therefore, if the worst correlation is no longer the worst, we consider it an improvement, and we restart the counter of iterations without improvement.

12.5 Experimental Results

For the experimentation, we used an Intel (R) Core(TM) i5 processor laptop with M 540 CPU @ 2.53 GHz, 6GB RAM, 64-bit, Windows 10, and 120GB SSD. We used Python 3.9.13 as a programming language and Pycharm for implementation. In this study, we evaluated the prediction model performance for detecting breast cancer by analyzing a confusion matrix. Table 12.6 shows the structure of our confusion matrix.

Where:

- TP: Positive class predicted as positive.
- TN: Negative class predicted as negative.
- FN: Positive class predicted as negative.
- FP: Negative class predicted as positive.

We used five widely-used [29, 31, 40] performance metrics from the literature to evaluate the model performance: accuracy, precision, sensitivity or also known as recall, specificity, and F1 Score, as presented in Eqs. (12.4)–(12.8).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (12.4)$$

Table 12.6 Confusion matrix

		Predicted	
		Positive	Negative
Actual	Positive	TP	FP
	Negative	FN	TN

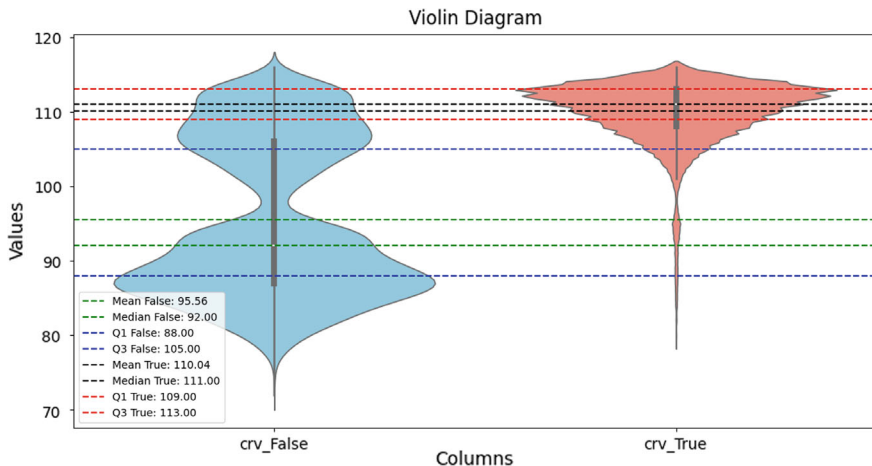


Fig. 12.4 Violin diagram representing negative and positive diagnosis

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12.5}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{12.6}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{12.7}$$

$$\text{F1 Score} = 2 \cdot \left(\frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \right) \tag{12.8}$$

The objective of this paper is to improve the classification rate compared with the Gail model. Therefore, we calculated the CRV average of the positive and negative cases to identify the threshold value that should separate positive from negative cases. We obtained a CRV average for all the cases and computationally tested values above and below the CRV average to find the best threshold. In Table 12.7, we show the accuracy and precision for the 101, 102, and 103 thresholds, where a threshold of 102 implies that CRVs above 102 will be classified as a positive case while below or equal to 102 the case will be classified as negative.

Additionally, Fig. 12.4 shows the distribution of positive and negative diagnoses. The positives show uniformly high numbers, clearly clustered above 102, while negatives fall into two groups; however, as we focused on the positive cases, the CRV distribution seems promising.

Table 12.7 presents the name of each metric in column one, while columns two to five display the corresponding percentages for each metric for both the Gail model and

Table 12.7 Comparative table of the Gail model and the proposed model with different thresholds and their respective evaluation metrics

	Gail model	Proposed model 101	Proposed model 102	Proposed model 103
Metrics	%	%	%	%
Accuracy (%)	60	80	81	81
Precision (%)	67	97	96	95
Sensitivity (%)	58	73	74	74
Specificity (%)	61	95	94	93
F1 score (%)	62	83	83	83

our proposed model with thresholds of 102. However, the threshold of 102 shows high sensitivity, thus achieving the objective of increasing the detection of positive cases and reducing the incidence of false positives in breast cancer diagnosis. Therefore, we select the threshold of 102.

Our model, with a threshold of 102, shows a significant improvement by obtaining a precision of 96% versus 67% for the Gail model. Here, it demonstrates that our proposal identifies positive cases with 29% higher precision. Additionally, our model shows an accuracy of 81%, surpassing the 60% of the Gail model, indicating that it improved in predicting both positive and negative cases. Boasting a sensitivity of 74% compared to the Gail model's 58%, our model effectively distinguishes between positive and negative cases. With a specificity of 94% versus 61% for Gail's model, our model effectively outperforms Gail's ability to identify negative cases. Finally, our model achieves an F1 score of 83%, suggesting a balanced identification of positive and negative cases compared to Gail's model of 62%. Table 12.8 shows 39 values for each element of each risk factor, representing the final weighting assigned by the harmony search algorithm.

Tables 12.9 and 12.10 show the classification results comparing Gail's model with our proposed model versus the real classifications. The sample for the real classifications comprises 132,042 cases diagnosed with breast cancer and 131,814 cases without cancer. The results show that our proposed model increased the number of true positive (TP) cases by approximately 29% compared to Gail, and the number of true negative (TN) cases improved by 13% over Gail.

Additionally, we applied the Wilcoxon test [2, 13], which showed that our proposed model effectively distinguishes between positive and negative cases. Highly significant differences were observed in positive and negative cases with ($p < 0.000$). Hence, our proposal is advisable for predicting breast cancer.

Table 12.8 Description and the final weighting of each attribute

#	Age	Weight	#	Age first birth	Weight
1	18–29	8	22	Age 25–29	10
2	30–34	8	23	Age ≥30	9
3	35–39	8	24	Nuliparous	10
4	40–44	9		<i>BIRADS</i>	
5	45–49	10	25	Almost entirely fat	8
6	50–54	11	26	Scattered fibroglandular densities	10
7	55–59	13	27	Heterogeneously dense	10
8	60–64	14	28	Extremely dense	6
9	65–69	14		<i>Hormone treatment</i>	
10	70–74	14	29	No	12
11	75–79	15	30	Yes	4
12	80–84	15		<i>Menopaus</i>	
	<i>Race</i>		31	Pre-menopausal	14
13	Non-hispanic white	9	32	Post-menopausal	15
14	Non-hispanic black	5	33	Surgical menopause	13
	<i>Family History</i>			<i>BMI</i>	
15	No	8	34	10–24.99	12
16	Yes	7	35	25–29.99	11
	<i>Age menarche</i>		36	30–34.99	10
17	Age ≥14	5	37	35 or more	10
18	Age 12–13	6		<i>Biopsy</i>	
19	Age > 12	6	38	No	0
	<i>Age first birth</i>		39	Yes	19
20	Age > 20	9			
21	Age 20–24	10			

Table 12.9 Gail model confusion matrix

Gail model			
		Predicted	
		Positive	Negative
Actual	Positive	87,996	44,046
	Negative	62,516	69,298

Table 12.10 Proposed model confusion matrix

		Proposed model 102	
		Predicted	
		Positive	Negative
Actual	Positive	126,642	5400
	Negative	45,571	86,243

12.6 Discussion

The method proposed aim is to increase the detection of positive cases and reduce the incidence of false positives in breast cancer diagnosis. Detection of positive cases allows earlier diagnosis and appropriate disease care. Additionally, reducing the incidence of false positives can minimize unnecessary anxiety and stress for patients and optimize health system resources by avoiding unnecessary tests and procedures. Experiments showed that weighting each element of attributes with the harmony search algorithm significantly improves the model classification.

Additionally, Table 12.8 provides some information regarding the proposed weights of the Harmony search algorithm. It is important to highlight that the larger the weight value, the higher the risk of breast cancer; i.e., we established the threshold of 102, where a CVR higher than 102 will be classified as a positive case, while lower values than 102 will be classified as a negative case.

Here, we can see that as age increases, the cancer risk also increases according to the proposed weights. According to [35], there is a lower risk in the group of women less than 39 years old due to protection factors like pregnancy and breast-feeding. However, the aging process is related to the deterioration of proteins and DNA in cells, which triggers oxidative stress and genomic instability accumulation. Oxidative stress happens when free radicals and reactive oxygen species damage biomolecules like lipids, proteins, and DNA. This damage disrupts their function and causes membrane permeability disorders, leading to cancer. These findings align with the proposed weights for this risk factor [20].

In breast cancer, several genes predispose to its development, with the most well-known and high-risk ones being BRCA1 and BRCA2 [46]. Additionally, there is a theory that mutations associated with medium and low risk occur in African American individuals for genes such as TOX3, APOBEC3, ATM, and NBN, among others, which might explain the low weights proposed by the HS algorithm on non-hispanic black women in comparison with non-Hispanic white women [4].

Regarding family history, the results show a one-point weight difference between individuals with and without a family history of cancer, proposing less risk to those without a family history; this contrasts with other case studies. We believe that individuals with a family history of cancer carry out more self-examination and diagnostic studies, which might introduce bias into previous studies. Another reason for these results might be an increase in breast cancer due to other factors not related

to genetics [4]; therefore, there would be a lower proportion of women with breast cancer who have a relative with the same condition than women with breast cancer without a relative with breast cancer.

As age at menarche decreases, the risk of breast cancer increases due to greater exposure to estrogens and increased hormone availability in the mammary gland tissue. This exposure to estrogens is directly related to an increased risk of breast cancer, which concurs with the weights proposed by the algorithm in this study. Regarding age at first birth, similar to previous weights, there is only a one-point difference among different age ranges. Women who gave birth at a younger age (below 20 years) or an age equal to or greater than 30 years are favored with the lower weight of 9. The theory suggests that protection through estrogen receptors begins ten years after the first pregnancy [20]. Considering this, we could argue that women who had their first pregnancy at or after 30 years of age would have protection up to 40 years or beyond, which, depending on the age-related risks, could contribute to the observed value of 9 by protecting women in those risky ranges of ages. However, we do not find an evident reason to explain the benefits of the group of women with their first pregnancy before the age of 20 compared to the other age groups. It could be an interesting analysis for future work.

Regarding BI-RADS breast density, it is shown that extremely high breast density provides a lower risk, followed by the category where breast density is almost entirely fat. Scattered fibroglandular and heterogeneously dense classifications contain significant glandular and fibrous connective tissues with less fatty tissue [25]. These dense breasts have a higher predisposition to estrogen receptors, which can contribute to cancer development; these findings concur with the proposed algorithm weightings. However, extremely dense breasts are more challenging to identify cancer, potentially leading to inadequate detection; that category also processes a substantial proportion of glandular and connective tissue [33]. For this last category, the algorithm proposes a lower weight than for the two previous categories, associating a lower risk of breast cancer; this result might be due to little data, about 10% of the population. There is also the possibility of other subjacent explanations yet to be found by scientists.

Regarding Hormone Treatment (HT), prolonged use of oral contraceptives (more than five years) may lead to an increased risk of breast cancer because estrogen and progesterone have a stimulating effect on proliferative mammary cells [9]. The only method that has not shown an increased risk of breast cancer is vaginal estrogen. However, individuals undergoing HT may have already undergone a prior study and consider their cancer risk to be low enough to use it; therefore, this might associate the use of HT with a low risk of breast cancer for our study.

Regarding menopause, specifically in pre-menopause, two key hormones are the luteinizing hormone (LH) and the follicle-stimulating hormone (FSH). These hormones play a crucial role in regulating ovarian function before menopause. During the transition to menopause, estrogen levels decrease while FSH and LH increase [25]. This hormonal imbalance can lead to alterations in menstrual cycles, including reduced frequency, irregularity, and changes in bleeding duration. Essentially, as ovulation becomes less efficient due to hormonal deficiency, menstrual cycles

become less predictable; therefore, other estrogen receptors, like in breasts, might be affected by some free estrogen. Furthermore, in post-menopause, the ovaries are no longer functional; therefore, estrogen levels, which should now be lower than before menopause, are no longer captured by the ovaries. Instead, they remain free and are taken explicitly by the mammary tissue due to their glandular and tissular structure [25]. Additionally, women who undergo Surgical Menopause through ovary removal (oophorectomy) or uterus removal (hysterectomy) experience menopause differently. On the one hand, oophorectomy poses a more abrupt menopause without the gradual transition. On the other hand, hysterectomy does not technically imply menopause until the ovaries cease functioning, even if ovulation has already stopped. Some studies suggest that women who had a hysterectomy and subsequently used conjugated equine estrogen had a decreased incidence and mortality rate of breast cancer [25]. Additionally, undergoing a hysterectomy before menopause is associated with a lower risk of invasive breast cancer in younger women. Therefore, regarding menopause, the proposed weights reflect the knowledge in the literature.

Regarding the body mass index (BMI), and according to the literature, having a larger size and weight, $BMI \geq 25 \text{ kg/m}^2$, is associated with an 82% increased risk of developing breast cancer in postmenopausal women. Obesity is linked to a higher risk of inflammatory breast cancer in premenopausal women. A significant weight gain is related to up to 28% more risk. For every 5 kg gained, there is a 23% increase in risk—especially in women who gained weight between the ages of 18 and the year before a positive breast cancer diagnosis [20]. Obesity is a significant risk factor because adipose tissue aids in the uptake of the enzyme aromatase, which is a crucial step in estrogen biosynthesis. Nonetheless, this behavior is not present in the proposed weights for this risk factor; on the contrary, there is a slight tendency for lower BMI values to have higher weight values, while higher BMI is associated with lower weight values. These weights might be related to the BI-RADS classification, where breasts that are mostly fat have a lower risk of developing breast cancer. On the other hand, other factors are still unknown to scientists, and further research could be helpful. Finally, if a patient performed a biopsy, it is most likely because a specialist found an anomaly in a mammogram and ordered a biopsy. Therefore, this represents an increase in the risk of breast cancer, which was correctly weighted by the HS algorithm.

According to this study, the weighting obtained using the HSA, as shown in Table 12.8, and the simple summation method to produce the CRV simplify a patient's consultation to determine her probability of developing breast cancer. For example, if a patient has the factors shown in Fig. 12.5, the sum of the values of the elements of each attribute will yield a result less than or greater than 102. In the example provided, the total is 87, i.e., less than 102, suggesting a low probability of developing breast cancer. Based on the result obtained, you can get help to take preventive measures. Additionally, specifically for BIRADS, if the patient does not know her BIRADS classification, she can use both the lowest and highest proposed weights, 6–10, and ponder the obtained results.

The results show the efficiency of the meta-heuristic approach adopted in this study. The reduction of false positives and the improvement of performance met-

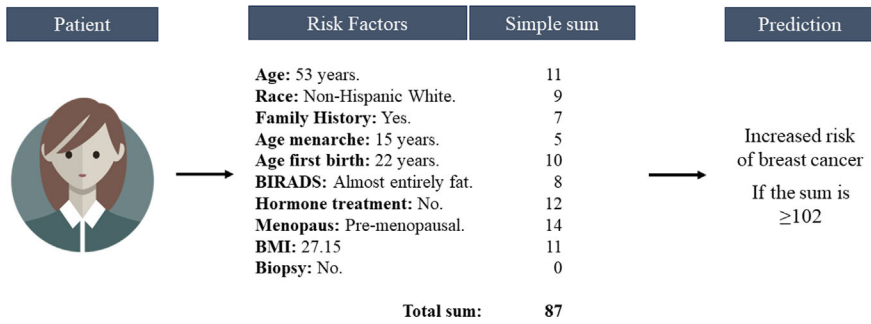


Fig. 12.5 This example case shows the simplicity of using the proposed model

rics demonstrate the ability of the model to distinguish between positive and negative cases more accurately. The results confirm the importance of assigning better weights to risk factors to improve performance. When comparing the Gail model in breast cancer prediction, it is possible to observe some limitations of classical models, demonstrating the importance of using advanced optimization techniques to obtain better weightings to evaluate breast cancer risk. Future research can apply this method to other meta-heuristic algorithms in various clinical contexts and consider the integration of multiple predictive models.

12.7 Conclusion and Future Work

In this paper, we present a model based on the HSA designed to determine the optimal weights for each value of the risk factors in a breast cancer dataset, aiming to improve disease prediction. This model offers a practical and easy-to-use alternative with a high patient-friendly emphasis, characterized by not requiring decimal numbers nor complicated calculus. Additionally, this study emphasizes the importance of achieving a high number of true positives while minimizing false positives. A high number of false positives could delay necessary treatment, putting patients' lives at risk. On the other hand, while false negatives are also concerning, they typically lead to further evaluation by an oncologist or additional medical tests; although not ideal, these scenarios do not pose an immediate risk to patients' lives.

As the results show, appropriate weighting of the elements associated with each risk factor presents an innovative and promising approach to improve accuracy. The results also show that the algorithm stands out for its ability to adapt, explore, and identify harmonies among different elements. The main contribution of this paper consists of finding new weights for each of the different elements of the risk factors using a simple approximation with integers, which facilitates the use of the model and improves the previous proposal using heuristic optimization techniques. As future work, we recommend adding demographic and clinical characteristics, performing

comprehensive comparisons with other metaheuristic optimization methods, and carrying out a variety of adjustments to the algorithm's parameters. Finally, it would be ideal to conduct clinical validation studies incorporating more diverse patient data to assess the performance of our proposal.

Acknowledgements Thanks to CONACYT for the support with the number 1010753.

References

1. A. Al-Omouh, A., A. Alsewari, A., S. Alamri, H., Z. Zamli, K.: Comprehensive review of the development of the harmony search algorithm and its applications. *Institute of Electrical and Electronics Engineers* **7**, 14233–14245 (2019). <https://doi.org/10.1109/access.2019.2893662>. <https://ieeexplore.ieee.org/abstract/document/8616762>
2. Alhussan, A.A., Abdelhamid, A.A., Towfek, S.K., Ibrahim, A., Eid, M.M., Khafaga, D.S., Saraya, M.S.: Classification of diabetes using feature selection and hybrid al-biruni earth radius and dipper throated optimization. *Diagnostics* **13**, 2038 (2023). <https://doi.org/10.3390/diagnostics13122038>
3. Ali, W., Ahmed, A.A.: Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting. *Institution of Engineering and Technology* **13**(6), 519–710 (2019). <https://doi.org/10.1049/iet-ifs.2019.0006>.
4. Amadou, A., Mejía, G.T., Hainaut, P., Romieu, I.: Breast cancer in latin america: global burden, patterns, and risk factors. *Salud Pública de México* **56**, 547–554 (2014). <https://doi.org/10.1016/j.semradonc.2015.09.004>
5. Baak, M., Koopman, R., Snoek, H., Klous, S.: A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *Computational Statistics and Data Analysis* **152**, 1–25 (2020). <https://doi.org/10.1016/j.csda.2020.107043>
6. Baskoro, S., Sunindy, W.D.: Predicting issue handling process using case attributes and categorical variable encoding techniques. 2019 International Conference on Data and Software Engineering (ICoDSE) pp. 1–5 (2019). <https://doi.org/10.1109/ICoDSE48700.2019.9092617>
7. Brownlee, J.: *Clever algorithms : nature-inspired programming recipes*, 1 edn. (2011)
8. Castellanos, A., Cruz-Reyes, L., Fernández, E., Rivera, G., Gomez-Santillan, C., Rangel-Valdez, N.: Hybridisation of swarm intelligence algorithms with multi-criteria ordinal classification: a strategy to address many-objective optimisation. *Mathematics* **10**(3), 322 (2022). <https://doi.org/10.3390/math10030322>
9. Chlebowski, R.T., Anderson, G.L., Aragaki, A.K., Manson, J.E., Stefanick, M.L., Pan, K., Barrington, W., Kuller, J.H., Simon, M.S., Lane, D., Johnson, K.C., Rohan, T.E., Gass, M.L.S., Cauley, J.A., Paskett, E.D., Sattari, M., Prentice, R.L.: Association of menopausal hormone therapy with breast cancer incidence and mortality during long-term follow-up of the women's health initiative randomized clinical trials. *Journal of Clinical Oncology* **324**, 369–380 (2020). <https://doi.org/10.1001/jama.2020.9482>
10. Doppala, B.P., Bhattacharyya, D., Chakkravarthy, M., hoon Kim, T.: A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset. *Distributed and Parallel Databases* **41**, 1–20 (2023). <https://doi.org/10.1007/s10619-021-07329-y>. <https://link.springer.com/article/10.1007/s10619-021-07329-y#citeas>
11. van den Ende, C., Oordt-Speets, A.M., Vroling, H., van Agt, H.M.E.: Benefits and harms of breast cancer screening with mammography in women aged 40–49 years: A systematic review. *International Journal of Cancer* **141**, 1295–1306 (2017). <https://doi.org/10.1002/ijc.30794>. <https://onlinelibrary.wiley.com/doi/epdf/10.1002/ijc.30794>

12. Fadhil, A.A., Hussein, H.A., Jawad, M., W Khaled, Y.A., Hussein, A.S., Jawad, M.A., Samein, L.H., Mohammed, N.M., Sherif, B.K., Obaid, A.J.: Identify breast cancer risk factors using the gail assessment model in iraq. *Arch Razi Inst.* **77**, 1901–1907 (2022). <https://doi.org/10.22092/ari.2022.359509.2436>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10133634/>
13. García, S., Luengo, J., Herrera, F.: *Data Preprocessing in Data Mining*. Intelligent Systems Reference Library (2015). <http://www.springer.com/series/8578>
14. Geem, Z.W.: *Music-Inspired Harmony Search Algorithm. Theory and Applications*, vol. 191, 1 edn. Springer (2009). <https://doi.org/10.1007/978-3-642-00185-7>
15. Geem, Z.W., Kim, J.H., Loganathan, G.V.: A new heuristic optimization algorithm: Harmony search. *Simulation* **76**, 60–68 (2001). <https://doi.org/10.1177/003754970107600201>
16. Grimm, L.J., Avery, C.S., Hendrick, E., Baker, J.A.: Benefits and risks of mammography screening in women ages 40 to 49 years. *Journal of Primary Care Community Health* **13**, 1–6 (2022). <https://doi.org/10.1177/21501327211058322>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8796062/>
17. Han, J., Kamber, M., Pei, J.: *Data Mining. Concepts and Techniques*, 3 edn. Elsevier (2012)
18. Herman-Saffar, O., Boger, Z., Libson, S., Lieberman, D., Gonen, R., Zeiri, Y.: Early non-invasive detection of breast cancer using exhaled breath and urine analysis. *Computers in Biology and Medicine* **96**, 227–232 (2018). <https://doi.org/10.1016/j.compbiomed.2018.04.002>. <https://www.sciencedirect.com/science/article/abs/pii/S0010482518300775>
19. Hou, J., Ye, X., Feng, W., Zhang, Q., Han, Y., Liu, Y., Li, Y., Wei, Y.: Distance correlation application to gene co-expression network analysis. *BMC Bioinformatics* **23**, 2–24 (2022). <https://doi.org/10.1186/s12859-022-04609-x>. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04609-x>
20. Ibarra, M.J.N., Juvera, G.C., Vélez, M.I.O., Villar, A.V.B., del Socorro Saucedo Tamayo, M.: Influencia de los factores reproductivos, la lactancia materna y la obesidad sobre el riesgo de cáncer de mama en mujeres mexicanas. *Nutrición hospitalaria* **32**, 291–298 (2015). <https://doi.org/10.3305/nh.2015.32.1.9049>. <http://www.nutricionhospitalaria.com/pdf/9049.pdf>
21. INEGI: Estadísticas a propósito del día internacional de la lucha contra el cáncer de mama. (2021). https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2021/EAP_LUCHACANCER-2021.pdf
22. INEGI: Estadísticas a propósito del día internacional de la lucha contra el cáncer de mama (19 de octubre). (2022). https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2022/EAP_CANMAMA22.pdf
23. INEGI: Estadísticas a propósito del día internacional de la lucha contra el cáncer de mama (19 de octubre). (2023). https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2023/EAP_CMAMA23.pdf
24. Jebli, I., Belouadha, F.Z., Kabbaj, M.I., Tilioua, A.: Prediction of solar energy guided by pearson correlation using machine learning. *Energy* **224**, 120109–120109 (2021). <https://doi.org/10.1016/j.energy.2021.120109>. <https://www.sciencedirect.com/science/article/abs/pii/S0360544221003583>
25. Kerlikowske, K., Cook, A.J., Buist, D.S., Cummings, S.R., Vachon, C., Vacek, P., Miglioretti, D.L.: Breast cancer risk by breast density, menopause, and postmenopausal hormone therapy use. *Journal of Clinical Oncology* **28**, 3830–3837 (2010). <https://doi.org/10.1200/JCO.2009.26.4770>
26. Khamparia, A., Bharati, S., Podder, P., Deepak, G., Khanna, A., Phung, T.K., H Thanh, D.N.: Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. *Multidimensional Systems and Signal Processing* **32**, 747–765 (2021). <https://doi.org/10.1007/s11045-020-00756-7>. <https://link.springer.com/article/10.1007/s11045-020-00756-7>
27. Khozama, S., Mayya, A.M.: Study the effect of the risk factors in the estimation of the breast cancer risk score using machine learning. *Biomedical Engineering* **22**, 3543–3551 (2021). <https://doi.org/10.31557/apjcp.2021.22.11.3543>. https://journal.waocp.org/article_89845.html
28. Kumar, N., Singh, V., Mehta, G.: Assessment of common risk factors and validation of the gail model for breast cancer: A hospital-based study from western india. *Tzu Chi Medical*

- J **32**, 362–366 (2020). https://doi.org/10.4103/tcmj.tcmj_171_19. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7605293/>
29. Larner, A.J.: The 2x2 matrix. (2021). <https://doi.org/10.1007/978-3-030-74920-0>
 30. Li, J., Guan, X., Fan, Z., Ching, L.M., Li, Y., Wang, X., Cao, W.M., Liu, D.X.: Non-invasive biomarkers for early detection of breast cancer. *Cancers* **12**, 2767–2794 (2020). <https://doi.org/10.3390/cancers12102767>. <https://www.mdpi.com/2072-6694/12/10/2767>
 31. Luque, A., Carrasco, A., Martín, A., de las Heras, A.: The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition* **91**, 216–231 (2019). <https://doi.org/10.1016/j.patcog.2019.02.023>. <https://www.sciencedirect.com/science/article/pii/S0031320319300950>
 32. Ming, C., Viassolo, V., Probst-Hensch, N., Dinov, I.D., Chappuis, P.O., Katapodi, M.C.: Machine learning-based lifetime breast cancer risk reclassification compared with the boadicea model: impact on screening recommendations. *British Journal of Cancer* **123**, 860–867 (2020). <https://doi.org/10.1038/s41416-020-0937-0>. <https://www.nature.com/articles/s41416-020-0937-0>
 33. Neira, P.: Densidad mamaria y riesgo de cáncer mamario. *Revista Médica Clínica Las Condes* **24**, 122–130 (2013). [https://doi.org/10.1016/S0716-8640\(13\)70137-8](https://doi.org/10.1016/S0716-8640(13)70137-8)
 34. Nouira, K., Maalej, Z., Rejab, F.B., Ouerfelly, L., Ferchichi, A.: Analysis of breast cancer data: a comparative study on different feature selection techniques. 2020 International Multi-Conference on: “Organization of Knowledge and Advanced Technologies” (OCTA) pp. 1–11 (2020). <https://doi.org/10.1109/octa49274.2020.9151824>. <https://ieeexplore.ieee.org/abstract/document/9151824>
 35. O’Brien, K.M., Sun, J., Sandler, D.P., DeRoo, L.A., Weinberg, C.R.: Risk factors for young-onset invasive and in situ breast cancer. *Cancer Causes & Control* **26**, 1771–1778 (2015). <https://doi.org/10.1007/s10552-015-0670-9>
 36. Olmos, J., Florencia, R., García, V., González, M.V., Rivera, G., Sánchez-Solís, P.: Metaheuristics for order picking optimisation: a comparison among three swarm-intelligence algorithms. *Technological and Industrial Applications Associated With Industry 4.0* pp. 177–194 (2022). https://doi.org/10.1007/978-3-030-68663-5_13
 37. OMS: Cáncer de mama (2023). <https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer>
 38. Rivera, G., Coello, C.A.C., Cruz-Reyes, L., Fernandez, E.R., Gomez-Santillan, C., Rangel-Valdez, N.: Preference incorporation into many-objective optimization: an ant colony algorithm based on interval outranking. *Swarm and Evolutionary Computation* **69**, 101024 (2022). <https://doi.org/10.1016/j.swevo.2021.101024>
 39. Saleh, B., Elhawary, M.A., Mohamed, M.E., Ali, I.N., Zayat, M.S.E., Mohamed, H.: Gail model utilization in predicting breast cancer risk in egyptian women: a cross-sectional study. *Breast Cancer Research and Treatment* **188**, 749–758 (2021). <https://doi.org/10.1007/s10549-021-06200-z>. <https://link.springer.com/article/10.1007/s10549-021-06200-z>
 40. Sammut, C., Webb, G.I.: *Encyclopedia of Machine Learning*. Springer (2011)
 41. Schober, P., Boer, C., Schwarte, L.A.: Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia* **126**, 1763–1768 (2018). <https://doi.org/10.1213/ane.0000000000002864>. https://journals.lww.com/anesthesia-analgesia/fulltext/2018/05000/correlation_coefficients__appropriate_use_and.50.aspx
 42. Toaza, B., Esztergár-Kiss, D.: A review of metaheuristic algorithms for solving tsp-based scheduling optimization problems. *Applied Soft Computing* **148**, 1568–1592 (2023). <https://doi.org/10.1016/j.asoc.2023.110908>. <https://www.sciencedirect.com/science/article/pii/S1568494623009262>
 43. Valero, M.G., Zabor, E.C., Park, A., Gilbert, E., Newman, A., King, T.A., Pilewskie, M.L.: The tyler-cuzick model inaccurately predicts invasive breast cancer risk in women with lcis. *Annals of Surgical Oncology* **27**, 736–740 (2019). <https://doi.org/10.1245/s10434-019-07814-w>. <https://link.springer.com/article/10.1245/s10434-019-07814-w>
 44. Vargas-Martínez, M., Rangel-Valdez, N., Fernández, E., Gómez-Santillán, C., Rivera, G., Balderas, F.: Mosa/do and mosad/do-ii: Performance analysis of decomposition-based algo-

- rithms in many objective problems. *SoftwareX* **25**, 101610 (2024). <https://doi.org/10.1016/j.softx.2023.101610>
45. Wang, L., Hu, H., Liu, R., Zhou, X.: An improved differential harmony search algorithm for function optimization problems. *Soft Computing* **23**, 4827–4852 (2018). <https://doi.org/10.1007/s00500-018-3139-4>. <https://link.springer.com/article/10.1007/s00500-018-3139-4>
 46. Wilbur, J.S., Colins, B.L., Penson, R.T., Dizon, D.S.: Breast cancer risk assessment: Moving beyond brca 1 and 2. *Seminars in Radiation Oncology* **26**, 3–8 (2016). <https://doi.org/10.1016/j.semradonc.2015.09.004>.
 47. Zhou, H., Wang, X., Zhu, R.: Feature selection based on mutual information with correlation coefficient. *Applied Intelligence* **52**, 5457–5474 (2021). <https://doi.org/10.1007/s10489-021-02524-x>. <https://link.springer.com/article/10.1007/s10489-021-02524-x>
 48. Zhu, Q., Tang, X., Elahi, A.: Application of the novel harmony search optimization algorithm for dbscan clustering. *Expert Systems with Applications* **178**, 1–12 (2021). <https://doi.org/10.1016/j.eswa.2021.115054>. <https://www.sciencedirect.com/science/article/abs/pii/S0957417421004954>

Part III
Data-Driven Decision

Chapter 13

Solidary Family Business, Intellectual Property and Sustainability in Rural Producers in Mexico: A Hybrid SEM-PLS and Fuzzy Approach



Miguel Reyna-Castillo[✉], Alejandro Santiago[✉], Xóchitl Barrios-del-Angel[✉], and Daniel Bucio-Gutierrez[✉]

Abstract Studies show that the actions of economies with a collaborative approach contribute to agricultural production enterprises adhering to the call for sustainable development with a triple bottom line. The sustainable call implies caring for the environment and the community but also developing the resources and capabilities that allow them a sustainable competitive advantage in the market. The World Intellectual Property Organization (WIPO) has considered that a more effective system of protection of intangibles will encourage investment in intellectual property development as a critical driver of economic well-being. This research aims to analyze the predictive capacity of the features of the solidarity economy and the culture of intellectual property on sustainability in rural family businesses in the agricultural sector in Mexico. The methodology was empirical-mathematical, based on a hybrid approach of structural equations and fuzzy evolutionary logic based on a survey of 88 rural family businesses. The results show that rural family businesses with solidarity economy traits tend to achieve sustainable development in their practices. Theoretical and managerial implications are presented.

M. Reyna-Castillo

Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCyT) & Faculty of Architecture, Design and Urbanism, Autonomous University of Tamaulipas, Ciudad Victoria, Mexico

e-mail: mreyna@docentes.uat.edu.mx

A. Santiago (✉)

Faculty of Engineering Tampico, Autonomous University of Tamaulipas, Ciudad Victoria, Mexico

e-mail: aurelio.santiago@uat.edu.mx

X. Barrios-del-Angel

Tampico School of Commerce and Administration at the Autonomous, University of Tamaulipas, Ciudad Victoria, Mexico

e-mail: axbarrios@docentes.uat.edu.mx

D. Bucio-Gutierrez

Faculty of Architecture, Design and Urbanism, Autonomous University of Tamaulipas, Ciudad Victoria, Mexico

e-mail: danielbucio@docentes.uat.edu.mx

Keywords Solidarity economy · Intellectual property · Sustainability · SEM-Fuzzy · Agricultural sector · Mexico

13.1 Introduction

Contemporary environmental and social compliance challenges have sparked significant interest in cooperative collaboration and intellectual property rights. These factors are crucial to open pathways toward a circular economy [1] and facilitate routes to Sustainable Development [2]. In this sense, the rural family business stands out from other organizations because it is a crucial nucleus for sustainability as it is considered with characteristic respect and love for nature where the capacity of its work is efficiently used [3] (Ministry of Agriculture, Livestock, Rural Development, Fisheries, and Food [SAGARPA] 2012). Its location generates the rural character since they are towns generally far from urbanity, and in these contexts, agricultural activity is seen as a preponderant activity [4, 5].

The family business is one of the most essential structures in Mexico because most organizations are created under this form of business [6]. The family business lives in a knowledge-based society [7] that has given way to the new economy, in which powerful forces such as global competition have developed [8]. The family business has the challenge of recognizing and taking advantage of its resources, both tangible and intangible, which lies in the aptitudes, skills, and talents of its human resources, partners, and other stakeholders, developing intellectual property, which will allow negotiation, anticipation, adaptation, proactivity, and flexibility [9].

For family businesses, the era of globalization has generated a more competitive environment in which the development of companies is based on knowledge. This intangible resource creates capital of transcendence, such as human and social capital, among others, so management models are now directed towards intangible resources and sustainable development [10]. In the case of rural family businesses, they are distinguished by having intangible human and social resources that generate knowledge, experience, and intensity in solidarity collaboration, which allows creative learning that generates a socio-emotional wealth that creates a strong cohesion with stakeholders [11, 12].

Management based on the social solidarity economy has experienced accelerated growth in the last decade. Characterized by cooperation and mutual aid, it is recognized as a crucial element for sustainable and competitive development in the global era. This is especially relevant when considering the fundamental role of intellectual intangibles [13–15]. Today, knowledge-based societies demand the effective management of tangible and intangible resources, focusing on human resources' aptitudes, skills, and talents. In this context, intellectual property becomes a strategic asset for organizations' anticipation, adaptation, and flexibility. This perspective acquires significant value for rural family businesses, which possess intangible human and social resources [16–18].

Intense collaboration and creative to learning generate deep knowledge and unique expertise. These intangibles and the socio-emotional richness that drives cohesion with stakeholders position rural family businesses as critical agents for sustainable development [15]. Their ability to manage benefits equitably, contribute to ongoing member training, and promote the social and solidarity economy further strengthens their role in building sustainable communities in a competitive global environment [19]. Specifically, the number of companies that have opted for the Social Economy paradigm has been growing in the Latin American region.

However, it has been slowly consolidated [20]. Recently, in Mexico, there have been significant efforts to strengthen a Social Solidarity Economy perspective aligned with the Sustainable Development Goals, such as those carried out with the generation of the Nodes for the Promotion of the Social and Solidarity Economy (NODESS) by the Mexican Federal Government [19]. The NODESS is the strategy that aims to help the Promotion of the Social Economy by generating the integration of a network of territorial alliances made up of at least three different actors: academic institutions, local governments, and Organizations of the Social Sector of the Economy (OSSE). The NODESS aims to develop social and solidarity economy ecosystems in their territories, through which territorial solutions to collective needs are proposed, designed, and implemented [21].

The literature has also shown the epistemological challenge in addressing the diffuse nature of the variables of the collaborative economy, intangibles, and sustainability within companies [22]. For example, Negash et al. [23], using a fuzzy Delphi method, analyzed the driving role of collaboration and innovation in values on the economic sustainability of suppliers. Kayikci et al. [24] explored the role of integration and collaboration between partners in promoting an innovative and sustainable circular economy in an automotive eco-cluster using the DEMATEL fuzzy method.

Other researchers used the fuzzy hierarchical technique, such as Shete et al. [25], who, from a Pythagorean AHP analysis, found that stakeholder collaboration and engagement, I+D, as well as linkage with university and government, are essential critical enablers to achieve sustainability in Indian manufacturing industries. In Vietnam, Tran et al. [26], also through an AHP analysis, found in their hierarchical analysis that collaboration between companies has a significant weight in the sustainable transport management of shipping companies. In the context of the rural social economy in agri-food cooperatives, Mozas-Moral et al. [27] and Antonio Parrilla González and Ortega Alonso [28], who used the methodological technique of Qualitative Comparative Analysis of Fuzzy Sets (fsQCA), showed how aspects of sustainability are favored by cooperative integration in Spain.

In response to the complexity of social phenomena and their mathematical and non-parametric dichotomy, some research has innovated by exploring state-of-the-art hybrid techniques using a mixed approach of Structural Equation Modeling (SEM) and fuzzy mathematics. Muñoz-Pascual et al. [29], using an SEM hybridization and a Comparative Fuzzy Ensemble Analysis fsQCA, explored the predictive relationship between knowledge management and sustainability-oriented performance in Human Resources in Portugal. In China, using the same mix of Structural Equation Modeling (SEM) using Partial Least Squares (PLS) and fsQCA approach, Wang et

al. [30] showed the mediating role of intellectual capital in predicting the performance of sustainability-oriented innovation. On the other hand, in the context of the Collaborative Economy, [31] showed with a hybrid SEM-fsQCA technique the predictive power of the cooperative culture on sustainability performance in innovation governance.

Barney's [32] Resource-Based Vision (RBV) theory offers a solid conceptual framework for understanding how intangible intellectual property and the culture of solidarity in the context of rural family firms can influence sustainable business performance. According to Barney, RBV suggests companies can achieve competitive advantages by effectively leveraging their internal resources and capabilities. In this case, knowledge management, as it is based on the identification and exploitation of intangible assets such as solidarity cooperation, can constitute a crucial source of sustainable performance in rural family businesses in the agricultural sector [26, 31, 33].

Within the context above, this research aims to analyze the predictive capacity of the features of the solidarity economy and the culture of intellectual property on sustainability in rural family businesses in Mexico's agricultural sector. In order to provide clarity to social variables, typically diffuse, this work uses a hybrid technique that has rarely been used until now. We explore a hybrid empirical-mathematical methodology based on a non-parametric SEM-PLS approach and a Fuzzy Inference System (FIS) based on a Genetic Algorithm (GA). The data were collected from a survey of 88 rural family businesses. The following sections present the methodological processes, results, conclusions, and social implications.

13.2 Methodology

This study is based on the epistemology proposed by Wacker [34], which advocates the complementarity and necessity of different methodologies to address the different dimensions of a phenomenon under study. The research uses three main approaches: empirical, statistical-sampling, and mathematical analytical.

The hybrid methodology combining SEM-PLS and Genetic Algorithms offers a powerful analytical tool that captures the complexity and uncertainty present in social phenomena such as sustainability in rural family businesses. This methodological combination is beneficial for several reasons: (i) First, SEM is widely recognized for its ability to model complex relationships between latent and observable variables, allowing the underlying structure of data to be explored and the validity of a theoretical model to be assessed without limiting itself to parametric issues such as data normality [35]. (ii) On the other hand, fuzzy logic is especially useful for handling inaccuracy and vagueness in data, which is common in social and economic contexts [36].

Combining these two techniques, our hybrid methodology leverages the best of both approaches [37, 38]. SEM allows us to establish causal relationships between variables and evaluate the relative importance of each by standardizing anomalous data. At the same time, fuzzy logic helps us manage uncertainty and capture the com-

plexity inherent in social interactions. In the context of our research on sustainability in rural family firms, this hybrid methodology provides a more detailed comparison of these two methodological techniques. While the SEM-PLS allows us to identify causal relationships between the solidarity economy, intellectual property, and sustainability, fuzzy logic helps us better understand these concepts' vague and fuzzy nature in the specific context of rural family businesses. This allows us to perform more sophisticated analyses and draw more robust conclusions about these critical social phenomena.

13.2.1 Data Collection and Instrument

A cross-sectional approach was adopted, collecting data at a specific time to analyze the relationships proposed in the theoretical model. The sample consisted of 88 surveys aimed at multi-branch rural family businesses in the southern area of Tamaulipas. This choice was justified considering the relevance of family businesses in rural areas and their essential contribution to the regional economy. The Likert-type questionnaires (0–4) were collected in person, thus guaranteeing the response of the interested party. The instrument was constructed through a literature review and a Delphi methodological approach in which representatives of rural family businesses participated.

A measurement instrument was developed with 38 items divided into three dimensions: Traits of a supportive family business [6, 10, 14, 21], Intellectual Property [39–41], Sustainable Development [33, 40, 42]. Initial individual and construct reliability and validity tests were performed to ensure the robustness of the instrument, using statistical analyses such as Cronbach's alpha, rhoA, composite reliability, and convergent and discriminant validity. Barrios Del Angel et al. [33] validated these measures in the Latin context, whose construct measures had a satisfactory Cronbach's alpha between 0.822 and 0.892.

13.2.2 SEM-PLS/FIS-GA Hybrid Technology

This work implemented a rarely used hybrid combination of two robust techniques focused on prediction: Partial Least Squares Structural Equation Modeling (SEM-PLS) and a Fuzzy Inference System (FIS) optimized by a genetic algorithm (GA). Although this hybrid approach is relatively uncommon, it is valuable and supportive, especially in handling the complexity of non-parametric, fuzzy, and hard-to-delineate empirical data, as is often the case with the social aspects of sustainability. Previous research has supported this hybrid approach [37, 38], highlighting its effectiveness in dealing with the complexity inherent in social data in the context of sustainability. This hybrid approach offers the ability to effectively handle often challenging data

in terms of its non-parametric and fuzzy nature, thus offering a valuable contribution to the field of study.

The non-parametric statistical technique was performed with the SmartPLS v. 4.1.0.0 program [43] and was based on Structural Equation Modeling (SEM) using Partial Least Squares (PLS). SEM-PLS is helpful for social science modeling that aims to explore multiple relationships with a non-parametric predictive approach [35]. The condition to consider a PLS-SEM measurement model valid must take into account six criteria: (1) The reliability of the indicators ($\lambda \geq 0.400$), (2) The internal consistency of the construct (Cronbachs alpha/ CR ≥ 0.700), (3) The reliability of the construct ($\rho_A \geq 0.708$), (4) The convergent validity of the construct ($\rho_A \geq 0.500$), (5) The discriminant validity between constructs (HTMT $\beta > 0.900$); a structural model with a predictive capacity relevant to the social sciences would also require (6) an explained variance of $R^2\beta > 0.100$ and (7) with standardized path coefficients of $\beta > [35, 44]$.

For the mathematical parametric approach, we implemented a fuzzy inference system from Mamdani [36] to predict Sustainable Development with our available database. The linguistic granularity is low, medium, and high, and the triangular shapes of the sets are fuzzy, which were also used in the Latin context by Reyna-Castillo et al. [37]. The idea is to discover a knowledge base that minimizes the error between the FIS's clear output and the training example's numerical values. After analyzing the relevant characteristics, we took the inputs related to the Solidary Family Business and the attributes of the Culture of Intellectual Property and, as a dependent (consequential) value, Sustainable Development in family businesses in the Mexican agricultural sector.

The values of inputs and outputs are normalized $\in [0, 1]$. The ruleset is an AND ruleset with two inputs. The number of repeating combinations equals k^n , where k is the number of entries and n is the number of granularity levels. The above calculation provides 9 DNA rules to be discovered by a genetic algorithm. The dataset was divided into training and validation sets at a respective percentage of 60 and 40%.

13.2.3 Research Design

In the methodology of this work, six stages were carried out: (i) A database obtained through a survey applied in the second half of 2023 to rural family entrepreneurs in the Mexican agricultural sector who operate under a Social Solidarity Economy model was processed and curated. (ii) According to the literature, the collected data were unified and validated, transforming them into first-order constructs by applying a Partial Least Squares Structural Equation Modeling (PLS-SEM) method using SmartPLS v. 4.1.0.0. (iii) The size and non-parametric significance of the structural measurement model was evaluated by obtaining the effects of the relationships. (iv) The latent values of the variables were obtained, and integers were used to represent the nominal values. (v) From the values of latent variables, a genetic algorithm (GA) was used to refine the fuzzy results of the Fuzzy Inference System (FIS), eliminat-

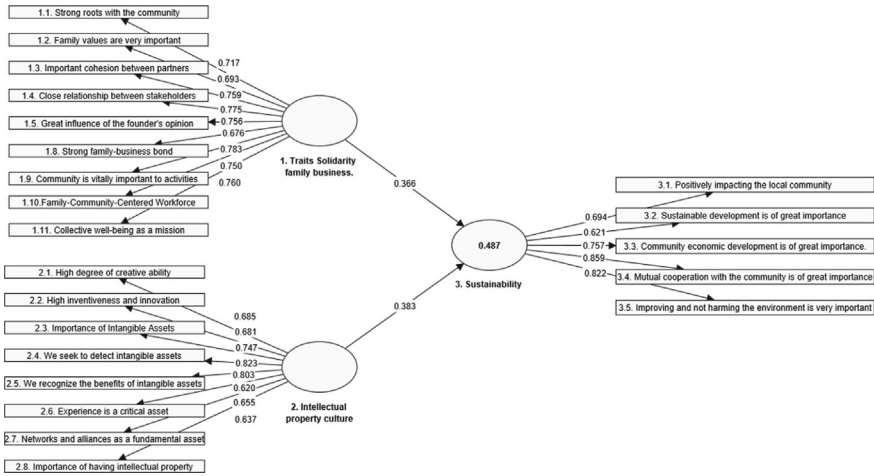


Fig. 13.1 Measures and constructs used in the study. Elaborated by the authors

ing the need for specialized knowledge. To assess the FIS of diffuse evolutionary attributes, sustainable development is predicted. (vi) Finally, the error of the exit FIS was assessed.

13.3 Results

13.3.1 Validation of the Measurement Model

Using the statistical program SmartPLS v. 4.1.0.0 [43], the values of the measurement model were obtained where external loads with the expected values $\lambda \geq 0.400$ were obtained. As shown in Fig. 13.1, all the external loads of the items are favorable. As for the power of the predictive capacity, it was also relevant, on the one hand, with an explained variance of R^2 0.487, exceeding the required threshold of $R^2 > 0.100$ and with path standardized coefficients of 0.366 and 0.383, according to the required parameter of $\beta > 0.200$ [35].

As for the quality criteria of the constructs of the measurement model, as shown in Table 13.1, the internal consistency of the construct ranges from a Cronbachs alpha of = 0.807–0.898 to a CR between 0.868 and 0.917, satisfactory values of Cronbachs Alpha/CR ≥ 0.700 . On the other hand, the reliability of the constructs presented values rhoA between 0.818 and 0.901. Likewise, the convergent validity of the construct (rhoA ≥ 0.500). Finally, discriminant validity between constructs met heterotrait-monotrait test (HTMT) criteria > 0.900 [45].

Table 13.1 Individual reliability and construct validity

Construct	Cronbachs Alpha	CR	rhoA	AVE	H	TM	T
					1	2	3
1. Traits solidarity family business	0.898	0.917	0.901	0.550	–	–	–
2. Intellectual property culture	0.857	0.889	0.864	0.504	0.838	–	–
3. Sustainability	0.807	0.868	0.818	0.571	0.745	0.765	–

Elaborated by the authors

Table 13.2 Latent score by observation (SEM-PLS Algorithm)

Observation	1. Traits solidarity family business	2. Intellectual property culture	3. Sustainability
1	0.534	1.466	0.408
2	1.096	1.097	– 0.29
3	1.096	0.852	1.249
4	0.044	– 1.235	– 1.24
5	0.508	0.971	1.249
6	0.856	0.618	0.783
7	0.763	1.466	1.249
8	0.811	1.466	– 0.335
9	– 3.581	– 2.845	– 1.869
10	– 1.383	– 1.766	0.388
11	– 1.624	– 1.046	– 1.563
12	– 0.642	– 1.492	– 0.91
13	0.83	0.222	– 0.031
14	0.616	1.111	1.249
15	0.472	– 0.368	– 0.226
16	1.096	1.466	0.24
17	– 0.132	0.187	0.783
18	– 0.299	– 0.25	– 0.657
19	– 2.024	– 1.45	0.388
20	– 0.492	– 0.087	0.51
(...) 88	0.657	0.377	0.993

Elaborated by the authors

The SEM-PLS Algorithm allows, in its first stage, the extraction of standardized information from the data by generating latent scores [46]. Table 13.2 represents the latent scores obtained for each of the 88 observations.

Following the research of Reyna-Castillo et al. [37], who explored the usefulness of the latent variables generated by the SEM-PLS Algorithm as a preliminary step

Table 13.3 The 9 rules of the fuzzy knowledge base

Rules	1. Traits solidarity family business	2. Intellectual property culture	3. Sustainability
1	LOW	LOW	MID
2	LOW	MID	LOW
3	LOW	HIGH	HIGH
4	MID	LOW	HIGH
5	MID	MID	MID
6	MID	HIGH	MID
7	HIGH	LOW	LOW
8	HIGH	MID	HIGH
9	HIGH	HIGH	HIGH

Fitness: 0.31553716772037016

Elaborated by the authors

to the use of fuzzy genetic algorithms, the estimated latent values were taken for the next phase of the rules for FIS calculation.

13.3.2 The Fuzzy Inference System

It is essential to remember that Genetic Algorithms (GA) are stochastic, so the fuzzy inference system (FIS) performance is intended to be determined. Since FIS is deterministic, we can calculate the exact value of the relative error produced by the best fuzzy knowledge base in the training phase against the test data with the antecedent and consequential data. The best fuzzy rules found by the AG are listed in Table 13.3. Using the rules in Table 13.3, we calculated an absolute error of 12.97, equivalent to about 13 misclassified examples out of 81, with a classification accuracy of 68.44%.

However, this absolute error could be the sum of slight differences between values, and for some decision-makers, these differences would be negligible. Therefore, the predictive power in Sustainable Development is close to 70%, based on aspects related to solidarity characteristics and the culture of intellectual property in rural family businesses. Our proposed FIS is a powerful tool for decision-makers and policy-makers in organizations.

13.4 Discussion

The question implicit in the objective of this research was: Is there a predictive capacity of the features of the solidarity economy and the culture of intellectual property on Sustainability in rural family businesses in the agricultural sector in

Mexico? To answer this question, we used a methodology based on a hybrid predictive approach that combined Partial Least Squares Structural Equation Modeling (SEM-PLS) and a Fuzzy Inference System (FIS) of genetic algorithms (GA). Initially, the results of our measurement model, based on the SEM-PLS Algorithm, allowed us to validate and curate our measurements and allowed us to approach the predictive potential of the relationships explored, as well as to obtain latent variables with viable standardized values for predictive verification based on the FIS-GA technique.

In general, regarding the predictive approach based on SEM-PLS, the results show a relevant predictive potential of the model with an explained variance of $R^2 \geq 0.487$. This result implies that, within the sample analyzed, the indicators of the variables of Aspects of the Solidarity Family Business and the Culture of Intellectual Property explain Sustainability by almost 50%, with standardized coefficients path $\beta = 0.366$ and 0.383 . On the other hand, when determining the FIS-GA performance and calculating the exact value of the relative error produced by the best fuzzy knowledge base, a classification accuracy of 68.44% was obtained. These results mathematically confirm the non-parametric predictive estimates obtained, allowing us to affirm that the collaborative features of rural family enterprises and their orientation to the participation of intellectual property rights have predictive power on their sustainable performance of triple social, environmental, and economic results.

From the theoretical perspective of the Resource-Based Vision (RBV) [32], our results imply that, as a whole and within the representative cases of the sample, the values of solidarity collaboration, the characteristics of rural family firms, as well as the culture of intellectual property rights are valuable resources that predict sustained performance in the agricultural sector. This theoretical support was consistent with the previous work of Barrios-DelÁngel et al. [33], where the theoretical hypothesis of the VBR confirmed the predictive relationship between intangible assets and sustainable development in rural family firms. It is also linked to the theoretical support of Sun et al. [31] who, under the Resource-Based theory, evidenced how Collaboration between companies is recognized as a valuable resource for the sustainable management of maritime transport in shipping companies in Vietnam. Regarding the specific results, the correlations obtained using the SEM-PLS Algorithm show item-based relationships with the dependent variable of Sustainability.

Figure 13.2 shows the correlation weights of the aspects of the variable “Traits of a solitary family business.” The item most correlated with Sustainability was “1.4. Close relationship between stakeholders” ($p = 0.591$). The results are consistent with the work of Shete et al. [25] in manufacturing companies in India and with Tran et al. [26] in shipping companies in Vietnam. They found that “The collaboration and commitment of a diverse group of stakeholders (such as suppliers, local people, customers, environmental specialists, NGOs)” is one of the most critical enablers for achieving sustainable nailing. Figure 13.2 also shows that the second item of the variable “Traits of a solitary family business” with the highest correlation with Sustainability was “1.9. The community is of vital importance to the activities” ($p = 0.511$). The results are consistent with Sun et al. [31], who highlighted cooperative culture as a fundamental factor for sustainable development performance in innova-

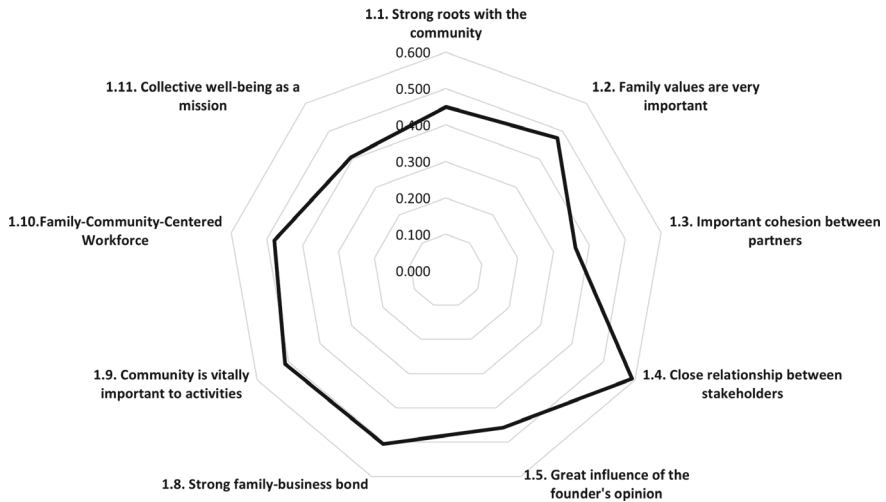


Fig. 13.2 PLS correlations: traits of supportive family business versus sustainability elaborated by the authors

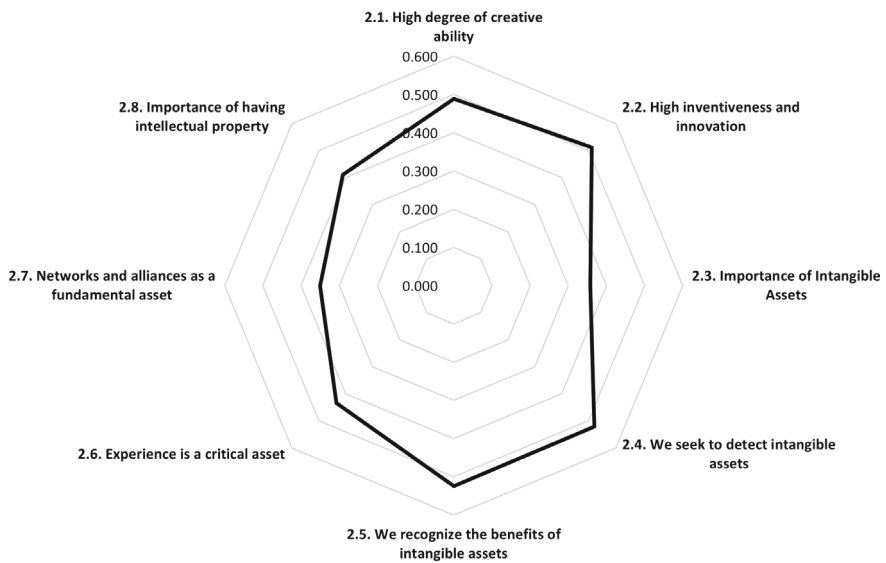


Fig. 13.3 PLS correlations: intellectual property culture versus sustainability. Elaborated by the authors

tion governance. On the other hand, Mozas-Moral et al. [27] found that the intensity of cooperative integration favors the degree of sustainable innovation.

On the other hand, regarding the specific results of the correlations of the dependent variable of Sustainability, Fig. 13.3 shows the correlation weights of the aspects

of the variable “Culture of intellectual property.” The item most correlated with Sustainability was “2.5. We recognize the benefits of intangible assets” ($p = 0.524$), and the second item with the highest correlation weight was “2.4. We seek to detect intangible assets” ($p = 0.521$). Within the studies analyzed, the role of intangibles was linked to the performance of triple-bottom Sustainability within companies [23, 26]. Business innovation was the backbone of the Sustainable Development Goals (SDGs). However, it has been shown that the intensity of its performance also depends on intangible values, such as the driving role of collaboration and innovation in values within the organization [29].

13.5 Conclusions

Our study focused on analyzing the predictive capacity of the characteristics of the solidarity economy and the culture of intellectual property on sustainability in rural family businesses in the agricultural sector in Mexico. We use an empirical-mathematical methodology, combining a hybrid approach of structural equations and fuzzy evolutionary logic from a survey of 88 rural family firms. The results reveal the model’s predictive capacity with a relevant explained variance. This implies that the characteristics of the solidarity economy and the culture of intellectual property explain a significant percentage of the sustainability of rural family businesses in the agricultural sector in Mexico. In addition, standardized route coefficients indicate a considerable influence of these variables on sustainability.

From the theoretical perspective of the Resource-Based Vision (RBV), our results support the idea that collaborative solidarity and a focus on intellectual property are valuable resources that predict sustained performance in the agricultural sector. These findings align with previous research highlighting the importance of these aspects in promoting business sustainability. In addition, a hybrid technique that combines partial structural equations and fuzzy evolutionary logic demonstrates its compatibility, relevance, and usefulness in analyzing complex phenomena such as sustainability in rural family firms. This approach makes capturing the complexity and uncertainty inherent in this context possible, thus offering a more complete and accurate view of the relationships between the variables studied.

In conclusion, our research underscores the predictive capacity of the characteristics of the solidarity economy and the culture of intellectual property on sustainability in the representative cases of rural family businesses in the agricultural sector in Mexico. This has significant implications for policymakers and supportive communities embarking on social enterprises, as they are urged to consider these characteristics as valuable resources for their long-term development, community collaboration, and stakeholder engagement. Rural family businesses in the agricultural sector in Mexico must not only sustain their business operations but also engage in continuous reflection to harness the intangible assets that enable them to sustain and differentiate themselves. Ensuring compliance with their intellectual property rights is a crucial aspect of long-term care for rural family businesses.

13.6 Limitations

It is important to acknowledge the limitations of our study. Firstly, the sample was confined to 88 rural family firms in a specific region of Mexico, which may limit the generalization of our results to other geographical areas or industrial sectors. Secondly, the cross-sectional nature of our research design precludes us from establishing causal relationships between the variables studied. Lastly, while we employ a combination of robust methodologies, such as partial structural equations and genetic algorithms, other techniques could potentially offer more comprehensive insights into the relationship between solidarity economy, intellectual property, and sustainability in rural family businesses.

Acknowledgements Funding is gratefully acknowledged to CONAHCYT under the Postdoctoral Fellowships for Mexico (2021-1) program with application number 2264959.

References

1. Eppinger, E., Jain, A., Vimalnath, P., Gurtoo, A., Tietze, F., Chea, R.H.: Sustainability transitions in manufacturing: the role of intellectual property. *Curr. Opin. Environ. Sustain.* **49**, 118–126, 4 (2021)
2. Siltaloppi, J., Ballardini, R.M.: Promoting systemic collaboration for sustainable innovation through intellectual property rights. *J. Coop. Organ. Manage.* **11**, 100200, 6 (2023)
3. SAGARPA Secretaria de Agricultura Ganadería Desarrollo Rural Pesca y Alimentación. *Agricultura familiar con potencial productivo en México* (2012)
4. Salgado, I.F., González, F.G.: La importancia de los factores socioculturales en la competitividad de la empresa rural. el caso de la empresa apícola miel tierra grande. *RICSH Revista Iberoamericana de las Ciencias Sociales y Humanísticas*, **9**, 70–93, 7 (2020)
5. Henry, C., McElwee, G.: Defining and conceptualising rural enterprise 8 (2014)
6. San-Martín, Durán, J.: *Radiografía de la empresa familiar en México* (2017)
7. Maditinos, D., Sevic, Z., Tsairidis, C.: Intellectual capital and business performance: an empirical study for the Greek listed companies. *Eur. Res. Stud. J.* **XIII**, 145–168, 11 (2010)
8. Bontis, N.: Intellectual capital: an exploratory study that develops measures and models. *Manage. Decis.* **36**, 63–76, 3 (1998)
9. OMC: Informe annual. organización mundial del comercio [omc] (2017)
10. Kim, D., Cho, W., Allen, B.: Sustainability of social economy organizations (SEOs): an analysis of the conditions for surviving and thriving. *Soc. Sci. J.*, pp. 1–17, 9 (2020)
11. Cesinger, B., Hughes, M., Mensching, H., Bouncken, R., Fredrich, V., Kraus, S.: A socio-emotional wealth perspective on how collaboration intensity, trust, and international market knowledge affect family firms' multinationality. *J. World Bus.* **51**, 586–599, 6 (2016)
12. Wanniarachchi, T., Dissanayake, D.G.K., Downs, C.: Community-based family enterprise and sustainable development in rural Sri Lanka. *Community Work Family*, pp. 1–19, 5 (2022)
13. Laghdas, M., García, E.C., Valverde, F.A.N.: Economía social y desarrollo territorial en cheffchaouen (marruecos): el papel de las cooperativas en el marco de la iniciativa nacional para el desarrollo humano. *Boletín de la Asociación de Geógrafos Españoles* **6** (2023)
14. Vélez, L.E.M.: Aportes conceptuales de la economía social y solidaria a la economía circular. *Cuadernos de Administración* **37**, e5010824, 8 (2021)

15. do Nascimento, F.S., Calle-Collado, Á., Benito, R.M.: Economía social y solidaria y agroecología en cooperativas de agricultura familiar en Brasil como forma de desarrollo de una agricultura sostenible. *CIRIEC-España, revista de economía pública, social y cooperativa*, p. 189, 4 (2020)
16. Chen, X., Huang, Q., Davison, R.M.: The role of website quality and social capital in building buyers loyalty. *Int. J. Inf. Manage.* **37**, 1563–1574, 2 (2017)
17. Lee, H.-T., Park, C.-S.: The role of social economy to revitalize rural-industrial complex in chungcheongnam-do. *Korean Rev. Corporation Manage.* **8**, 93–111, 12 (2017)
18. Virlanuta, F.O.: Comparative analysis state of development of the social economy in the EU and in Romania. *Procedia Econ. Finance* **23**, 335–340 (2015)
19. Martínez-Gutiérrez, R., Solís-Quinteros, M.M., Sánchez-Hurtado, C., Carey-Raygoza, C.E.: Challenges for an Observatory of the 2030 Goals, SDG and Social Economy, in Northern Mexico (2021)
20. Serrano-Serrato, L.V., Benavides, O.T.: Análisis del crecimiento del sector de economía solidaria en el área de agricultura y su aplicación a la formación. *PUBLICACIONES* **52**, 357–378, 1 (2022)
21. INAES: Nodos de impulso a la economía social y solidaria nodess, 1 (2023)
22. Cisneros, L., Rivera, G., Florencia, R., Sánchez-Solís, J.P.: Fuzzy optimisation for business analytics: a bibliometric analysis. *J. Intell. Fuzzy Syst.* **44**(2), 2615–2630 (2023)
23. Negash, Y.T., Hassan, A.M., Lim, M.K., Tseng, M.-L.: Sustainable supply chain finance enablers under disruption: the causal effect of collaboration value innovation on sustainability performance. *Int. J. Logist. Res. Appl.*, pp. 1–25, 1 (2024)
24. Kayikci, Y., Kazancoglu, Y., Lafci, C., Gozacan, N.: Exploring barriers to smart and sustainable circular economy: the case of an automotive eco-cluster. *J. Clean. Prod.* **314**, 127920, 9 (2021)
25. Shete, P.C., Ansari, Z.N., Kant, R.: A Pythagorean fuzzy AHP approach and its application to evaluate the enablers of sustainable supply chain innovation. *Sustain. Prod. Consumption* **23**, 77–93, 7 (2020)
26. Tran, T.M.T., Yuen, K.F., Li, K.X., Balci, G., Ma, F.: A theory-driven identification and ranking of the critical success factors of sustainable shipping management. *J. Clean. Prod.* **243**, 118401, 1 (2020)
27. Mozas-Moral, A., Bernal-Jurado, E., Fernández-Uclés, D., Medina-Viruel, M.: Innovation as the backbone of sustainable development goals. *Sustainability* **12**, 4747, 6 (2020)
28. Parrilla-González, Juan, Ortega-Alonso, Diego: Sustainable development goals in the Andalusian olive oil cooperative sector: heritage, innovation, gender perspective and sustainability. *New Medit.* **21**, 6 (2022)
29. Muñoz-Pascual, L., Galende, J., Curado, C.: Human resource management contributions to knowledge sharing for a sustainability-oriented performance: a mixed methods approach. *Sustainability* **12**, 161, 12 (2019)
30. Wang, N., Xie, W., Huang, Y., Ma, Z.: Big data capability and sustainability oriented innovation: the mediating role of intellectual capital. *Bus. Strategy Environ.* **32**, 5702–5720, 12 (2023)
31. Sun, Y., Wang, T., Gu, X.: A sustainable development perspective on cooperative culture, knowledge flow, and innovation network governance performance. *Sustainability* **11**, 6126, 11 (2019)
32. Barney, J.: Firm resources and sustained competitive advantage. *J. Manage.* **17**, 99–120, 3 (1991)
33. Barrios-DelÁngel, A.-X., Reyna-Castillo, M., Bucio-Gutiérrez, D.: Activos intangibles y la competitividad sostenible en las empresas familiares. *Revista de Ciencias Sociales* **28**, 94–109 (2022)
34. Wacker, J.G.: A definition of theory: research guidelines for different theory-building research methods in operations management. *J. Oper. Manage.* **16**, 7 (1998)
35. Hair, J.F., Risher, J.J., Sarstedt, M., Ringle, C.M.: When to use and how to report the results of PLS-SEM. *Eur. Bus. Rev.* **31**, 2–24, 1 (2019)
36. Cerdón, O.: A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: designing interpretable genetic fuzzy systems. *Int. J. Approximate Reasoning* **52**, 894–913, 9 (2011)

37. Reyna-Castillo, M., Santiago, A., MartÃnez, S.I., Rocha, J.A.C.: Social sustainability and resilience in supply chains of Latin America on COVID-19 times: classification using evolutionary fuzzy knowledge. *Mathematics* **10**, 2371, 7 (2022)
38. Ringle, C.M., Sarstedt, M., Schlittgen, R., Genetic algorithm segmentation in partial least squares structural equation modeling. *OR Spectrum* **36**, 251–276, 1 (2014)
39. Chang, H.H., Chuang, S.-S.: Social capital and individual motivations on knowledge sharing: participant involvement as a moderator. *Inf. Manage.* **48**, 9–18, 1 (2011)
40. Joo, J.-H.: A mediating role of social capital between corporate social responsibility and corporate reputation: perception of local university on CSR of KHNP. *J. Ind. Distrib. Bus.* **11**, 63–71, 3 (2020)
41. Tejedo Romero, F., de AraÃjo, J.F.F.E.: Human capital information: generating intangibles and social responsibility. *Cuadernos de gesti3n* **16**, 125–144 (2016)
42. Sallah, C.A., Caesar, L.D.: Intangible resources and the growth of women businesses. *J. Entrepreneurship Emerg. Econ.* **12**, 329–355, 1 (2020)
43. Ringle, C.M., Wende, S., Becker, J.-M.: Smartpls 4. <http://www.smartpls.com> (2022)
44. Hair Jr., J.F.: Next-generation prediction metrics for composite-based PLS-SEM. *Ind. Manage. Data Syst.*, ahead-of-print 10 (2020)
45. Henseler, Ringle, C.M., Sarstedt, M.: A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Market. Sci.* **43**, 115–135, 1 (2015)
46. Sarstedt, M., Hair, J.F., Cheah, J.-H., Becker, J.-M., Ringle, C.M. How to specify, estimate, and validate higher-order constructs in PLS-SEM. *Australas. Market. J.* **27**, 197–211, 8 (2019)

Chapter 14

Learning Analytics in Reading Comprehension



Maritza Bustos-López , Isaac Machorro-Cano , Giner Alor-Hernández , Jonathan Hernández-Capistran , and José Oscar Olmedo-Aguirre 

Abstract Emerging technologies enable the acquisition of substantial volumes of data produced through user engagement with Internet-connected learning platforms. Hence, it is feasible to analyze and convert such data into valuable insights to maximize the effectiveness of any instruction, be it for children, adolescents, or adults. Using Learning Analytics in this context enables the derivation of conclusions from the data, the identification of new variables that may assist an academic institution in responding more effectively to the various situations that students encounter, and the formulation of decisions based on the information analysis. By leveraging Learning Analytics, numerous benefits can be realized, including but not limited to obtaining vital student data, generating predictions, reducing attrition, increasing revenue, and enhancing course offerings. This chapter of the book explores the utilization of Learning Analytics in the context of reading comprehension to quantify various learning-related behaviors. The primary graphs utilized in Learning Analytics are delineated, accompanied by a case study on reading comprehension.

Keywords Data analytics · Data visualization · Education · Learning analytics · Reading comprehension

M. Bustos-López · G. Alor-Hernández (✉) · J. Hernández-Capistran
Tecnológico Nacional de México/I. T. Orizaba, Orizaba, Veracruz, México
e-mail: giner.ah@orizaba.tecnm.mx

M. Bustos-López
e-mail: maritbustos@gmail.com

J. Hernández-Capistran
e-mail: jonathan.hc@orizaba.tecnm.mx

I. Machorro-Cano
Universidad del Papaloapan, San Juan Bautista Tuxtepec, Oaxaca, México
e-mail: imachorro@unpa.edu.mx

J. O. Olmedo-Aguirre
Escuela Superior de Física y Matemáticas del IPN, Ciudad de México, México
e-mail: jolmedoa@ipn.mx

14.1 Introduction

Education is a fundamental right for every human being that enables a person to participate fully in society. The right to quality education throughout life ensures that a person is socially and economically integrated. However, UNESCO data show that 244 million children and youth worldwide are out of school for social, economic, or cultural reasons, and an estimated 771 million youth and adults lack basic literacy skills, two-thirds of whom are women [1].

Writing, reading, reading comprehension, and reading competency are essential skills that all educational systems strive for and are used in every context that individuals encounter. Children and young individuals face reading difficulties due to inadequate reading skills and limited comprehension and information-processing abilities. These challenges negatively affect their academic performance, the acquisition of meaningful knowledge, and overall school achievements. In the 2018 Program for International Student Assessment (PISA) results, which reflect the pre-pandemic examination procedure, all Latin American nations scored below average in the PISA Reading Comprehension test. The lowest average score recorded was 487 points [2]. Based on the above, lacking skills prevents a person from participating effectively and productively in a modern, cosmopolitan society.

Although AI is currently present in many social contexts, it has grown, developed, and implemented several utilities in the educational sector to provide solutions to the education system. Technologies such as Big Data, Learning Analytics, machine learning algorithms, deep learning, computer vision, natural language processing, and neural networks are integrated within the educational context. Specifically, learning analytics generates a large volume of information that contributes to identifying findings within the existing information data and new variables that rethink pedagogical strategies to face situations particular to the teaching–learning process. Learning Analytics is an area of research that uses computational data analysis from the learning process to understand and improve learning. Learning Analytics uses data about students and their activities in various contexts to help understand and improve educational management processes and, crucially, to improve students' learning.

Learning Analytics tools enable the collection, analysis, and dissemination of data generated in virtual environments, including information such as exam results (both correct and incorrect answers), frequency of content access, time spent on specific pages, and utilization of supplementary video materials, among other variables. The obtained data undergoes analysis and transformation into information, which aids in identifying both collective and individual behaviors, thus enhancing the teaching process. The primary goal is to establish a database that aids management in assessing performance and making choices about all educational activities conducted.

In recent times, there has been a growing inclination among students toward online education and the use of electronic devices such as computers, tablets, and smartphones. This trend creates a “digital footprint,” which may be automatically examined and merged with information on their history, previous academic achievements, or

job goals. Learning Analytics may be used to evaluate information and customize educational material, activities, or procedures to provide a more customized and improved learning experience for the student. This approach can potentially be more advantageous for students than the conventional educational strategy widely adopted in higher education.

Moreover, educational institutions may get advantages from using Learning Analytics, including (1) identifying students who are in danger of discontinuing their studies, (2) providing course recommendations tailored to individual students, and (3) facilitating personalized learning experiences. This chapter of the book provides a comprehensive examination of data analytics in the education sector, with a specific emphasis on a case study that investigates the reading comprehension of university students. This case study aims to demonstrate the advantages and benefits of using data analytics in decision-making processes within higher education.

The structure of this chapter is as follows: Sect. 14.2 provides a comprehensive overview of relevant research; Sect. 14.3 outlines the primary graphs used in Learning Analytics; Sect. 14.4 describes the APIs for Learning Analytics; Sect. 14.5 includes a case study on the reading comprehension abilities of university students; and lastly, Sect. 14.6 delivers the findings.

14.2 Related Works

This section provides a comprehensive survey of the existing literature on Learning Analytics. The articles were categorized into four distinct aspects: (1) Learning Analytics is a research field that uses computational data analysis in the learning process to improve and optimize learning environments [3]; (2) Social Learning Analytics is a subset of Learning Analytics that focuses on understanding how social and cultural factors contribute to students' knowledge gain [4]; (3) Mobile Learning Analytics provides insights and alternatives to enhance students' mobile learning experiences, particularly outside the classroom [5]; (4) Big Data and Learning Analytics focuses on collecting and analyzing a large amount of information from various sources such as online interactions, user logs, and social networks [6].

14.2.1 *Learning Analytics*

One of the primary goals of Learning Analytics (LA) is to provide students and instructors with feedback on their acquired knowledge, optimize teaching and learning behavior, and foster the growth of expertise in the field to gain a deeper understanding of education. Leitner and Ebner [7] examined the design and development of a dashboard for LA in Higher Education. The dashboard analysis delineated three primary objectives: (1) to satisfy the requirements of various stakeholders; (2) to optimize the potential for integration and transferability to alternative contexts;

and (3) to ensure universal accessibility through open-source development. A separate investigation by Jivet et al. [8] employed a mixed-methods research approach to create Learning Analytics Dashboards (LAD). This methodology entailed an initial qualitative pre-study and a comprehensive quantitative study involving 247 university students. The methodology was implemented with a self-regulated learning approach, various immediate learning feedback strategies, and the student's objectives to accommodate college students' varying abilities and requirements.

The examination and assessment of literature about LA research in the educational setting involves five primary steps, as outlined by Campbell et al. [9]. These steps encompass capture, report, predict, act, and refine. Institutional-level decisions are pivotal in determining key performance indicators for student success, forming the foundation for operational reports used in decision-making processes. Practical data analysis in LA is crucial for achieving objectives. In a systematic review by Nunn et al. [10], various methods were identified, including visual data analysis, social network analysis, and semantic and educational data mining. These methods involve prediction, clustering, relationship mining, discovery with models, and data separation for human judgment when analyzing data.

Mattingly et al. [11] analyzed the role of learning and academic analytics in distance education, focusing on undergraduate and graduate programs. Their objective was to explore the various variables within academic analytics, aiming to predict student success by examining the nature of students' learning experiences and the support provided by academic programs and institutions. The study delved into evaluating the effectiveness of current student support mechanisms, identifying improvement areas, and offering educators insights. Furthermore, the researchers discussed leveraging this data to develop new metrics and facilitate a continuous improvement cycle.

Conversely, El-Alfy et al. [12] introduced a typology associated with educational domains, categorizing dispersed endeavors that examine the advantages and obstacles of Learning Analytics in areas such as management, teaching and learning processes, and related research. Through their literature review on learning analytics, the authors discovered that 50% of the research comprises empirical studies, 30% consists of conceptual studies, and only 20% constitutes systematic literature reviews. Despite their significance in the educational context, this distribution highlights a restrained focus on systematic literature reviews.

In the ongoing examination of LA within higher education, Leitner et al. [13] employed mixed methods to identify information sources, utilizing libraries such as the Learning Analytics and Knowledge (LAK) conference, SpringerLink, and the Web of Science (WOS) databases. The analysis yielded insights into (1) diverse techniques employed in studies and projects related to Learning Analytics, (2) emerging trends in various projects, and (3) discussions addressing the limitations, future directions, and challenges of these studies. Similarly, Wong [14] conducted a comprehensive review of relevant case studies sourced from Scopus spanning 2007–2016, revealing (1) the benefits of Learning Analytics for effective decision-making in institutions, (2) its role in evaluating pedagogies and instructional designs to enhance and monitor student learning, (3) its predictive capacity for student performance, (4)

ability to identify undesirable learning behaviors and emotional states, (5) capability to recognize at-risk students, enabling timely intervention and support, and (6) provision of insightful data to enhance the personalization and engagement of student's learning experiences, fostering reflection and improvement.

Zhang et al. [15] conducted a systematic analysis utilizing bibliometric and visualization methods to elucidate the evolution of Learning Analytics (LA) in higher education. The study amalgamated the literature review with analysis tools to scrutinize the trajectory of the LA field. Employing bibliometric analysis, the research delineated the developmental stages of the primary methods in LA. The study categorizes four principal methods: (1) social Learning Analytics, (2) content analysis, (3) dispositional analysis, and (4) discourse analysis. Examining social Learning Analytics, Díaz-Lázaro et al. [16] conducted a study within the Primary Education program at the University of Murcia. The focus was on understanding how students learn and collaborate in online environments, specifically through social networks like Weblog, Twitter, and Facebook. The research methodology included (1) participant observation, (2) quantitative analysis primarily derived from Learning Analytics data, and (3) qualitative analysis involving the scrutiny of student content from comments.

Viberg et al. [17] conducted a literature review examining 252 articles on LA in higher education (HE) published between 2012 and 2018. The analysis considered four key propositions: (i) the impact of Learning Analytics on learning outcomes, (ii) their role in supporting learning and teaching, (iii) the extent of their implementation, and (iv) their ethical usage. The study also delved into the research methodologies employed and the evidence provided by LA research in the context of learning in HE. The findings revealed that the field of LA is dynamic, characterized by a prevalence of descriptive studies and interpretive methods for data collection, primarily conducted online. Notably, there has been a discernible shift towards gaining a deeper understanding of students' learning experiences in recent years.

Tsai and Gasevic [18] conducted a literature review focusing on the policies relevant to the implementation and challenges of LA in higher education. The study highlighted key findings, emphasizing the necessity to (1) enhance communication channels among stakeholders and adopt pedagogy-based approaches for LA, (2) address the lack of guidance in developing data literacy among end users and assessing the progress and impact of Learning Analytics, and (3) establish formalized guidelines for monitoring the robustness, effectiveness, and legitimacy of Learning Analytics. Additionally, Cerratto Pargman and McGrath [19] systematically reviewed the research literature on ethical issues within the context of LA in higher education. The review outlined three inclusion criteria: (1) characterization of empirical research on LA ethics, (2) identification of major ethical areas addressed in the literature, and (3) identification of knowledge gaps. The authors concluded that the existing empirical work on LA ethics is limited, and further studies are necessary as Learning Analytics systems continue to evolve, requiring a comprehensive understanding through future investigations involving various institutional constituents.

Adejo and Connolly [20] highlighted that LA is an emerging field that proposes integrating data mining techniques and student data for informed decision-making, aiming to enhance the student learning experience and study environment. They emphasized using the Technology-Organisation-Human framework for determining a successful LA implementation pathway in higher education institutions. In a separate study, Kuhnel et al. [21] introduced the MyLA (My Learning Analytics) application to assess its usability among potential users. MyLA gathered learning behavior and personality traits data, combining Learning Analytics with mobile learning (m-learning) and incorporating personalized learning elements. This approach allowed learners to monitor their progress over time.

14.2.2 Social Learning Analytics

Social Learning Analytics (SLA) is a specific branch of Learning Analytics that reveals emerging skills and ideas developed and transmitted through interactions and collaborations within the educational process. It emphasizes the importance of identifying learning within its contextual framework. Khousa et al. [22] introduced a model integrating professional preparation into higher education, creating a novel learning approach based on Communities of Practice (CoP) through Learning Analytics and social computing techniques. The primary objective is to measure, instill, and track the development of professional competencies in higher education students. The proposed model comprises three key modules: (1) career preparation, (2) career prediction, and (3) career development. The study employed a semi-supervised aggregation method, specifically the Fuzzy Pairwise-Constraints K-Means (FCKM) algorithm. Addressing the practicality of SLA in the learning process, Hernandez-Garcia et al. [23] conducted a study exploring the correlation between social network analytics parameters and student performance outcomes and the relationship between social network parameters and overall course performance. Their research demonstrates how SLA visualizations can facilitate observing visible and invisible interactions in online distance education, aiding decision-making, and predicting academic performance.

In another study, Manca et al. [24] identified the possibilities and educational challenges of applying Big Data techniques to massive online learning platforms and social networks. The research underscored vital aspects, including identifying tools and methodological instruments for SLA, addressing ethical concerns, and devising mechanisms to ensure user privacy and security within online learning environments. Similarly, Hernández-García and Conde-González [25] introduced an approach involving integrating three systems—Moodle, GraphFES, and Gephi. The objective was to showcase the potential of employing social network analysis methods and visualizations in computer-assisted collaborative learning environments, aiming to enhance and expand the scope of this field of study.

SLA harnesses the potential of utilizing data from students' activity logs in online settings to discern behaviors indicative of their performance in learning environments. Considering this, Doleck et al. [26] assessed an algorithm's validity, aiming to gauge the effectiveness of social learning networks within discussion forums accompanying conventional Massive Open Online Courses (MOOCs). The study sought to optimize online learning environments, enhancing social learning and deepening insights into indirect online learning. This study considered forum users who engage in knowledge-seeking behaviors like reading and searching, even if they do not actively share knowledge. Aguilar et al. [27] integrated SLA tasks into autonomous cycles in a smart classroom to identify students' learning styles. They utilized a course for analyzing external data from the web, particularly from social networks like Twitter, to construct knowledge models about students, ultimately employing Semantic Mining, Text Mining, and Data Mining techniques to develop SLA tasks to enhance learning processes.

A growing number of scholars are directing their attention to SLA as an emerging trend in education. Systematic literature reviews serve as platforms to present best practices in this field. Kaliisa et al. [28] conducted an extensive systematic review, analyzing 36 SLA-related studies from 2011 to 2020. Their focus encompassed methodological characteristics, educational approaches, and theoretical perspectives within the studies. The results underscored the prevalence of SLA in formal and fully online settings, with social network analysis being the predominant analytical technique. Notably, most SLA studies aimed to comprehend students' learning processes, adopting a social constructivist perspective to interpret learning behaviors. However, specific gaps were identified: (i) limited teacher involvement in developing SLA tools and infrequent sharing of SLA visualizations for teaching support, (ii) some SLA studies needing more theoretical frameworks, and (iii) a restricted number of studies integrating multiple analytic approaches. Additionally, (iv) few studies explored innovative network approaches, and (v) temporal patterns of student interactions were infrequently studied to understand the evolution of social and knowledge networks over time. Moreover, SLAs were recognized as a relatively recent extension of LA, finding applications in computer-supported collaborative learning environments. Rienties and Toetenel [29] presented a study derived from preliminary work on learning design, emphasizing its significance in predicting and comprehending learner performance in blended and online settings. The study aimed to unravel the intricate connections between learning design, learning processes, and outcomes, suggesting the potential for further analysis by combining datasets to examine SLA.

Chen et al. [30] pioneered the development of an early-stage design-based tool dedicated to crafting student-centric SLA within a postsecondary environment. Their research focused on fostering meaningful student discussions in online courses, introducing analytical tools designed to transform discussion forum data into actionable insights for student reflection. The findings underscored the imperative for SLAs to prioritize user-friendly tools, fostering student data literacy and seamless alignment between analytics and pedagogical designs. In parallel, Verdu et al. [31] introduced MSocial, an integrated tool within the Moodle platform, enabling students to engage in social networks without disconnecting from the Moodle environment. MSocial

actively monitors student activities on social platforms, computes key social network analytics metrics, and promptly showcases results on the Moodle course platform. This monitoring empowers educators to easily comprehend, visualize, and scrutinize student social interactions, facilitating the integration of these insights to enhance the overall learning process.

14.2.3 Mobile Learning Analytics

Presently, educational landscapes are in continual flux, witnessing dynamic transformations. Mobile learning is experiencing a growing embrace within the realm of education. Hybrid learning environments, seamlessly integrating formal and informal learning aspects, are emerging in diverse activities spanning distributed environments, physical spaces, and virtual realms. The advent of technology encourages us to navigate and adjust to these evolving educational settings adeptly. Quintero et al. [32] undertook a comprehensive study comparing technologies for creating mobile applications specifically designed for visualizing Learning Analytics. They developed prototypes of mobile applications using a case involving competency-based assessment analytics, wherein decision criteria were established to guide the selection of the mobile application type. The study emphasized the significance of defining the application's scope to adjust the time and cost of development. Regarding mobile visual analytics development, the researchers concluded that a profound understanding of the context is crucial, facilitating the systematic exploration and classification of existing visualization libraries pertinent to the identified problem. Furthermore, the study highlighted that developing native applications is not obligatory.

Pishtari et al. [33] employed supervised machine learning (SML) algorithms to categorize textual content automatically within mobile learning (m-learning) analytics designs. The pedagogical classifications employed in this study were pertinent to the learning tasks integrated into the designs. Avastusrada and Smartzoos served as tools, providing the dataset and features essential for crafting mobile learning applications. The research employed EstBERT and Logistic Regression algorithms for optimization, comparison, and achieving optimal performance. The outcomes of the SML implementation indicated an accuracy exceeding 0.86 and Cohen's kappa surpassing 0.69.

Shorfuzzaman et al. [34] introduced a cloud-based mobile learning framework employing Big Data analytics to extract insights from substantial volumes of learner data within mobile learning environments. Their empirical study proposed a model for adopting mobile learning, extending the Technology Acceptance Model (TAM). The suggested framework addresses the processing limitations of mobile devices by offloading computationally intensive tasks to the cloud, thereby leveraging ample computational and storage capabilities.

Kabassi and Alepis [35] concentrated on integrating learning analytics data derived from diverse modes of human-computer interaction and contemporary smartphones. By amalgamating data from various modalities, more precise insights could

be derived, ultimately enhancing and supporting the learning journey through tailored software design solutions. Their approach involved the application of multiple theories, including Multi-Criteria Decision-Making (MCDM), Analytic Hierarchy Process (AHP), and Simple Additive Weighting (SAW).

Aljohani and Davis [36] outlined the theoretical benefits of employing LA methods to improve learning in mobile and ubiquitous learning settings. Their study introduced the Mobile and Ubiquitous Learning Analytics Model (MULAM), designed to examine learner data within mobile environments. The model utilized the five-step Learning Analytics approach proposed by Campbell and Oblinger, encompassing Capture, Inform, Predict, Act, and Refine.

Pishtari et al. [37] conducted a comprehensive literature review examining the intersection of Learning Design (LD) and LA within mobile and ubiquitous learning (m/u-learning) environments. The study offered a contemporary snapshot of the field and revealed shared interests between the learning design approach and learning analytics in developing m/u-learning environments. This identified convergence establishes a symbiotic relationship, fostering potential mutual benefits in LD and LA domains.

Viberg et al. [38] introduced the MALLAS conceptual framework, aiming to enhance second language learning by integrating LA and self-regulated learning (SRL). This innovative framework incorporates assisted application and service design within mobile learning (m-learning) environments, offering a holistic approach to support learners in their language learning journey.

Tabuenca et al. [39] conducted a longitudinal study investigating the impact of monitoring learning time with a mobile tool on self-regulated learning. The study uncovered significant findings: (1) Positive effects of learning time tracking on enhancing time management skills in online courses, (2) Assistance for students in developing learning-to-learn competence through a real-time Learning Analytics feedback approach utilizing two channels—notifications and graphical visualization, and (3) Detailed specifications and practical insights for instructional designers and teachers to implement analogous approaches in the educational process.

14.2.4 Big Data and Learning Analytics

The utilization of Big Data analytics extends to both business and educational domains. Big Data enables educators to assess class-wide performance and individual student achievements. The primary goal is to formulate pedagogical strategies for the entire class, specifically identifying individual strengths and weaknesses. Seufert et al. [40] proposed a comprehensive design framework for LA, encompassing crucial dimensions and facilitating the creation of LA services supporting the educational process. The framework operates on a two-dimensional scale, distinguishing between individual vs. social and reflective vs. predictive aspects. The research identifies four foundational approaches to LA: (1) studying performance prediction, (2) exploring formative individual assessment and feedback services, (3)

investigating Social Learning Analytics, and (4) examining the proficient use of LA applications. These approaches aim to enhance the learning process and outcomes. The study concludes with a discussion on model validation through case studies and provides insights into future research prospects concerning LA in the educational context.

Ang et al. [41] conducted a comprehensive literature review examining current and emerging trends in educational Big Data and data analytics. The analysis delved into the architectural and social challenges of the educational Big Data approach. Emphasis was placed on diverse data sources originating from educational platforms, encompassing Learning Management Systems (LMS), Massive Open Online Courses (MOOC), Learning Object Repository (LOR), Open Course Ware (OCW), Open Educational Resources (OER), Social Networks Linked Data, and Mobile Learning. These varied sources collectively contribute to the formation of Big Education Data. In a parallel literature review, Sin and Muthu [42] explored the implementation of Big Data technologies in education, specifically delving into educational data mining and LA within learning environments. Their findings highlighted the increasing role of data mining in education, shaping the learning landscape for new generations. This evolution focuses on predicting student performance through data mining techniques, introducing LA at the higher education level, and observing the utilization of Learning Analytics in social learning contexts.

Klašnja-Milićević et al. [6] introduced an architectural framework emphasizing the pivotal role of Big Data and Learning Analytics in education. This framework serves a dual purpose, facilitating the management of reform initiatives in higher education while aiding instructors in enhancing teaching and learning. It focuses on constructing efficient learning systems for students, instructors, course designers, and institutions. In a separate study, Romero and Ventura [43] conducted a comprehensive survey outlining the current landscape of Educational Data Mining (EDM) and LA. The evolution of EDM is evident in its diverse nomenclature in academic discourse, including Academic Analytics, Institutional Analytics, Teaching Analytics, Data-Driven Education, Data-Driven Decision-Making in Education, Big Data in Education, and educational data science. The research analysis highlights the convergence of EDM and LA as two collaborative communities striving to enhance learning through data. Furthermore, the study identifies two future trends: the imperative need for versatile EDM/LA tools capable of addressing various educational challenges through a unified interface, emphasizing the enhancement of model portability, and the incentive of a data-centric culture within educational institutions to enhance decision-making processes and improve the overall teaching–learning paradigm.

In educational contexts, numerous studies underscore the significance of big data analytics. Picciano [44] analyzed the evolving landscape of Big Data and analytics in U.S. education. This examination searches through the conceptual framework, practical applications, implementation strategies, and the growth trajectory of emerging big data and Learning Analytics methodologies. The objective is to provide educational administrators with insights for assessing the performance and integration of these innovative technologies. Reyes [45] contributed a study specifically concentrating on Learning Analytics, elucidating the pivotal roles played by stakeholders

in the teaching and learning process. The study seeks to optimize and enhance the learning experience by introducing a learner-centered analytics approach. This approach addresses technological challenges and ethical considerations and aspires to revolutionize the teaching–learning paradigm. Roy and Singh [46] delivered a comprehensive review of Big Data tools and technologies within Learning Analytics and Educational Data Mining. The review emphasizes the predominant focus of many authors on predicting student performance in educational environments.

Aguilar [47] delineated the potential impact of Learning Analytics on fostering equitable and socially just educational outcomes. By attending to the individual needs of each student, the approach aims to cultivate a more personalized, learner-centered learning environment. Khan et al. [48] crafted a systematic literature review protocol in the realm of Learning Analytics, shedding light on applications, challenges, existing solutions, and future directions concerning the application of Big Data techniques. In a study by Huang et al. [49] conducted across three universities in Taiwan and Japan, the objective was to explore the correlation between students' online learning actions and academic performance using Learning Analytics applications. The study incorporated machine learning to train the model, utilizing seven datasets. Results identified essential factors influencing the predictive performance of classification methods, including the number and categories of significant features and Spearman correlation coefficient values. Additionally, eight classifiers (GaNB, SVC, linear-SVC, LR, DT, RF, NN, and XGBoost) and five evaluators (accuracy, recall, precision, F1-measure, and AUC) were employed to assess the predictive performance of the classification methods.

Rabelo et al. [50] introduced SmartLAK, a significant Big Data software architecture designed to enhance Learning Analytics services. Utilizing an ontology grounded in the Experience API specification, SmartLAK adeptly captures and semantically represents the data streams from learners engaged in course-related learning activities. To facilitate efficient processing of substantial data volumes within virtual learning environments, SmartLAK employs an RDF database, ensuring high-performance accessibility for Learning Analytics services. The validation of SmartLAK took place at the Faculty of Education, University of Santiago de Compostela. Looking ahead, the authors propose enhancing the architecture by incorporating an Enterprise Service Bus. This future development aims to empower SmartLAK to integrate diverse data flow sources seamlessly.

14.3 Data Visualization in Learning Analytics

Recently, the escalating abundance, velocity, diversity, and magnitude of data necessitated the adoption of advanced technologies and methodologies for enhanced visualization and analysis. Graphics, esteemed for their efficacy in discerning crucial content-related information, have become invaluable. The pursuit of graph-based methodologies has gained prominence in various domains, including education,

driven by the desire for a more profound comprehension of contextual intricacies. This evolving trend aims to uncover strategies and avenues for continual improvement.

Although simple graphs can indicate, for example, where there is over- or under-performance, they cannot elucidate the underlying causes crucial for effective decision-making. Graphs delineate intricate relationships among elements, unraveling the nature and structure of links within specific data sets. These links are imperative to discern the how and why of elements, constituting a key advantage in visualizing and analyzing graphs to enhance decision-making. Visualizing these links is paramount, aiding comprehension in specific cases identified through graph analysis or scenarios involving unprocessed information. Therefore, data visualization seeks a profound understanding of data swiftly. Furthermore, owing to the proliferation of Big Data, the utilization of graph visualization and analysis is on the rise as a method to derive novel insights from numerous unverified, irregular, or complex interconnected data streams.

In the contemporary landscape, open-source cloud-based tools exhibit enhanced capacities for efficiently gathering substantial data and presenting it visually within seconds. This functionality streamlines the data analysis process, enabling the extraction of additional insights to enhance knowledge and contribute to effective decision-making, as outlined by Brath and Jonker [51].

Subsequent sections provide a detailed description of the graphs utilized for data visualization and analysis.

3D pie chart. It graphically represents segmented pie slices of varying colors, each indicating relative frequencies or magnitudes. The size of each sector corresponds proportionally to the quantity it represents. This chart is presented in three dimensions, incorporating depth, width, and height, thus termed three-dimensional [52].

Area chart. It is a valuable tool for identifying trends or variations in data across time or diverse categories. Resembling a line chart, it distinguishes itself by shading the lower space beneath the lines with a specific color. This shading enhances the visualization of trend magnitudes or potential variations [53].

Bar chart. It visually displays a dataset using rectangular bars, either vertically or horizontally, with lengths proportional to the represented values, depicted in various colors. Additionally, it asymmetrically represents measures of central tendency. This chart type proves advantageous for illustrating multiple data series and capturing trends across time [54].

Line chart. It facilitates the identification of volatility, acceleration, or trends over time. Comprising a series of points connected to form a complete line, it effectively communicates the changes in a variable. Particularly beneficial when visualizing the behavior of one or more quantitative variables across a period, this chart employs different colors to distinguish points and lines, enhancing analytical clarity [55].

Scatterplot matrix. It combines all attributes of scatterplots in a matrix, utilizing colors to differentiate variables. This form of visualization facilitates comparisons between dimensions and allows for measuring differences among different variables vertically or horizontally. Moreover, this visual representation effectively identifies linear correlations among the variables [56].

Hierarchical pie chart. Also known as a sunburst chart, it is closely related to the tree map. In this hierarchical pie chart, intermediate levels are organized in consecutive rings, and each intermediate level is analyzed based on its color and size. This chart provides a visual representation of hierarchical relationships, with each ring representing a distinct level of hierarchy, making it a valuable tool for hierarchical data analysis [51].

Timeline chart. It visualizes a sequential order of events within a narrative, process, or story. Timelines, presented either horizontally or vertically, offer a simplified method for comprehending the roles of various events, processes, and actions played within a specified time frame. Moreover, they allow the visualization of concurrent events, their durations, relationships, and the specific points in time when they occurred [57].

Network graph. This graphical representation illustrates a network of interconnected nodes and edges based on a dataset. It reveals the flow of information, the spatial arrangement of network components, and their interactions. Nodes or vertices, often depicted as small dots or circles (sometimes using icons), represent entities, while linking lines portray connections, elucidating the nature of relationships within a group of entities. Links are depicted as simple lines connecting nodes [58].

Sankey diagram. In this scenario, limited quantity data is presented and examined from left to right, where the thickness of a given link indicates the minimum or maximum quantity. Incoming links intersect a specific node at right angles on one side, while outgoing links similarly exit the node on the opposite side. On both sides of each node, links are grouped at distinct entry and exit points corresponding to each node, representing the total amount of entry and exit [51].

Gauge chart. Also referred to as a clock chart or speedometer chart, this type of chart utilizes needles to indicate specific data as if reading an analog clock. It comprises a gauge axis with interval markers, data ranges, color ranges, needles, and a central dynamic list point. Each needle's value is interpreted within the colored data range or concerning the chart axis. This graph compares values among a few variables through one or multiple needles on the same indicator or by using several indicators [59].

Treemap. Employing a hierarchical structure reminiscent of a tree, the Treemap visually organizes information into nested rectangles that fully occupy the plot. The size of each rectangle corresponds to the volume of the category or subcategory, termed as nodes, which can be of root or leaf type. The root node is the initial one, and the leaf node, the final in the hierarchy, detaches from another. Consequently, the

aggregate of the root node (category) encompasses the sum of all leaf nodes (subcategories). Notably, Treemap enables the comparison of subcategories and categories through distinct colors for enhanced differentiation [51].

Heatmap. This graphical representation of data utilizes colors as aesthetic elements to convey varying activity levels. Dark hues represent low activity, while light or vibrant colors denote high activity. Heatmaps can be presented through rectangles or spatial layouts. This visualization technique proves beneficial for handling and analyzing extensive datasets, facilitating the detection of patterns and fluctuations within the data [60].

Bubble chart. Though resembling scatter charts, bubble charts introduce an additional variable indicating the size of the circles (bubbles). These bubbles, represented by markers, vary in size to convey relative importance and are colored to depict differences between categories or represent additional data variables. This chart exposes and compares relationships among categorized or labeled bubbles through dimensional aspects or positioning. The overview box facilitates the examination of correlations or patterns, and interactivity can be incorporated for each bubble to provide additional information, apply filters, or rearrange the bubbles, enhancing the user's engagement [61].

Tag cloud chart. The tag cloud chart highlights words or keywords based on their frequency by taking the shape of a cloud or other imaginative forms. Also known as a tag cloud, it visually represents tags or labels of varying sizes and colors on a webpage, blog, or website. Larger-sized tags with more intense colors indicate a higher frequency of appearance, providing a visual snapshot of the most prominent terms [62].

Doughnut Chart. A variation of the pie chart, the doughnut chart features a central hole resembling a ring and represents categories through colored arcs rather than sectors. Each value's portion of the ring occupied is proportional to its frequency, meaning larger values occupy a more significant portion of the chart. The central ring aims to prevent confusion regarding the area parameter. An advanced version, the disaggregated or multiple-ring graph diagram, introduces additional complexity by incorporating one or more segments detached from the central ring [63]. The subsequent section describes the APIs for Learning Analytics.

14.4 APIs for Learning Analytics

Application Programming Interfaces (APIs) for Learning Analytics are crucial for learning analytics as they enable interoperability between different systems, allowing for easy data exchange. Their flexibility facilitates custom data extraction based on specific analytical needs. APIs automate data collection, reducing manual errors and allowing real-time data analysis for immediate feedback. They also ensure secure access to sensitive data, maintaining privacy and regulatory compliance. Lastly, APIs

support scalable analysis, handling large datasets from various sources. Below is a brief description of six learning analytics APIs, Learning Analytics, Open Education, Moodle Analytics, Google, Learning Analytics Machine, and edX Data Analytics, along with their key features.

Learning Analytics API

The Learning Analytics API is a key component in Learning Analytics, offering features and capabilities to enhance educational data analysis. The Learning Analytics API consists of several data tables. A calculator is integrated into the API, processing learning analytics data and calculating a status value. The API's calculator analyzes learning analytics data and computes a status value. This status value likely represents a summarized metric or indicator derived from the analyzed data. The API is designed with a focus on modularity and scalability. It maintains a separation between data capture and the analytics/reporting processes. This separation allows for a more flexible and scalable Learning Analytics Infrastructure. The Learning Analytics API is engineered to support the development of scalable Learning Analytics Infrastructures. This scalability ensures that the API can handle varying volumes of data and accommodate the needs of different educational contexts. One of the primary goals of the API is to facilitate the building of scalable Learning Analytics Infrastructure. Providing the necessary tools and functionalities enables developers to create robust analytics systems. Learning Analytics API Methods:

- a. `getStudentData(studentId)`: Returns demographic, academic, and other relevant information for a single student.
- b. `getCourseData(courseId)`: Gives full details about a course—its structure, assignments, tests, etc.
- c. `getPerformanceData(studentId, courseId)`: Provides a student's performance details within a particular course—grades, activity, engagement levels, etc.
- d. `getPredictedPerformance(studentId, courseId)`: Uses machine learning models to predict a student's future performance based on historical data.
- e. `getPersonalizedRecommendations(studentId)`: Returns a list of personalized learning recommendations for a student.

The previous methods show some of the more distinctive features of the API, like the predicted performance of a student and the corresponding personalized recommendations for improving the student's performance [64].

Open Education API (Open Onderwijs)

Open Onderwijs API is a collaborative effort involving education institutes and suppliers in the Netherlands to create an open API for educational purposes. The API emphasizes openness and shared resources in the educational domain. There is a reference implementation of the Open Onderwijs API in Python (Django).

Open Education API Overview: The API focuses on relationships between specified objects, emphasizing concepts like `programOffering`, `courseOffering`, and `conceptOffering`.

Endpoints and Concepts. The offering endpoint encompasses smaller concepts like `programOffering`, `courseOffering`, and `conceptOffering`. Program relations are not presented as a separate endpoint; instead, they can be explored within the program endpoint.

Tagged Functionalities. The API includes various tagged functionalities, such as service metadata, academic sessions, associations, buildings, courses, components, education specifications, groups, news, offerings, organizations, persons, programs, and rooms, among others.

Models. The API defines several models for different entities, including services, education specifications, programs, courses, components, program offerings, course offerings, component offerings, associations, persons, groups, academic sessions, organizations, buildings, rooms, news feeds, and news items [65].

Moodle Analytics API

Moodle, a popular open-source Learning Management System (LMS), does have built-in analytics and reporting features, which can be extended by various plugins. However, there is no designated “Moodle Analytics API” as such. However, the Moodle core does provide various APIs, including the Web Services API, with which developers can extract and process analytical data.

Moodle Web Services API: Web services are a collection of defined protocols that allow communication between Moodle and other systems. This API is designed for service-oriented architectures and machine-to-machine communication. It enables other systems to log in to Moodle and create, retrieve, update, and delete entities such as users, courses, etc.

Methods: This API gives developers the ability to call core Moodle functions such as CRUD operations for users, courses, grades, and more—for instance, `core_user_create_users`, `core_course_create_courses`, `core_grades_get_grades`, etc.

Protocols: Moodle Web Services API supports multiple protocols, including REST, XML-RPC, and SOAP.

Security: Each web service call should be accompanied by a user token for authentication. Also, Moodle uses access control based on defined capabilities.

Language Support: Moodle provides multilanguage support so that communication can be conducted in various languages.

Data Formats: The API allows for data exchange in multiple formats, including XML and JSON.

Using web services, analytics data can be fetched and processed as needed. There are plugins in Moodle, like “Moodle Learning Analytics” or “Inspire Analytics”, that offer machine learning-backed predictions and insights. These, however, may not necessarily provide dedicated APIs and might be more about enhancing Moodle’s built-in analytics functionalities [66].

Google Analytics API

Google Analytics API offers a way for developers to programmatically interact with their Google Analytics data, enabling them to build custom dashboards, automate

complex reporting tasks, and integrate Google Analytics data into their business applications.

Data Access. Google Analytics APIs consist of tools for accessing Analytics data: Reporting API v4 for pulling data associated with users, sessions, and events; Real-Time Reporting API for accessing real-time data; Multi-Channel Funnels Reporting API for accessing conversion path data.

Management. The Google Analytics Management API allows administrators to configure accounts, manage properties and views, set user permissions, and handle segments, goals, filters, etc.

Data Insertion. Google Analytics Measurement Protocol allows developers to send data to Google Analytics from any device connected to the web.

Embed API. Used for creating and embedding interactive dashboards and report components into your apps using JavaScript.

Security. Every request must be authorized, typically via OAuth2.

Language Support. Client libraries are offered in various programming languages, and the HTTP/HTTPS interface can be used directly.

Data Formats. Mostly, JSON is used for request and response data [67].

Learning Analytics Machine API

The X5GON project aims to unify Open Educational Resources (OER) by offering freely available innovative technology. It features the Learning Analytics Machine (LAM), which can process multi-lingual OER collections, provide insights into resource usage across different languages and cultures, and enhance the visibility of your content globally. The X5GON LAM API is a REST Flask Python web API complemented by auto-generated swagger documentation. This API allows users to test the endpoints directly on a web page. Users can access the latest learning analytics work package results and retrieve content analytics made on the OERs through various endpoints. Analytics are based on AI models applied and scrutinized on the X5gon corpus, which gathers different OERs.

The LAM API, a part of the X5GON project, consists of three main parts:

1. **The Services.** They include endpoints built upon Learning Analytics (LA) models and heuristics, providing solutions for various LA issues. Updated versions can be found in the X5GON LAM API repository of the University of Nantes team.
2. **The X5gonlamtools.** They are the underlying algorithms that use OERs in diverse formats to calculate LA metrics and generate models needed for services. These tools tackle various LA problems like determining difficulty and concept continuity within OERs. Updated versions of these tools can be found in the X5GON LAM Tools repository of the University of Nantes team.
3. **The X5gonlammodels.** They are AI-based LA models calculated based on the OERs and the algorithms of X5gonlamtools. Updated versions of these models can be found in the X5GON LAM Models repository of the University of Nantes team.

The X5GON LAM API is a REST Flask Python web API complemented by auto-generated swagger documentation. This API allows users to test the endpoints

directly on a web page. Users can access the latest results of the learning analytics work package and retrieve content analytics made on the OERs through various endpoints. Analytics are based on AI models applied and scrutinized on the X5gon corpus, which gathers different OERs [68].

edX Data Analytics API

The edX Data Analytics API provides a convenient way to interact programmatically with the analytical data of edX, an open-source Learning Management System (LMS) targeted toward higher education. edX Data Analytics API key characteristics:

Data Accessibility. The edX Data Analytics API allows developers to access various data types, such as course activity, student enrollment, and learner profile data.

RESTful Interface. It offers a RESTful interface that provides a standard way for systems to interact with the edX platform.

Security. Requests to the API are authenticated using OAuth2 tokens, ensuring the privacy and security of the data.

JSON Format. The API communicates using the JSON data format both for requests and responses.

End Points. The edX analytics API provides several endpoints like the Course Activity endpoint, which presents data about learner activity, the Course Enrollment endpoint, which tracks learner enrollments in courses over time, among others.

Rate-Limiting. Each API endpoint has a specific rate limit to protect the system.

Pagination. The edX Analytics API uses pagination to handle large data requests [69].

The subsequent section presents a case study on data analytics for reading comprehension in higher education, utilizing data visualization through the graphs.

14.5 Case Study: Learning Analytics for Reading Comprehension in Higher Education

The study centered on implementing a reading comprehension assessment among higher education students. Reading comprehension, a crucial skill, involves grasping the significance of the words within a text and comprehending the overall message conveyed. In middle and higher education, the scope of this skill extends to a nuanced understanding, requiring students to comprehend, analyze, justify, and encapsulate the information presented in the text.

The outlined case study aimed to discern and assess the significance of integrating Learning Analytics into assessing students' reading comprehension skills. The primary goal was to assess the extent of reading comprehension, allowing for the identification of areas for improvement and weak points in individual students. This information served as a foundation for tailoring pedagogical strategies in a personalized manner, aiming to enhance reading comprehension abilities. Notably,

the presented graphs are accessible to both students and teachers, providing a detailed, analytical visualization of the student's reading comprehension levels.

In the examination, students read a narrative text and respond to related questions. The featured case study's reading comprehension assessment encompasses two focal domains: word readings and language comprehension. Notably, this study introduces an additional dimension by evaluating the correctness of responses and considering diverse variables. These include the time taken for execution and the methodologies employed in test development. This broader scope aims to provide insights into the general context of test execution, enhancing the overall understanding of reading comprehension assessment dynamics.

In the primary findings, the cohort of successful test-takers comprised 1065 students, with a gender distribution of 49.9% female and 50.1% male. The outcomes of those who excelled in the word reading section are visually represented in Fig. 14.1 using a treemap. This graphical representation illustrates the count of students for each assessed dimension, including decoding (15 women and 19 men), word recognition (55 women and 50 men), fluency (84 women and 72 men), phonological awareness (88 women and 98 men), and knowledge of written language (67 women and 64 men). A notable observation surfaced: a minority of students succeeded in decoding, while phonological awareness emerged as the dimension with the most successful students. The Treemap is also used for the representation and visualization of the most popular courses among students, among other aspects.

Moreover, Fig. 14.2 illustrates a horizontal bar chart showcasing the highest-performing questions on the assessment. The analysis discerned that most students demonstrate proficiency in understanding the concepts of written language, inferences, and text structure, with 96 students excelling across all three categories. Conversely, a subset of students, comprising 59 individuals, possesses limited knowledge of syntax and grammar.

Additionally, the horizontal bar chart helps to visualize specifically the questions or topics of some type of assessment where students present the highest number of errors, which indicates a lack of knowledge.

Similarly, Fig. 14.3 illustrates an area graph depicting the average time three students responded to five questions in the reading section. This graph facilitates a clear visualization of the comparative time allocation among the three students, mainly when focusing on the designated area corresponding to the question: What is the structure of the text?. Additionally, the area graph is used to visualize and observe student age trends and the total number of students by age when selecting any of them. This type of graph is also commonly used to present experiment results in academic papers.

Contrastingly, Fig. 14.4a illustrates an interactive timeline chart presenting the completion times of assessments for six students. Within the graph depicted in Fig. 14.4a, upon selecting a student's rectangle, the evaluation's initiation and conclusion dates are visible, with the time scale denoted in months. Additionally, the graph allows for dynamic expansion, enabling daily visualization of students' evaluation completion times, as shown in Fig. 14.4b.

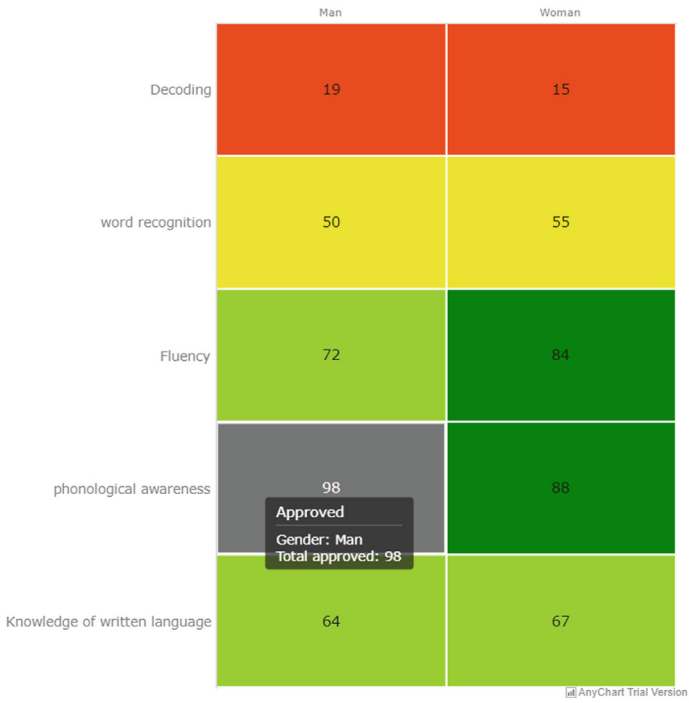


Fig. 14.1 Treemap of the distribution of passing students in the word reading section

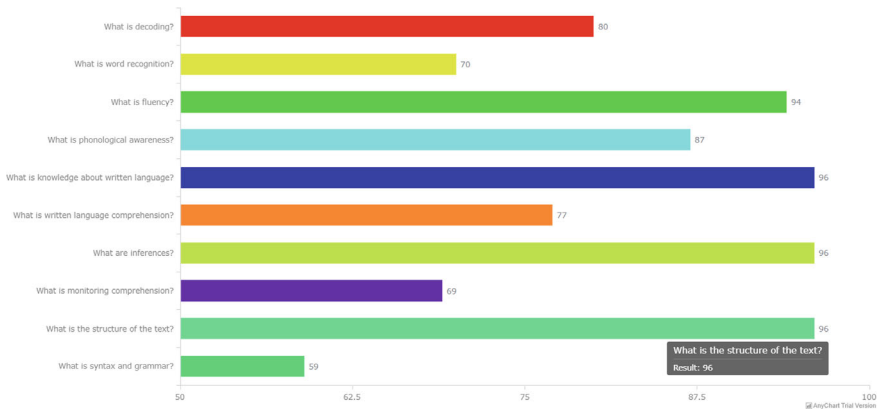


Fig. 14.2 Horizontal bar chart showing the most successful questions in the test

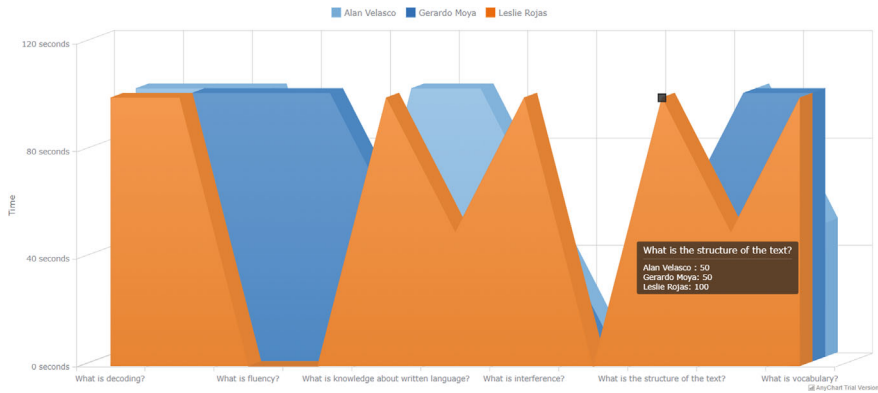


Fig. 14.3 Area chart showing the average time of the reading section

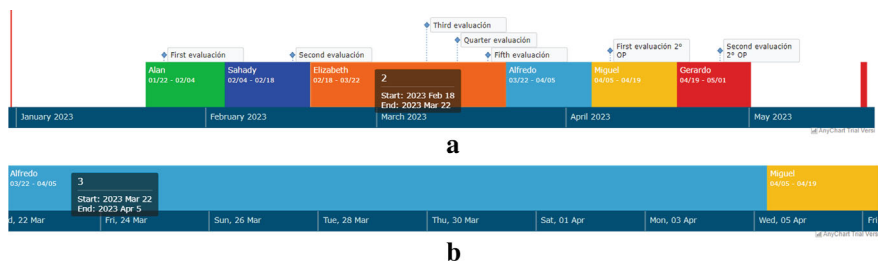


Fig. 14.4 a Timeline chart displaying the completion time of evaluations in months. b Timeline chart showing the time of completion of the evaluations in days

Additionally, a timeline chart can display learners’ browsing history, and it can be combined with text and other graphics to produce infographics, which present statistics, historical events, and any other type of information.

Figure 14.5 illustrates a horizontal bar chart presenting the performance outcomes of the top ten students in the language comprehension section. This graph enables a detailed examination of individual student scores in the assessed aspects by selecting the respective bar. For instance, the selected student on the graph achieved scores as follows: inferences (100), comprehension supervision (90), text structure (100), syntax and grammar (100), and vocabulary (90). Likewise, a horizontal bar chart visualizes the results of the most successful and least successful students in some type of assessment, where areas of opportunity are identified that are useful to maintain or improve the contents of a given topic.

The relationship and linkage of teachers concerning educational resources is visualized in the network graph in Fig. 14.6, where the amount of educational resources available to each teacher is clearly identified and observed, for example, when selecting the node of professor José Lauro, his relationship with the educational resources is indicated by the orange links.

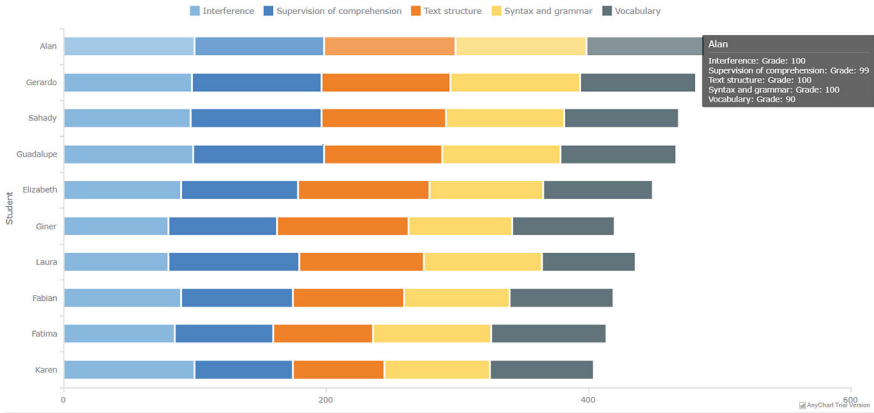


Fig. 14.5 Horizontal bar chart revealing the Top 10 of the language comprehension section

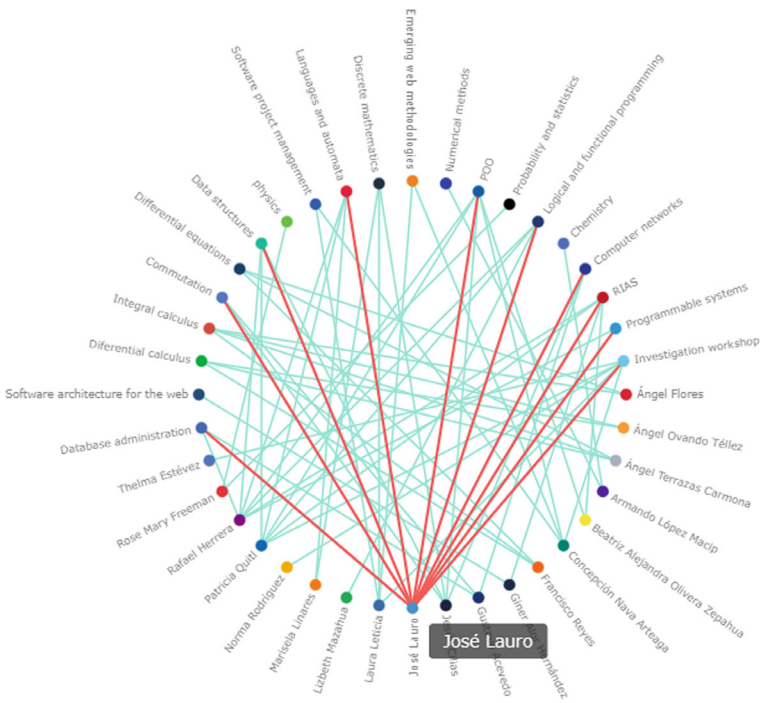


Fig. 14.6 Network graph: teachers with more educational resources

Regularly in a network graph, nodes are drawn as small dots or circles, but icons are also used. Links are drawn as simple lines connected between nodes. However, in some cases, not all nodes and links are created equally: i.e., additional variables are displayed, for example, by assigning a value to the node size. In addition, a network graph also makes it possible to visualize the competencies of learners concerning a subject or topic, making it possible to observe the aspects in which the learner requires some reinforcement.

Figure 14.7 displays a Sankey diagram depicting the teachers most favored or followed by students, providing a visual representation of their preferences. The diagram also illustrates the relationships between these teachers. A Sankey diagram is also used to visualize students’ preferences regarding the subjects they take in their professional preparation. In addition, the Sankey chart is used in various contexts, such as finance, cyber security, criminal investigations, sales process, Web analytics, and supply chain management.

Similarly, Fig. 14.8 presents a gauge chart depicting the antiquity of teachers’ access to the digital platform for academic task assignments. The visualization illustrates that teachers with lower antiquity exhibit a more extensive utilization of the digital platform. This same type of chart is used to visualize the time students access the educational resources or activities the teacher assigns on the digital platform.

Figure 14.9 displays a Treemap chart showing the most popular courses among students. The visualization distinctly reveals that organic chemistry, methods, and thermodynamics are among the most popular courses. Additionally, a Treemap can be used to visualize the distribution of passing students in a particular subject, unit or topic.

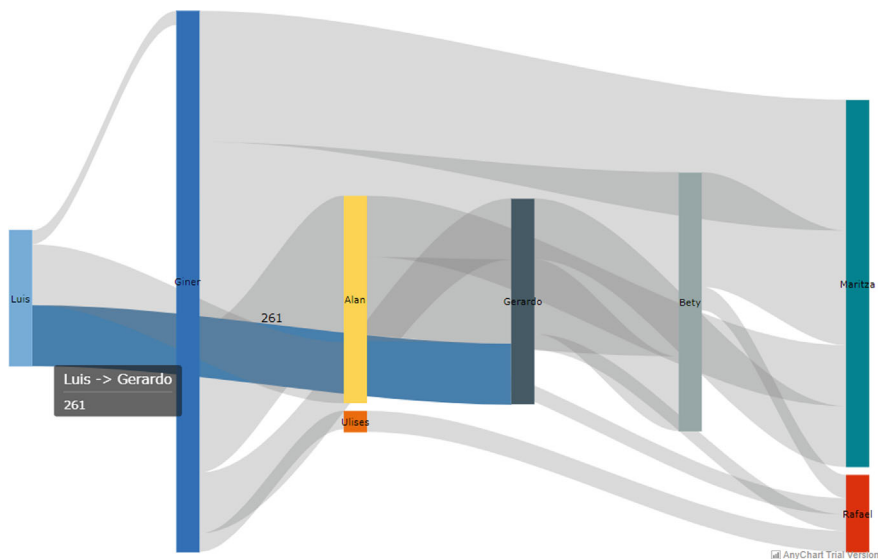


Fig. 14.7 Sankey diagram revealing the teachers most followed by students

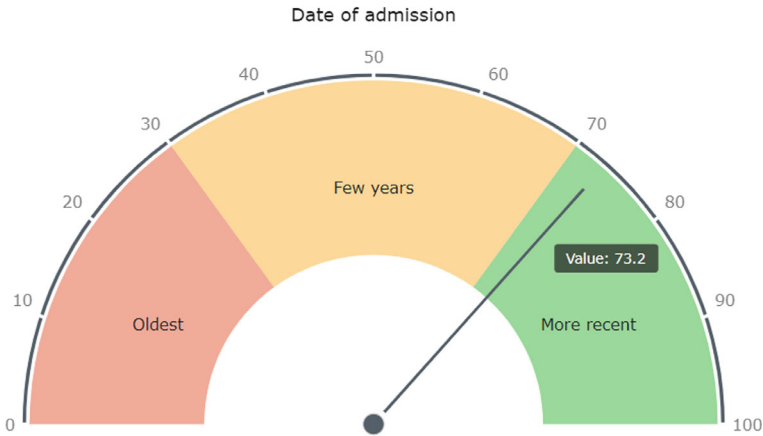


Fig. 14.8 Gauge chart showing the antiquity of the teachers' access to the digital platform

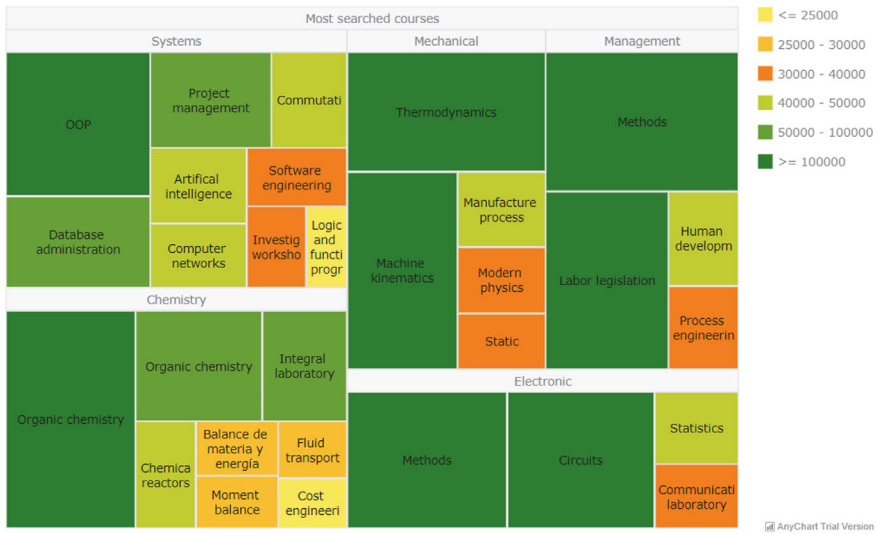


Fig. 14.9 Treemap chart showing the most popular courses among students

Conversely, Fig. 14.10 depicts a bubble chart designed to illustrate the devices students employ for accessing educational courses. Notably, the visualization reveals that the cell phone is a predominant choice among students, registering 19,103,423 views. The detailed information on the total visualizations for each device becomes visible upon clicking the respective bubble. Additionally, both the outline and the entire bubble adopt a darker hue. This allows teachers to make decisions and design and adjust the didactic material to the context of the device most used by students. In addition, this type of graph is also used to visualize the type of activities that are most

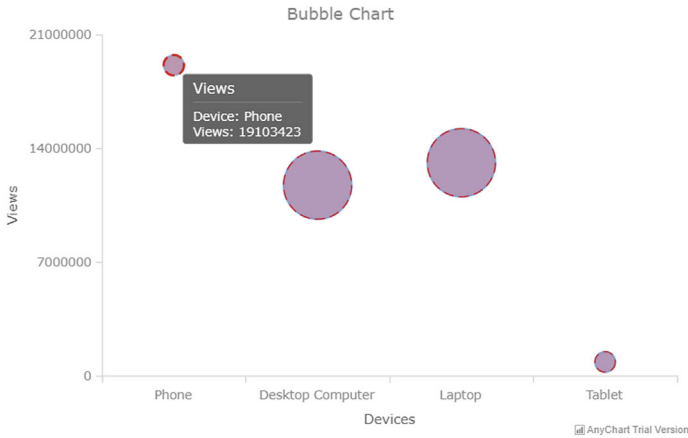


Fig. 14.10 Bubble chart indicating devices used by students to visualize educational courses

preferred or consulted by students, identifying areas for improvement in activities that are little consulted.

Furthermore, Fig. 14.11 presents a word cloud chart to illustrate the courses most favored by students. The prominent courses identified through the visualization include Object-Oriented Programming (OOP), Thermodynamics, Manufacturing Processes, and Database Administration. Additionally, a word cloud chart proves advantageous in accentuating crucial concepts within a topic, facilitating a swift identification of keywords within information.

The Tag cloud chart is continuously used in the educational field because it is beneficial for highlighting the main ideas of a topic and quickly identifying the keywords in some information. In addition, this type of graphic is widely used in visual marketing because, with the different colors and sizes of the labels or tags, it is easier to associate them with certain concepts.

Figure 14.12 illustrates a doughnut chart designed to depict the courses most frequently downloaded by students. The visualization reveals that Thermodynamics is the most downloaded course, amassing 9,101 downloads. The doughnut chart is also useful for visualizing the subjects, types of activities, or resources most preferred by students and those least consulted, allowing us to identify areas for improvement.

Figure 14.13 presents a Heatmap designed to depict the courses predominantly accessed by students every month. Notably, the visualization reveals that, for instance, during October, Integral recorded the highest number of visualizations at 9,012,371, followed by Physics with 6,425,674 visualizations.

A Heatmap is useful for representing and analyzing large data sets and identifying patterns and changes. It is also valued because it allows the optimization of the usability of an educational website, providing information on the exact points on which students focus their attention based on interactions and navigation behavior. The most commonly used Heatmaps are based on user clicks. However, it is also possible to use other techniques (eye-tracking, among others) to monitor the exact

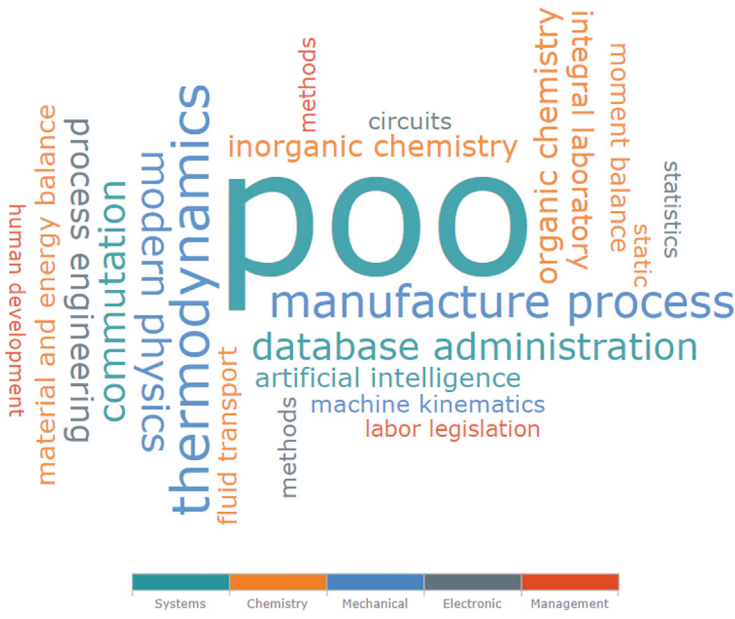


Fig. 14.11 Tag cloud chart exhibiting the most popular courses among students

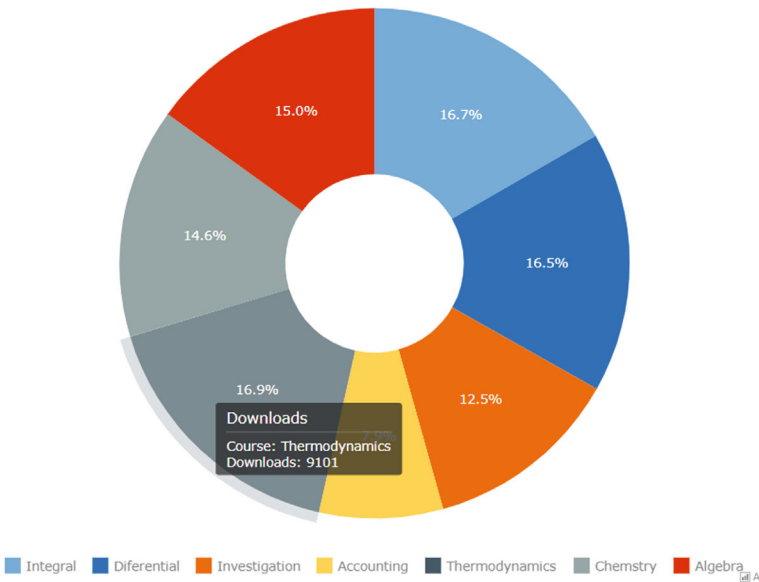


Fig. 14.12 Doughnut chart displaying the courses most downloaded by students

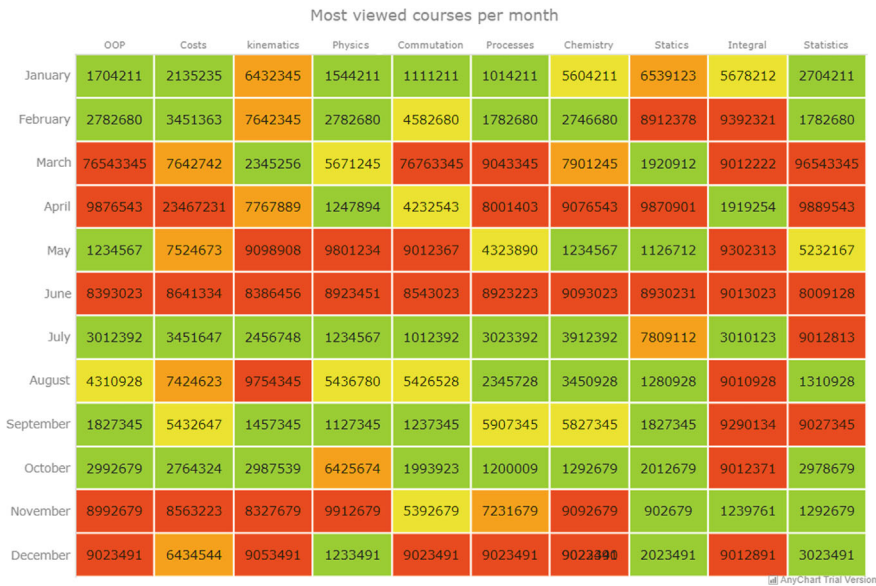


Fig. 14.13 Heatmap revealing the courses most viewed by students

points of the educational website on which the student’s gaze is focused. Additionally, it is possible to use heatmaps to represent the movement of the mouse or how far down the screen the learner reaches. Furthermore, it is important to note that heatmap analysis is often combined with other digital tools or resources to contribute to better decision making.

The visual representations were displayed on a webpage to enhance their visibility and analytical capabilities. These visuals were created using the AnyChart tool, specifically version 8.11.1. Subsequently, the succeeding section encompasses the primary findings, forthcoming research directions, and expressions of gratitude.

14.6 Conclusion

Learning Analytics involves measuring, collecting, analyzing, and presenting data concerning learners and their contexts to enhance understanding and optimize the learning process and its surrounding environments. This methodology enables the identification of patterns and trends within data, facilitating the generation of recommendations or alerts for instructors to address programmatic issues. Additionally, Learning Analytics encompasses the creation of reports, graphs, and visualizations to elucidate various behaviors in the learning process. This approach aids in pinpointing the root causes of data patterns or trends. For instance, if exam grades are consistently low, Learning Analytics can identify the underlying issue, such as

insufficient content depth. Real-time data analysis allows prompt adjustments to the content, contributing to improved instructional design. By scrutinizing specific content effectiveness, personalized feedback can be tailored, enhancing the overall learning experience.

This chapter elucidates the advantages of employing Learning Analytics in enhancing reading comprehension. The presented case study facilitates the identification of challenging areas for students, enabling timely interventions and targeted support from the institution. Such support may encompass supplementary resources like tutorials or study guides. Continuous monitoring of student performance allows institutions to discern trends and patterns, informing decisions to enhance the instructional program. Adjustments, such as curriculum modifications, increased assessments, or alterations in the teaching–learning approach, can be implemented based on this ongoing assessment. Furthermore, Learning Analytics personalizes learning experiences by analyzing performance data, interests, and preferences. Organizations can then tailor learning paths, offer customized feedback, and provide resources tailored to the specific needs of learners.

Learning Analytics, an evolving discipline, is reshaping educational methodologies. By meticulously collecting and analyzing student performance data, educational institutions can pinpoint areas of difficulty, monitor longitudinal progress, and customize learning experiences for each student.

Examining a comprehensive set of meaningful variables in Learning Analytics offers profound insights into the factors influencing actively engaged users' teaching and learning dynamics. This intricate analysis surpasses the limitations of basic bar chart data, providing a wealth of information essential for devising pedagogical strategies aimed at achieving specific learning objectives.

In future work, we intend to monitor an online course, workshop, or tutorial given in higher education to identify areas of opportunity, such as reasons for dropout or failure, the most exciting topics, the type of content most visited, the most viewed or the least viewed, and to improve the content, among other aspects.

Acknowledgements This research chapter was sponsored by Mexico's National Council of Humanities, Science and Technology (CONAHCYT) and Mexico's Secretariat of Public Education (SEP) through the PRODEP program. Authors also thank Tecnológico Nacional de México (TecNM), Sistema de Universidades Estatales de Oaxaca (SUNEO) and Instituto Politécnico Nacional (IPN) for supporting this work.

References

1. UNESCO: The right to education. Every Human Being Has the Right to Quality Education and Lifelong Learning Opportunities (2022). <https://www.unesco.org/en/right-education>. Accessed 17 Nov 2023
2. OECD: PISA (Programme for International Student Assessment) (2018). <https://www.oecd.org/pisa/>. Accessed 14 Nov 2023

3. Clow, D.: An overview of learning analytics. *Teach. Higher Educ.* **18**(6), 683–695 (2013). <https://doi.org/10.1080/13562517.2013.827653>
4. Ferguson, R., Buckingham Shum, S.: Social learning analytics: five approaches. In: Paper presented at the 2nd International Conference on Learning Analytics and Knowledge, Vancouver, British Columbia (2012). <https://doi.org/10.1145/2330601.2330616>
5. Fulantelli, G., Taibi, D., Arrigo, M.: A semantic approach to mobile learning analytics. In: *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality* (pp. 287–292). ACM, Salamanca (2013). <https://doi.org/10.1145/2536536.2536579>
6. Klačnja-Milićević, A., Ivanović, M., Budimac, Z.: Data science in education: big data and learning analytics. *Comput. Appl. Eng. Educ.* **25**(6), 1066–1078 (2017). <https://doi.org/10.1002/cae.21844>
7. Leitner, P., Ebner, M.: Development of a dashboard for learning analytics in higher education. *Learn. Collab. Technol. Technol. Educ. LCT* **12**, 293–301 (2017). https://doi.org/10.1007/978-3-319-58515-4_23
8. Jivet, I., Scheffel, M., Schmitz, M., Robbers, S., Specht, M., Drachsler, H.: From students with love: an empirical study on learner goals, self-regulated learning and sense-making of learning analytics in higher education. *Internet Higher Educ.* **47**, 100758 (2020). <https://doi.org/10.1016/j.iheduc.2020.100758>
9. Campbell, J.P., DeBlois, P.B., Oblinger, D.G.: Academic analytics: a new tool for a new era. *Educ. Rev.* **42**(4), 40–57 (2007)
10. Nunn, S., Avella, J.T., Kanai, T., Kebritchi, M.: Learning analytics methods, benefits, and challenges in higher education: a systematic literature review. *Online Learn. J.* **20**(2), 1–17 (2016)
11. Mattingly, K.D., Rice, M.C., Berge, Z.L.: Learning analytics as a tool for closing the assessment loop in higher education. *Knowl. Manag. E-Learn.* **4**(3), 236 (2012)
12. El-Alfy, S., Marx Gómez, J., Dani, A.: Exploring the benefits and challenges of learning analytics in higher education institutions: a systematic literature review. *Inform. Discov. Deliv.* **47**(1), 25–34 (2019). <https://doi.org/10.1108/IDD-06-2018-0018>
13. Leitner, P., Khalil, M., Ebner, M.: Learning analytics in higher education—a literature review. In: Pea-Ayala, A. (ed.) *Learning Analytics: Fundamentals, Applications, and Trends*, pp. 1–23. Springer, New York (2017). https://doi.org/10.1007/978-3-319-52977-6_1
14. Wong, B.T.M.: Learning analytics in higher education: an analysis of case studies. *Asian Assoc. Open Univ. J.* **12**(1), 21–40 (2017). <https://doi.org/10.1108/AAOUJ-01-2017-0009>
15. Zhang, J., Zhang, X., Jiang, S., Ordóñez de Pablos, P., Sun, Y.: Mapping the study of learning analytics in higher education. *Behav. Inform. Technol.* **37**(10–11), 1142–1155 (2018). <https://doi.org/10.1080/0144929X.2018.1529198>
16. Díaz-Lázaro, J.J., Solano Fernández, I.M., Sánchez-Vera, M.M.: Social learning analytics in higher education: an experience at the primary education stage. *J. New Approaches Educ. Res.* **68**(2), 119–126 (2017). <https://doi.org/10.7821/naer.2017.7.232>
17. Viberg, O., Hatakka, M., Bälter, O., Mavroudi, A.: The current landscape of learning analytics in higher education. *Comput. Hum. Behav.* **89**, 98–110 (2018). <https://doi.org/10.1016/j.chb.2018.07.027>
18. Tsai, Y., Gasevic, D.: Learning analytics in higher education: challenges and policies—a review of eight learning analytics policies. *ACM International Conference Proceeding Series*, pp. 233–242. Association for Computing Machinery (2017). <https://doi.org/10.1145/3027385.3027400>
19. Cerratto Pargman, T., McGrath, C.: Mapping the ethics of learning analytics in higher education: a systematic literature review of empirical research. *J. Learn. Anal.* **8**(2), 123–139 (2021). <https://doi.org/10.18608/jla.2021.1>
20. Adejo, O., Connolly, T.: Learning analytics in higher education development: a roadmap. *J. Educ. Pract.* **8**(15), 156–163 (2017)
21. Kuhnel, M., Seiler, L., Honal, A., Ifenthaler, D.: Mobile learning analytics in higher education: usability testing and evaluation of an app prototype. *Interact. Technol. Smart Educ.* **15**(4), 332–347 (2018). <https://doi.org/10.1108/ITSE-04-2018-0024>

22. Khousa, E.A., Atif, Y., Masud, M.M.: A social learning analytics approach to cognitive apprenticeship. *Smart Learn. Environ.* **2**(1), 14 (2015). <https://doi.org/10.1186/s40561-015-0021-z>
23. Hernández-García, Á., González-González, I., Jiménez-Zarco, A.I., Chaparro-Peláez, J.: Applying social learning analytics to message boards in online distance learning: a case study. *Comput. Hum. Behav.* **47**, 68–80 (2015). <https://doi.org/10.1016/j.chb.2014.10.038>
24. Manca, S., Caviglione, L., Raffaghelli, J.E.: Big data for social media learning analytics: potentials and challenges. *J. E-Learn. Knowl. Soc.* **12**(2), 27–39 (2016)
25. Hernández-García, Á., Conde-González, M.A.: Bridging the gap between LMS and social network learning analytics in online learning. *J. Inform. Technol. Res.* **9**(4), 1–15 (2016). <https://doi.org/10.4018/JITR.2016100101>
26. Doleck, T., Lemay, D.J., Brinton, C.G.: Evaluating the efficiency of social learning networks: perspectives for harnessing learning analytics to improve discussions. *Comput. Educ.* **164**, 104–124 (2021). <https://doi.org/10.1016/j.compedu.2021.104124>
27. Aguilar, J., Buendía, O., Pinto, A., Gutiérrez, J.: Social learning analytics for determining learning styles in a smart classroom. *Interact. Learn. Environ.* **30**(2), 245–261 (2022). <https://doi.org/10.1080/10494820.2019.1651745>
28. Kaliisa, R., Rienties, B., Mørch, A.I., Kluge, A.: Social learning analytics in computer-supported collaborative learning environments: a systematic review of empirical studies. *Comput. Educ. Open* **32**, 100073 (2022). <https://doi.org/10.1016/j.caeo.2022.100073>
29. Rienties, B., Toetenel, L.: The impact of 151 learning designs on student satisfaction and performance. In: Proceedings of the Sixth International Conference on Learning Analytics and Knowledge LAK '16, pp. 339–343 (2016). <https://doi.org/10.1145/2883851.2883875>
30. Chen, B., Chang, Y.-H., Ouyang, F., Zhou, W.: Fostering student engagement in online discussion through social learning analytics. *Internet Higher Educ.* **37**, 21–30 (2018). <https://doi.org/10.1016/j.iheduc.2017.12.002>
31. Verdu, M.J., De Castro, J.-P., Regueras, L.M., Corell, A.: MSocial: practical integration of social learning analytics into Moodle. *IEEE Access* **9**, 23705–23716 (2021). <https://doi.org/10.1109/ACCESS.2021.3056914>
32. Quintero, C.A., Florian-Gaviria, B., Pabon, O.S.: Comparative study of technologies for mobile learning analytics. In: Proceedings of the 9th Computing Colombian Conference (9CCC), pp. 82–89 (2014). <https://doi.org/10.1109/ColumbianCC.2014.6955361>
33. Pishtari, G., Prieto, L.P., Rodríguez-Triana, M.J., Martínez-Maldonado, R.: Design analytics for mobile learning. *J. Learn. Anal.* **9**(2), 236–252 (2022)
34. Shorfuzzaman, M., Hossain, M.S., Nazir, A., Muhammad, G., Alamri, A.: Harnessing the power of big data analytics in the cloud to support learning analytics in mobile learning environment. *Comput. Hum. Behav.* **92**, 578–588 (2019). <https://doi.org/10.1016/j.chb.2018.07.002>
35. Kabassi, K., Alepis, E.: Learning analytics in distance and mobile learning for designing personalised software. *Intell. Syst. Ref. Libr.* **158**, 185–203 (2020). https://doi.org/10.1007/978-3-030-13743-4_10
36. Aljohani, N.R., Davis, H.C.: Learning analytics in mobile and ubiquitous learning environments. In: *CEUR Workshop Proceedings*, pp. 70–77 (2012)
37. Pishtari, G., Rodríguez-Triana, M.J., Sarmiento-Márquez, E.M., Pérez-Sanagustín, M., Ruiz-Calleja, A., Santos, P., et al.: Learning design and learning analytics in mobile and ubiquitous learning: a systematic review. *Br. J. Educ. Technol.* **51**(4), 1078–1100 (2020). <https://doi.org/10.1111/bjet.12944>
38. Viberg, O., Wasson, B., Kukulska-Hulme, A.: Mobile-assisted language learning through learning analytics for self-regulated learning (MALLAS): a conceptual framework. *Austr. J. Educ. Technol.* **36**(6), 34–52 (2020)
39. Tabuenca, B., Kalz, M., Drachslar, H., Specht, M.: Time will tell: the role of mobile learning analytics in self-regulated learning. *Comput. Educ.* **89**, 53–74 (2015). <https://doi.org/10.1016/j.compedu.2015.08.004>
40. Seufert, S., Meier, C., Soellner, M., Rietsche, R.: A pedagogical perspective on big data and learning analytics: a conceptual model for digital learning support. *Technol. Knowl. Learn.* **24**(4), 599–619 (2019). <https://doi.org/10.1007/s10758-019-09399-5>

41. Ang, L.M., Ge, F., Seng, K.: Big educational data and analytics: survey, architecture and challenges. *IEEE Access* **8**, 116392–116414 (2020). <https://doi.org/10.1109/ACCESS.2020.2994561>
42. Sin, K., Muthu, L.: Application of big data in education data mining and learning analytics: a literature review. *ICTACT J. Soft Comput.* **6956**, 1035–1049 (2015)
43. Romero, C., Ventura, S.: Educational data mining and learning analytics: an updated survey. *WIREs Data Min. Knowl. Discov.* **10**(3), 1355 (2020). <https://doi.org/10.1002/widm.1355>
44. Picciano, A.G.: The evolution of big data and learning analytics in American higher education. *J. Asynchr. Learn. Netw.* **16**(3), 9–20 (2012)
45. Reyes, J.A.: The skinny on big data in education: learning analytics simplified. *TechTrends* **59**(2), 75–80 (2015). <https://doi.org/10.1007/s11528-015-0842-1>
46. Roy, S., Singh, S. N.: Emerging trends in applications of big data in educational data mining and learning analytics. In: *Proceedings of the 2017 7th International Conference on Cloud Computing, Data Science and Engineering-Confluence*, IEEE, pp. 193–198 (2017). <https://doi.org/10.1109/confluence.2017.7943148>
47. Aguilar, S.J.: Learning analytics: at the nexus of big data, digital innovation, and social justice in education. *TechTrends* **62**, 37–45 (2018). <https://doi.org/10.1007/s11528-017-0226-9>
48. Khan, S.U., Bangash, S.A.K., Khan, K.U.: Learning analytics in the era of big data: a systematic literature review protocol. In: *Proceedings of the 2017 International Symposium on Wireless Systems and Networks (ISWSN)*, pp. 1–7 (2017). <https://doi.org/10.1109/ISWSN.2017.8250033>
49. Huang, A.Y.Q., Lu, O.H.T., Huang, J.C.H., Yin, C.J., Yang, S.J.H.: Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Interact. Learn. Environ.* **28**(2), 206–230 (2020). <https://doi.org/10.1080/10494820.2019.1636086>
50. Rabelo, T., Lama, M., Amorim, R.R., Vidal, J.C.: SmartLAK: A big data architecture for supporting learning analytics services. In: *Proceedings of the 2015 IEEE Frontiers in Education Conference (FIE)*, pp. 1–5 (2015). <https://doi.org/10.1109/FIE.2015.7344147>
51. Brath, R., Jonker, D.: *Graph Analysis and Visualization: Discovering Business Opportunity in Linked Data*. Wiley, New York (2015)
52. De, P.: Automatic data extraction from 2D and 3D pie chart images. In: *Proceedings of the 2018 IEEE 8th International Advance Computing Conference (IACC)*, pp. 20–25 (2018). <https://doi.org/10.1109/IADCC.2018.8692104>
53. Davila, K., Setlur, S., Doermann, D., Kota, B.U., Govindaraju, V.: Chart mining: a survey of methods for automated chart analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(11), 3799–3819 (2021). <https://doi.org/10.1109/TPAMI.2020.2992028>
54. Newman, G.E., Scholl, B.J.: Bar graphs depicting averages are perceptually misinterpreted: the within-the-bar bias. *Psychon. Bull. Rev.* **19**, 601–607 (2012). <https://doi.org/10.3758/s13423-012-0247-5>
55. Archambault, S.G., Helouvy, J., Strohl, B., Williams, G.: Data visualization as a communication tool. *Library Hi Tech News* **32**(2), 1–9 (2015). <https://doi.org/10.1108/LHTN-10-2014-0098>
56. Nguyen, Q.V., Miller, N., Arness, D., Huang, W.D., Huang, M.L., Simoff, S.: Evaluation on interactive visualization data with scatterplots. *Vis. Inform.* **4**(4), 1–10 (2020). <https://doi.org/10.1016/j.visinf.2020.09.004>
57. Hasan, K.T., Addullah, S., Ahmed, R., Giunchiglia, F.: The history of temporal data visualization and a proposed event centric timeline visualization model. *Int. J. Comput. Appl.* **10**(27), 27–33 (2013)
58. Cherven, K.: *Network Graph Analysis and Visualization with Gephi*. Packt Publishing Ltd., Chicago (2013)
59. García, M.Á., Harmsen, B., Redmond, S., Pover, K.: *QlikView: Advanced Data Visualization: Discover Deeper Insights with Qlikview by Building Your Own Rich Analytical Applications from Scratch*. Packt Publishing, Reino Unido (2018)
60. Gu, Z.: Complex Heatmap Visualization. *iMeta* **1** (2022). <https://doi.org/10.1002/imt2.43>

61. Aspin, A.: Pro Power BI Desktop. Apress, United States (2017). <https://doi.org/10.1007/978-1-4842-3210-1>
62. Dunaiski, M., Greene, G.J., Fischer, B.: Exploratory search of academic publication and citation data using interactive tag cloud visualizations. *Scientometrics* **110**, 1539–1571 (2017). <https://doi.org/10.1007/s11192-016-2236-3>
63. Cai, X., Efstathiou, K., Xie, X., Wu, Y., Shi, Y., Yu, L.: A study of the effect of doughnut chart parameters on proportion estimation accuracy. *Comput. Graph. Forum* **37**(6), 300–312 (2018). <https://doi.org/10.1111/cgf.13325>
64. LA-API: The Learning Analytics API (2024). <https://beyondlms.org/tools/LA-API/>. Accessed 4 Jan 2024
65. Open-education-api: Learning-Analytics (2024). <https://github.com/open-education-api/learning-analytics>. Accessed 9 Jan 2024
66. Moodle: Moodle Analytics API (2024). <https://moodledev.io/docs/apis/subsystems/analytics>. Accessed 12 Jan 2024
67. Google Analytics: Analytics (2024). <https://analytics.google.com/analytics/web/provision/#/provision>. Accessed 15 Jan 2024
68. LAM API: Learning Analytics Machine API (2024). <https://github.com/X5GON/lamapi>. Accessed 18 Jan 2024
69. edX: edX Data Analytics API (2024). <https://edx.readthedocs.io/projects/edx-data-analytics-api/en/latest/overview.html>. Accessed 22 Jan 2024

Chapter 15

Improving Home Loan Predictions: A Fusion of PCA, Decision Tree and Random Forest Approaches



S. S. Thakur , Soma Bandyopadhyay , Sudip Kumar Bera,
and Mahika Thakur

Abstract Loan default happens when a borrower receives funds from banks but fails to refund the loan. Debt is commonly utilized by individuals to acquire assets like homes and vehicles that would otherwise be unaffordable. As global economies become more interconnected and interdependent, the demand for capital has significantly increased. Loans can offer financial benefits when used wisely, but they also pose notable challenges. The last decade saw a surge in retail, small and medium-sized enterprises (SME), and commercial borrowers, with rising defaults impacting financial institutions. In this proposed research, the authors aim to develop a predictive model for identifying potential loan defaulters in the consumer lending sector. Dataset from Kaggle i.e. loan prediction based on customer behavior in .CSV format has been used which contains valuable data on previous client behavior, including demographic characteristics of each customer and a target variable indicating loan default or non-default. Principal Component Analysis (PCA) is used in our work which identifies principal components and helps in dropping features of less importance. Leveraging this data, our objective is to forecast the credit risk of new consumers and distinguish between higher-risk and lower-risk individuals. The result shows that 89% accuracy has been achieved using Random Forest (RF) classifier which is bit higher than the accuracy of Decision Tree (DT) classifier which is 88%. This analysis would assist financial institutions in making informed decisions when acquiring new customers, enabling them to effectively manage risk and optimize lending strategies.

S. S. Thakur (✉) · S. Bandyopadhyay · S. K. Bera · M. Thakur
MCKV Institute of Engineering, Liluah, Howrah, West Bengal, India
e-mail: subroto_thakur@yahoo.com

S. Bandyopadhyay
e-mail: somabanmuk@yahoo.co.in

S. K. Bera
e-mail: sudip7407@gmail.com

M. Thakur
e-mail: thakur.mahika02@gmail.com

Keywords Loan defaults · Principal component analysis · Random forest · Decision tree · Prediction

15.1 Introduction

In today's global scenario, individuals worldwide depend on banks for providing with loans to overcome financial constraints and accomplish personal goals. The dynamic nature of the economy and intensifying competition in the financial industry have made loan acquisition a necessity. Furthermore, banking institutions of all sizes depend on lending activities to generate profits, manage their operations, and navigate financial challenges. A substantial portion of banks' revenue is derived from loans provided to customers. Interest is levied on these loans that are disbursed to customers. Thus, loans serve as a primary income source for banks, with a significant portion of their assets stemming from the interest earned on these loans. While lending loans offers substantial benefits for both borrowers and lenders, it is not without risks. These risks primarily manifest credit risk signifies the potential for a borrower to default on their loan obligations [1]. The credit lending sector within the banking industry has experienced significant expansion, driven by both established banks and a wave of emerging credit start-ups. This growth has coincided with fierce competition among these players. However, the surge in loan applications and borrowing has also led to a rise in bad credit losses, posing challenges for lenders. Credit loans are monetary provisions furnished by financial institutions to individuals. These loans are typically repayable within a specified period, with or without interest. Commonly, these loans must be paid back within a designated timeframe, either with or without an additional cost known as interest. They are pursued for a range of reasons, such as personal expenditures, educational costs, medical bills, vacations, and business necessities [2]. The primary goal of banks is to invest their assets in customers considered to be low-risk.

Currently, numerous banks follow a loan application process that involves verification and validation steps. However, no bank has been able to provide a guarantee regarding the safety of a customer selected for a loan application [3]. The increase in loan defaults is one of the factors contributing to the financial crisis in banking sector, resulting in tighter regulations being implemented for loan approval. The importance of forecasting loan defaults has grown due to banks' efforts to comply with laws and regulations, extend credit to qualified clients, decrease credit risks for ineligible clients, and optimize the efficiency of their application procedures. Effective credit risk management plays a vital role in ensuring a bank's enduring viability and expansion, given that lending constitutes a fundamental pillar of the banking industry. Over the past few years, the rise in customer loan risk coupled with the worsening impact of the pandemic has resulted in an elevated level of customer default risk. As a result, financial institutions and banks are now placing significant research emphasis on identifying high-risk customers. Customer creditworthiness serves as the benchmark for evaluating loan amounts and interest rates, making the

swift identification of customer information a prominent research area [4]. Due to advancements in the banking sector, there has been an upsurge in loan applications from individuals. However, banks have limited resources and can only grant loans to a select few. Therefore, determining the eligible candidates who pose lower risks for the bank has become a standard procedure.

Nowadays, numerous banks and financial firms assess loan applications through a regression process of authentication and validation. However, there is still no guarantee regarding the selection of the most deserving and suitable applicant from among all the applicants. In recent times, there has been a persistent rise in reported instances of financial fraud in India. These frauds, compared to traditional methods, have seen a significant increase in terms of frequency, complexity, diversity, and financial impact. As a result, these matters raise substantial worries for regulatory authorities. The strength and stability of a nation's financial infrastructure are pivotal in shaping the appeal of its economy for potential investments. In addition to this they function as markers of the citizens' welfare, safety, and quality of life. As a result, when the banking sector grapples with heightened levels of Non-Performing Assets (NPAs), it becomes a crucial issue as it reflects the financial challenges faced by borrowing customers. The Indian economy is significantly affected by these challenges. Significant number of individuals apply for loans on a daily basis, seeking funding for various purposes. However, not all applicants are genuine, and not everyone can be approved for credit. Therefore, it is of utmost importance to assess the associated risks by carefully analyzing the demographic data of the applicants [5]. The approval of loans in financial organizations presents challenges that impact the efficiency of the financial process, mostly due to inaccurate prediction or inadequate data. As a result, banks strive to minimize credit risks by conducting thorough evaluations of loan statuses to mitigate potential issues. In this context, accurate estimation based on given data and collected information play significant role. The field of data mining, especially machine learning, presents a hopeful strategy for providing accurate and prompt determinations regarding the approval or rejection of loans. The main aim of this research work is to explore the method of loan prediction through the utilization of diverse machine learning methods. Various machine learning approaches like Decision Trees and Random Forest can be employed to forecast the suitable candidate eligible for a loan. In addition to these machine learning techniques mentioned above, statistical technique Principal Component Analysis (PCA) has been applied in the dataset which contributes in dimensionality reduction.

The subsequent sections of this paper are organized in the following manner. Section 15.2 provides a concise literature review of prior research conducted on loan evaluation and credit risk assessment. The proposed methodology was included in Sect. 15.3. Section 15.4, outlines the model development which includes PCA, Decision Tree and the Random Forest approaches. The results and discussion have been explained in Sect. 15.5. Finally, Sect. 15.6 provides the concluding remarks.

15.2 Literature Review

Lee undertook a research investigation to assess the efficiency of credit scoring through the application of two data mining methods: classification and regression Tree (CART), and multivariate adaptive regression splines (MARS). To assess the practicality and efficacy of employing CART and MARS to develop models for credit scoring, a credit scoring task is conducted using a dataset from a bank's credit card records. The focus of the study primarily revolves around utilizing demographic variables as independent variables in the analysis [6]. Ince and Aktan explored credit scoring and assessed a bank's credit card policy by utilizing four separate methodologies. To evaluate the viability and effectiveness of these methods, a credit scoring exercise is performed using a dataset extracted from a bank's credit card records. This research examines the effectiveness of credit scoring models, emphasizing on both conventional approaches and artificial intelligence methods. The examined traditional methods encompass discriminant analysis and logistic regression, whereas the artificial intelligence approaches comprise neural networks (NNs) and classification and regression trees. By conducting experimental analyses using real-world datasets, the results revealed that both classification and regression trees, along with neural networks, showcase enhanced performance in contrast to the conventional credit scoring models. Particularly, these models excel in predictive accuracy and demonstrate a reduced occurrence of Type II errors. This research highlights the potential benefits of leveraging artificial intelligence techniques for credit scoring, indicating their ability to enhance predictive accuracy and improve decision-making in the credit assessment process [7].

M. V. J. Reddy and B. Kavitha introduced a method to predict class labels using NNs by incorporating attribute relevance analysis. A notable advantage of this approach lies in its capacity to minimize the necessary neural network units. By doing so, the prediction speed for new data instances can be increased. The proposed technique involves utilizing attribute relevance analysis to identify and eliminate irrelevant attributes from being used as inputs to the neural network. A simple neural network was utilized by the authors to assess how well this method performs in predicting class defaulters. The obtained results indicate the feasibility and effectiveness of this approach [8]. In their research, Odeh et al. employed the fuzzy simplex generic algorithm, a multi-objective optimization algorithm, to formulate decision rules for forecasting loan defaults within a representative credit institution. The data used in this research were sourced from the loan database of customers within the Seventh Farm Credit District. The results suggest that, among the top five rules, a consistently dependable indicator of default status is a low working capital percentage. This means that a company experiencing challenges in meeting its day-to-day operational funding requirements would likely struggle to fulfill its debt obligations, making default highly probable. Furthermore, the outcome highlights that a poor repayment history which is characterized by low refund capacity and low owners' equity, should be given significant consideration during credit assessments.

If both of these factors are significantly lower in comparison to the credit institutions' database, it could strongly suggest a likelihood of default [9]. Zhou and Wang conducted a study where they enhanced the original Random Forest algorithm by allocating weights to Decision Trees and employing a weighted majority approach for prediction. They demonstrated that this weighted majority approach improved the performance of Random Forests. Regarding overall accuracy and balanced accuracy, it outperformed other classifiers like SVM, KNN, and C4.5 [10].

Kemalbay and Korkmazoğlu conducted a survey on a dataset comprising 2331 randomly selected customers. The dataset included binary explanatory variables that were highly correlated with each other. Their objective was to model for determining the approval status of customers' housing loan applications. To address the issue of multicollinearity among the categorical explanatory variables and predict the binary response using logistic regression, researchers proposed the implementation of categorical PCA. This approach allowed them to effectively handle the multicollinearity problem and improve the predictive power of the logistic regression model [11]. Archana Gahlaut and her colleagues conducted credit risk prediction through the utilization of classification mining models. The dataset used in their study was obtained from the dataset contributed by Hans Hofmann and available on the UCI Machine Learning data repository. It comprises various variables, including Credit, Balance_credit_acc, Rate, Duration, Occupation, Age etc. Based on the analysis of the graph, the area covered under the curve they concluded that Random Forest outperformed other algorithms in terms of predictive classification modeling. On the contrary, the Neural Network did not perform well for both datasets. In conclusion, the study suggests that among the various algorithms, Random Forest stands out as the most promising choice for constructing an effective predictive classification model for the given risky credit prediction task [12].

Tariq et al. conducted a comprehensive study on predicting loan default and developed a methodology that incorporated KDD, CRISP-DM, and SEMMA techniques. After careful consideration and evaluation of various schemes, they selected the most promising approach for estimating loan default in the financial sector. This chosen scheme achieved a precision of 79.8%. However, it was deemed unsuccessful due to its unfavourable ROC score and inadequate performance in the specific area of interest [13]. Obare et al. conducted research focusing on loan default in Kenya, employing a logistic regression model to assess instances of nonpayment for individual loans. The research employed a mathematical analysis procedure that considered the characteristics of borrowers as indicators of loan default. The prototype achieved a precision of 77.27 and 73.33% for the training and test data respectively. The logistic regression model demonstrated a precision of 84.40% for the train statistics and 82.44% for the test statistics. However, a major drawback of this model was its high rate of false positives, indicating a significant number of incorrect predictions for loan nonpayment [14].

In the research conducted by Suliman Mohamed Fati, a comparison is made between three widely recognized machine learning algorithms: logistic regression, Decision Tree and Random Forest. The results strongly indicate that logistic regression outperforms the other two algorithms across various evaluation metrics,

including accuracy, precision, recall, F1 score, and AUC. These findings underscore the importance of carefully selecting the appropriate algorithm for loan status prediction. Logistic regression clearly stands out as the preferred choice due to its exceptional predictive capabilities. The significance of these results further reinforces the value of employing Logistic regression for accurate and reliable loan status predictions [15]. C. N. Sujatha et al. implemented a model for predicting loans using three different algorithms: Decision Tree, logistic regression, and K-Nearest Neighbors (K-NN). The researchers achieved an accuracy of 84.55% using logistic regression. However, the accuracy obtained with Decision Tree and K-NN were 70.73% and 65.04% respectively [16].

Doko et al. explored various machine-learning techniques, including logistic regression, Decision Tree, Random Forest, support vector machines (SVM), and neural network, to classify credit risk data. Their findings revealed that maximum accuracy was achieved using the Decision Tree model, both with and without scaling, when dealing with imbalanced data. The next best performing models were Random Forest and linear regression [17]. Kokate et al. developed an automated credit score classification technique for customers using several machine learning algorithms, comprising gradient boosting, Random Forest, and the integration of feature selection techniques with Decision Trees. The goal was to accurately classify customers as either valid or invalid for loan purposes. In this model, the Decision Tree algorithm attained an accuracy level of 80%. To further improve the accuracy, a gradient boosting model with a voting classifier algorithm was trained and employed. The voting classifier algorithm merged the Decision Tree and gradient boosting techniques, allowing them to work together as an ensemble learner. This ensemble learner employed either a weighted vote with the highest weight or the average of predicted probabilities, referred to as a soft vote, to make predictions regarding the class labels within the dataset [18].

In research carried out by Adebisi et al., an Artificial Neural Network (ANN) algorithm was utilized to create a loan prediction system. The researchers collected user information from Igboora Micro Finance Bank, including the credit history of the users with this bank. The results of their study indicated that the developed system achieved an impressive accuracy rate of 92%. This high accuracy demonstrates the system's ability to effectively anticipate whether a loan applicant is prone to default on repayment or not. Additionally, this prediction system exhibited the capability to identify loans that have a high risk of becoming bad debtor payments. The findings of this study provide evidence of the system's strong predictive performance and its potential to assist in making informed lending decisions [19]. B. R. Puneeth applied various machine learning techniques, including logistic regression, Decision Trees, Random Forest, and XGBoost, to forecast loan-related information. However, regardless of the different methods used, the accuracy achieved so far remains limited to 81.17% [20]. In their research, Luo et al. explored the utilization of a Deep Belief Network (DBN) in conjunction with Restricted Boltzmann Machines to address the credit scoring challenge. They compared the classification performance of DBN with

three other techniques: Multinomial Logistic Regression (MLR), Multilayer Perceptron (MLP), and Support Vector Machine (SVM). The authors utilized a collection of CDS data to evaluate the performance in corporate credit rating [21].

15.3 Proposed Methodology

In this work, we have used a dataset namely “loan-prediction-based-on-customer-behavior” from Kaggle which is in .CSV format. This dataset comprises of 252,000 records and includes 13 significant attributes were employed to address the loan prediction problem, which involves binary classification to determine whether the applicant qualifies for a loan or not. The block diagram of our proposed system for home loan prediction is depicted in Fig. 15.1.

Pre-processing activities were executed, encompassing exploratory data analysis to grasp attribute interconnections, addressing missing values, and detecting and eliminating outliers. This study encompasses thirteen primary explanatory factors. These variables are Id, Income, Age, Experience, Married/Single, House_Ownership, Car_Ownership, Profession, City, State, Current_Job_years, Current_House_Years and Risk_Flag. Label encoder is employed for converting categorical data of Married/Single and Car_Ownership into numerical data while for House_Ownership column one-hot encoder is used to convert it into numerical data. The snapshot of data frame with all the attributes is depicted in Fig. 15.2. Similarly, for Profession, City and State are categorical attributes, they also need to be converted into numerical data. From the data frame it can be observed that Id, Income, Age, Current_Job_years, Current_House_Years and Risk_Flag its type is integer, whereas the remain six attributes namely Experience, Married/Single, House_Ownership, Car_Ownership, Profession, City, State are of object type.

Figure 15.3 illustrates the specifics of the dataset. Figure 15.4 depicts the Marital status of the individuals who applied for loans. It has been observed that out total 252,000 loan customers, 226,272 customers are Single, whereas 25,728 customers are married.

Figure 15.5, shows the House_Ownership status and can be observed that out total 252,000 customers, 231,898 customers lived in rented accommodation, 12,918 customers lived in their owned accommodation and rest 7184 customers are living in norent_noown accommodation. Figure 15.6 shows the Profession details and can be observed that in total there are 51 Professions, along with 5957 Physician and details of other profession are also mentioned. Figure 15.7 shows City details and can be observed that in total there are 317 cities, out of which the city Vijayanagaram has 1259 loan applicants and other details are shared.

Figure 15.8, shows State details and can be observed that that in total there are 29 states and each state e.g. Uttar_Pradesh has 28,400 loan holders. In the next section, we are going to discuss about Principal Component Analysis, Decision Tree classifier and Random Forest.

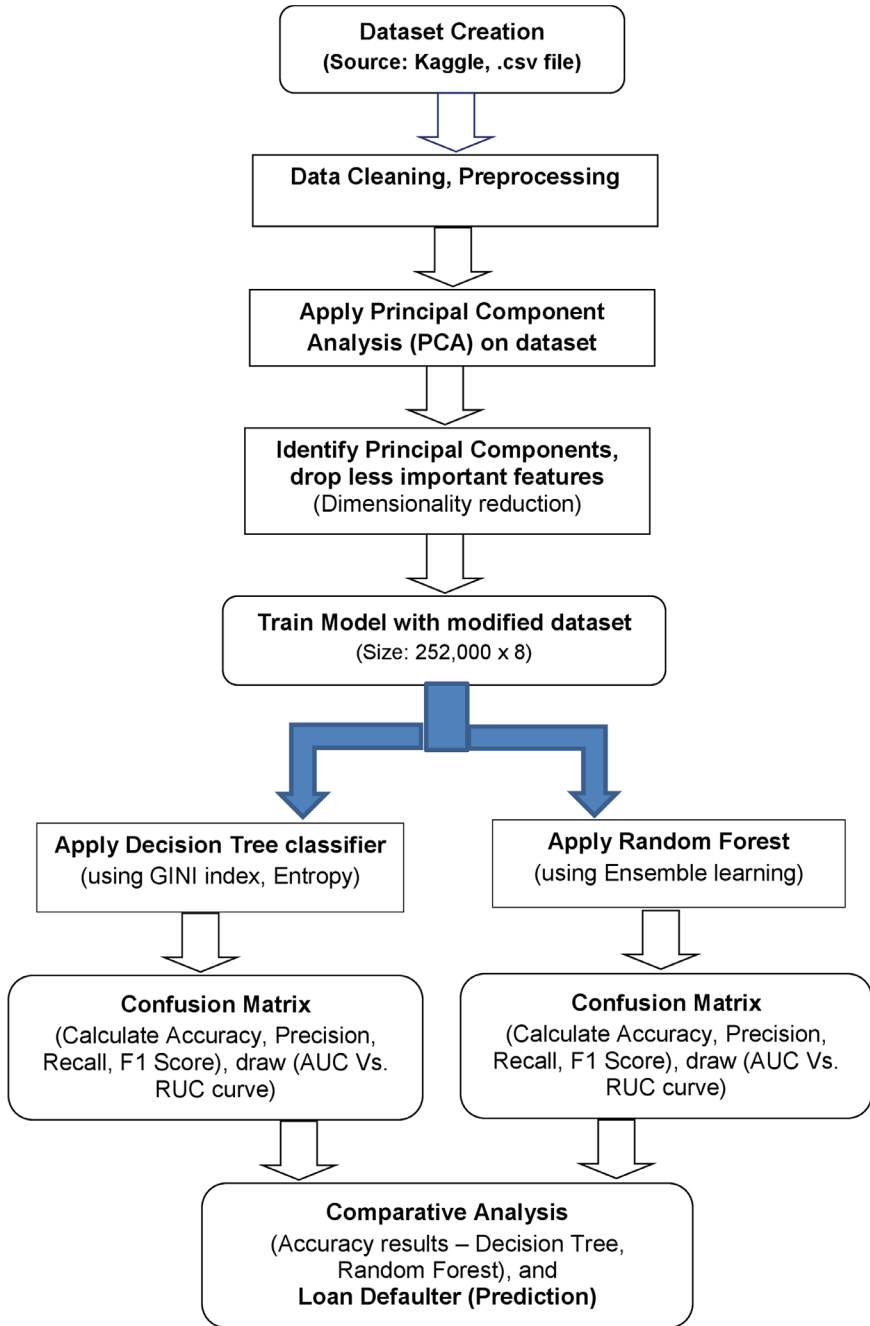


Fig. 15.1 Block diagram of the proposed home loan prediction system

```

↳ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 252000 entries, 0 to 251999
Data columns (total 13 columns):
#   Column                Non-Null Count  Dtype
---  ---                -
0   Id                    252000 non-null int64
1   Income                252000 non-null int64
2   Age                   252000 non-null int64
3   Experience             252000 non-null int64
4   Married/Single        252000 non-null object
5   House_Ownership       252000 non-null object
6   Car_Ownership         252000 non-null object
7   Profession             252000 non-null object
8   CITY                  252000 non-null object
9   STATE                 252000 non-null object
10  CURRENT_JOB_YRS       252000 non-null int64
11  CURRENT_HOUSE_YRS    252000 non-null int64
12  Risk_Flag             252000 non-null int64
dtypes: int64(7), object(6)
memory usage: 25.0+ MB

```

Fig. 15.2 Snapshot of data frame with all the attribute

15.4 Model Development

All experiments conducted in this work were done on a host with an AMD Ryzen 5 5600U CPU with Radeon graphics 2.30 GHz, 16 GB of RAM/512 GB SSD with \times 64-based processor. The software environment is performed under Windows 11 OS, using Pandas, Matplotlib, Sklearn and Graphviz framework which corresponds to Python 3.10.12 version.

15.4.1 Principal Component Analysis

Principal Component Analysis (PCA) is a frequently used method to reduce the dimensions of large datasets. Its objective is to convert a collection of multiple variables into a smaller subset that maintains a significant portion of the pertinent information found in the initial dataset. Although dimensionality reduction sacrifices a certain degree of accuracy, the key concept is to trade some accuracy for simplicity. By reducing the number of variables, PCA facilitates easier exploration, visualization, and analysis of data. It also improves the efficiency of machine learning algorithms by eliminating extraneous variables. As the number of principal components matches the count of variables in the data, these components are constructed in a way that the first component captures the maximum possible variance present in the dataset. In summary, PCA aims to reduce the quantity of variables within a dataset while retaining a significant amount of valuable information. For this work, a

Id	Income	Age	Experience	Married/Single	House_Ownership	Car_Ownership	Profession	CITY	STATE	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS	
0	1	1303834	23	3	single	rented	no	Mechanical_engineer	Rewa	Madhya_Pradesh	3	13
1	2	7574516	40	10	single	rented	no	Software_Developer	Parbhani	Maharashtra	9	13
2	3	3991815	66	4	married	rented	no	Technical_writer	Alappuzha	Kerala	4	10
3	4	6256451	41	2	single	rented	yes	Software_Developer	Bhubaneswar	Odisha	2	12
4	5	5768871	47	11	single	rented	no	Civil_servant	Truchirappalli[10]	Tamil_Nadu	3	14



Fig. 15.3 The details of the dataset

Fig. 15.4 Marital status

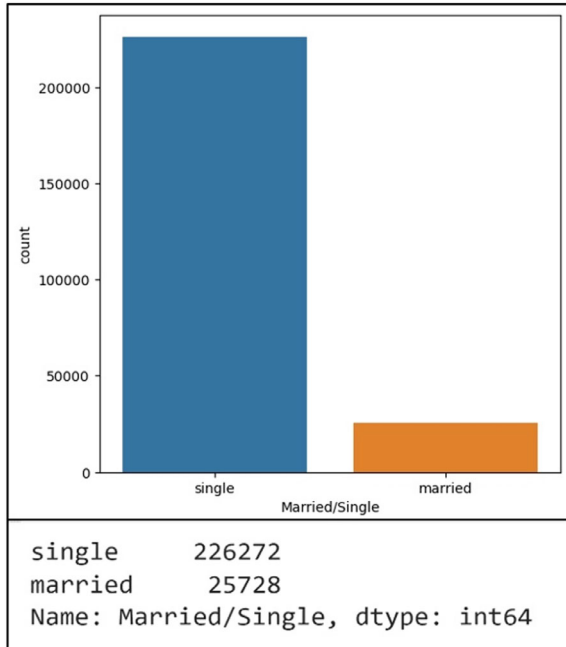


Fig. 15.5 House_ Ownership status

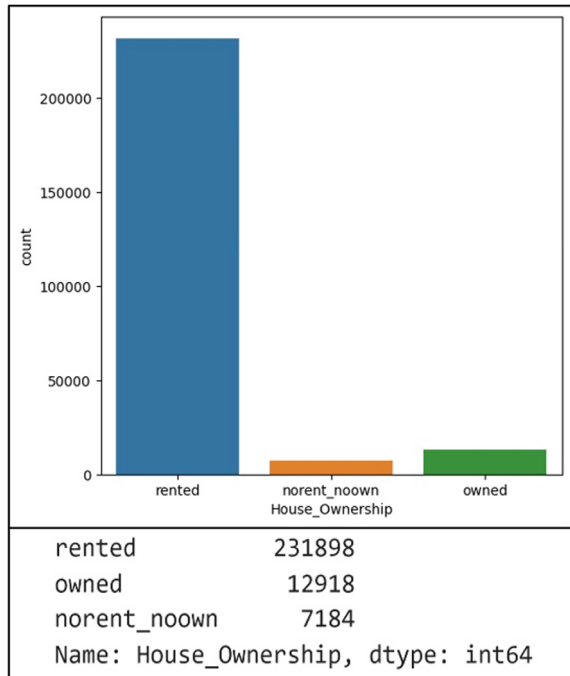


Fig. 15.6 Profession details

Physician	5957
Statistician	5806
Web_designer	5397
Psychologist	5390
Computer_hardware_engineer	5372
Drafter	5359
Magistrate	5357
Fashion_Designer	5304
Air_traffic_controller	5281
Comedian	5259
Industrial_Engineer	5250
Mechanical_engineer	5217
Chemical_engineer	5205
Technical_writer	5195
Hotel_Manager	5178
Financial_Analyst	5167
Graphic_Designer	5166
Flight_attendant	5128
Biomedical_Engineer	5127
Secretary	5061
Software_Developer	5053
Petroleum_Engineer	5041
Police_officer	5035
Computer_operator	4990
Politician	4944
Microbiologist	4881

dataset named “loan-prediction-based-on-customer-behavior” in CSV format which is available on Kaggle is used. This dataset comprises 252,000 records and includes 13 significant attributes as shown in Fig. 15.2.

Step 1: Standardization

The initial step involves standardizing the range of continuous variables. Prior to employing PCA, standardization is vital due to the technique’s sensitivity to the variances inherent in the original variables. If there are significant differences in the ranges of the initial variables, those with larger ranges will have a greater influence compared to those with smaller ranges. This dominance of variables with larger ranges can introduce bias and result in skewed outcomes.

Fig. 15.7 City details

```

↳ Total categories in CITY: 317

Vijayanagaram      1259
Bhopal              1208
Bulandshahr         1185
Saharsa[29]         1180
Vijayawada          1172
...
Ujjain              486
Warangal[11][12]   459
Bettiah[33]         457
Katni                448
Karaikudi            431
Name: CITY, Length: 317, dtype: int64

```

Fig. 15.8 State details

```

Total categories in STATE: 29

Uttar_Pradesh      28400
Maharashtra         25562
Andhra_Pradesh     25297
West_Bengal        23483
Bihar               19780
Tamil_Nadu         16537
Madhya_Pradesh     14122
Karnataka           11855
Gujarat             11408
Rajasthan           9174
Jharkhand           8965
Haryana             7890
Telangana           7524
Assam               7062
Kerala              5805
Delhi               5490
Punjab              4720
Odisha              4658
Chhattisgarh       3834
Uttarakhand         1874
Jammu_and_Kashmir  1780
Puducherry          1433
Mizoram             849

```

The equation used to calculate the standard deviation is shown in Eq. (15.1), z-score is shown in Eq. (15.2):

$$standard\ deviation = \sqrt{\frac{\sum(x_i - x)^2}{N}} \tag{15.1}$$

where,

- x_i = data values in the set.
- x = mean of the data.
- N = number of data values

$$z = \frac{value - mean}{standard\ deviation} \tag{15.2}$$

Hence, the transformation of data onto comparable scales through standardization helps mitigate this issue and ensures fair representation of all variables in the PCA analysis as shown in Fig. 15.9. During training the Risk_Flag has been dropped from the dataset as it is an independent variable.

The ‘Label_Encoder’ is used to transform categorical variables ‘Married/Single’ and ‘Car_Ownership’ into numerical labels. This transformation likely maps each unique category in these columns to an integer. The ‘OneHotEncoder’ is used to perform one-hot encoding on the ‘House_Ownership’ column. It is a technique used to convert categorical variables as binary vectors, where each vector becomes a binary feature column.

Step 2: Covariance Matrix Computation

The purpose of computation of the covariance matrix is carried out to assess the manner in which variables within the dataset vary in relation to each other, with the goal of identifying any relationships or correlations between them. This analysis helps to determine if certain variables contain redundant or overlapping information due to high correlation. Insights into the interdependencies and patterns among the variables

	Income	Age	Experience	Married/Single	House_Ownership	Car_Ownership	CURRENT_JOB_YRS	CURRENT_HOUSE_YRS	Profession_count	CITY_count	STATE_count
0	1303834	23	1303834	3	1	0.0	0	3	5217	798	14122
1	7574516	40	10	1	0.0	0	9	13	5053	849	25562
2	3991815	66	4	0	0.0	0	4	10	5195	688	5805
3	6256451	41	2	1	0.0	1	2	12	5053	607	4658
4	5768871	47	11	1	0.0	0	3	14	4413	809	16537
...
251995	8154883	43	13	1	0.0	0	6	11	4772	1033	23483
251996	2843572	26	10	1	0.0	0	6	11	4661	798	14122
251997	4522448	46	7	1	0.0	0	7	12	4729	741	25562
251998	6507128	45	0	1	0.0	0	0	10	5166	897	1433
251999	9070230	70	17	1	0.0	0	7	11	5806	667	16537

252000 rows × 11 columns

Fig. 15.9 Data transformation

can be gained by calculating the covariance matrix, enabling the identification and understanding of the underlying relationships within the data.

Step 3: Computation of Eigenvectors and Eigenvalues

After computing covariance matrix, its eigenvectors and eigenvalues are calculated to identify the principal components. The eigenvectors signify the directions or axes in the dataset space, while the eigenvalues representing the extent of variance explained by each eigenvector. The eigenvectors are sorted based on their corresponding eigenvalues in descending order. This sorting ensures that the principal components, which capture the most variance, are ranked first. The desired number of principal components is chosen based on the amount of variance one aims to preserve within the dataset.

Typically, the top-k eigenvectors that explain a significant portion of the total variance are selected. The data is transformed by performing a matrix multiplication of the original data with the selected eigenvectors. This transformation projects the data onto the new coordinate system defined by the principal components as shown in Fig. 15.10.

Step 4: Feature Vector

The feature vector, as mentioned refers to a matrix comprising of the remaining principal components, after discarding certain components with low significance. These remaining components capture the most significant variance in the dataset and are considered important for further analysis or modeling.

At this stage, a choice is made whether to retain all of these components or eliminate those with lower significance (manifesting as low eigenvalues), resulting

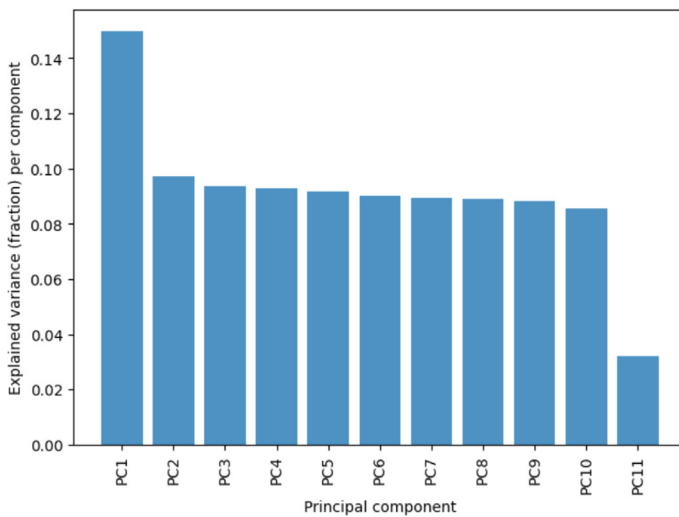


Fig. 15.10 Plot showing principal component versus explained variance

in the creation of a vector matrix comprising the remaining ones, referred to as the feature vector. Here we have decided to discard three components i.e. City_count, State_count and Profession_count from the dataset, as these features are less important and doesn't contribute much as they have low eigen values. In this work, we retain the remaining principal components as our feature vector.

Step 5: Transform the Data Along the Axes of the Principal Components

In this step, the aim is to use the feature vector formed using the eigenvectors of the covariance matrix, and to reorient the data from the original axes to the ones represented by the principal components, hence the name Principal Component Analysis. This can be accomplished by performing multiplication of the transpose of the initial dataset, with the transpose of the feature vector, as demonstrated in Eq. (15.3).

$$Feature = FeatureVector^T \times StandardizedOriginalDataSet^T \quad (15.3)$$

StandardizedOriginalDataSet refers to the original dataset after it has been standardized. Standardization is a preprocessing step mostly used in machine learning and data analysis, and to transform the features of the dataset to obtain a mean value of 0 and a standard deviation of 1. This raw dataset containing our observations and features, which include columns such as income, age, experience, marital status, etc.

The important features are shown in Fig. 15.11, now the dataset comprises of 252,000 entries along with 8 columns i.e. important features which is used for training the proposed model.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 252000 entries, 0 to 251999
Data columns (total 8 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Income                252000 non-null  float64
1   Age                   252000 non-null  float64
2   Experience             252000 non-null  float64
3   Married/Single        252000 non-null  float64
4   House_Ownership       252000 non-null  float64
5   Car_Ownership         252000 non-null  float64
6   CURRENT_JOB_YRS       252000 non-null  float64
7   CURRENT_HOUSE_YRS    252000 non-null  float64
dtypes: float64(8)
memory usage: 15.4 MB
```

Fig. 15.11 Dataset with 8 features

15.4.2 *Decision Tree*

Decision Trees find application as versatile tools utilized for tasks involving both classification and regression. These structures resemble like an inverted tree-like flowchart, with root node at the top and subsequent nodes branching out through feature-based splits. This tree-like structure allows for easy visualization of the decision-making process. In practical terms, Decision Trees can be thought of as a collection of if-else statements. At each node, a condition is evaluated, and based on the result, the tree follows the appropriate branch to the next connected node. This process continues until a leaf node is reached, providing us the final decision or prediction based on the path followed through the tree as shown in Fig. 15.12, i.e. Decision Tree using Gini index.

In Figs. 15.12 and 15.13, the blocks serve as visual aids to illustrate the hierarchical structure of decision trees and how nodes are separated within the tree. Each block represents a node in the decision tree, with lines connecting them to show the flow of decision-making, starting from the root node until it reaches to the leaf nodes. The blocks in the figures represent decision nodes where specific conditions or features are evaluated to determine the path the tree takes. These conditions help to partition the data into subsets depending on the values of the features. The separation of nodes in the tree reflects, how the decisions are made based on the features of the dataset.

In Fig. 15.12, the Decision Tree using the Gini index, likely employs the Gini impurity measure to evaluate the purity of a node. The Gini index which measures the probability of incorrectly classifying a randomly chosen element, if they were randomly classified based on the distribution of labels in the node. It is commonly used in decision tree algorithms, especially in classification tasks.

In Fig. 15.13, the Decision Tree was generated using Information Gain and Entropy, which utilizes information gain and entropy measures to decide the splitting criteria at each node. Information gain quantifies the decrease in entropy achieved by partitioning the data based on a particular feature. Entropy, in this context, represents the uncertainty or disorder in the dataset. The decision tree algorithm aims to minimize entropy at each split, resulting in more homogeneous subsets.

The objective of machine learning is to minimize uncertainty and disorder within datasets, and Decision Trees are employed for this purpose. Entropy quantifies the uncertainty or disorder present in a dataset. By using entropy, the impurity of a specific node can be assessed. However, it remains unclear whether the entropy of the parent node or a specific node has decreased. To address this, a new metric called “Information Gain” is introduced as shown in Fig. 15.13, i.e. Decision Tree using Information Gain and Entropy. Information Gain quantifies the decrease in entropy achieved through data division based on a specific feature. It serves as a decisive factor in selecting the attribute to be used, as the decision node or root node in the Decision Tree. Information Gain provides insights into how much the parent node’s entropy has diminished after the split, helping determine the optimal feature for creating decision boundaries and improving overall predictive accuracy.

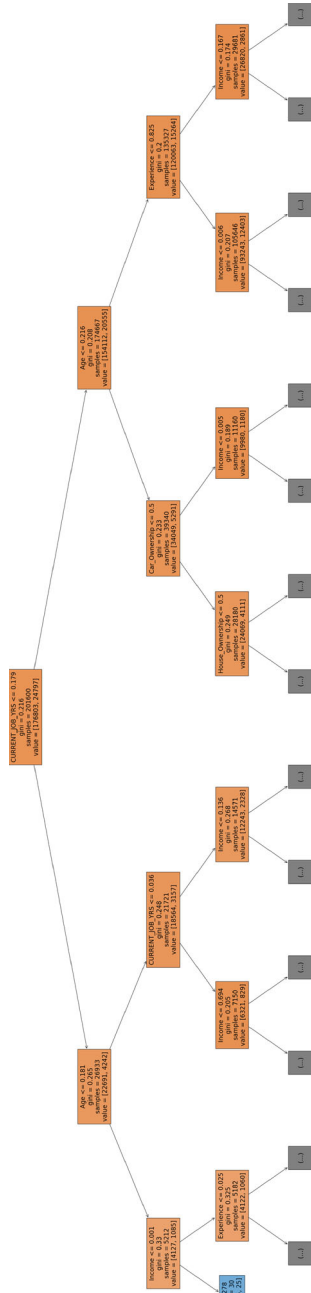


Fig. 15.12 Decision tree using Gini index



Fig. 15.13 Decision tree using information gain and entropy

The ID3 algorithm, which stands for Iterative Dichotomiser, utilizes entropy to determine Information Gain. The Information Gain $IG(A)$ of an attribute A with respect to a set S is shown in Eq. (15.4) and calculated as:

$$IG(A) = Entropy(S) - \sum (|S_v|/|S|) * Entropy(S_v) \quad (15.4)$$

where S_v is the subset of examples in S for which attribute A has the value v . The summation is taken over all possible values v of attribute A . By calculating the Information Gain for each attribute, the ID3 algorithm identifies the attribute that provides the most useful and discriminative information for splitting the data. This attribute is chosen as the splitting attribute, and the tree is recursively built by repeating this process for the subsets of examples resulting from the split.

The choice between using the Gini index or information gain/entropy, which depends on the specific requirements of the problem and the available characteristics of the dataset. Both metrics aim to minimize impurity in the resulting nodes, but they may behave differently under various conditions. Therefore, the depiction of decision trees using different metrics in Figs. 15.12 and 15.13 allows for a comparative understanding of how different splitting criteria influence the structure and decision-making process of the tree. It provides insights into the trade-offs and considerations involved in selecting appropriate metrics for constructing decision trees based on the dataset and problem domain.

Advantages of using Decision Tree is that, it is easy to interpret, and computationally efficient, which can handle both numerical and categorical data.

Disadvantages are prone to overfitting, sensitive to small variations in the data, lack of robustness.

15.4.3 *Random Forest*

In the realm of data science, Random Forest stands out as a widely favored and extensively employed algorithm. As a supervised machine learning algorithm, it finds extensive usage in both classification and regression tasks. This algorithm functions by creating multiple Decision Trees on different samples from the dataset, and by then combining their predictions through majority voting for classification or averaging for regression. The working principle of Random Forest is shown in Fig. 15.14. A key advantage of the Random Forest algorithm is its ability to handle datasets, which comprises of both continuous and categorical variables. This flexibility of RF makes it suitable for a diverse range of applications, as it can effectively handle various types of data. In classification problems, it can accommodate categorical variables, while in regression tasks, it can handle continuous variables. This versatility contributes to the widespread adoption and effectiveness of Random Forest in real-world scenarios. The Random Forest algorithm in machine learning operates on the principle of ensemble learning. Ensemble learning involves combining multiple

models to make predictions, as opposed to relying on a single model. In case of the Random Forest algorithm, a collection or ensemble of models, specifically Decision Trees, is utilized to generate predictions. Each Decision Tree within the Random Forest is trained using distinct subsets of the dataset, ensuring diversity among the models. When making the predictions, the Random Forest algorithm combines the predictions from all the individual Decision Trees through methods such as majority voting in classification or averaging in regression. This process can be divided into two stages. In the initial stage, a random selection of “k” features is made from a pool of “m” total features to construct the Random Forest. In this stage, the following steps are followed:

Stage1: K features are selected among m features where $k < m$. Subsequently, node “d” is determined by utilizing the optimal split obtained from the selected “k” features. This node is split into daughter nodes using the best split. Then the forest is constructed by iteratively repeating the aforementioned steps “n” times, resulting in the creation of “n” trees.

Stage2: Test features are utilized with the decision rules from each randomly generated Decision Tree to foresee the outcome. These predicted outcomes are then recorded and saved. After that the total number of votes for each predicted target is computed. Ultimately, the predicted target with the highest number of votes is adopted as the final output generated by the Random Forest (RF) algorithm. Figure 15.14 depicts the Decision Tree 1 of Random Forest, where Current_Job_Yrs has been selected as a root node for splitting the data.

Figure 15.15 depicts the Decision Tree 2 of Random Forest, where experience has been selected as a root node and Fig. 15.16 depicts the Decision Tree 3 of Random Forest, where Income has been selected as a root node for splitting the data and tree generation. By leveraging the collective wisdom of multiple models, the Random Forest algorithm (RF) enhances prediction accuracy and robustness. The ensemble nature of the Random Forest enables it to capture a wider range of patterns and effectively handle complex datasets. This approach has made the Random Forest

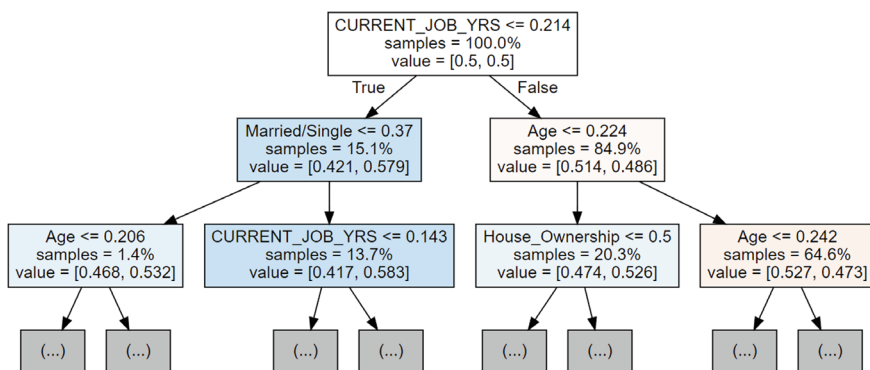


Fig. 15.14 Decision Tree 1 of Random forest

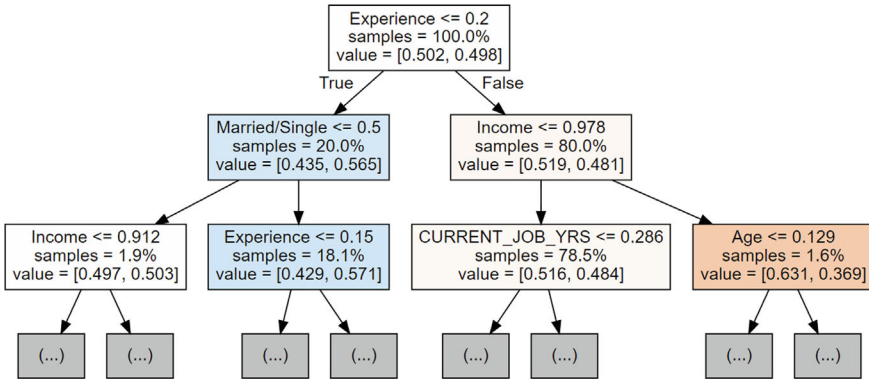


Fig. 15.15 Decision tree 2 of Random forest

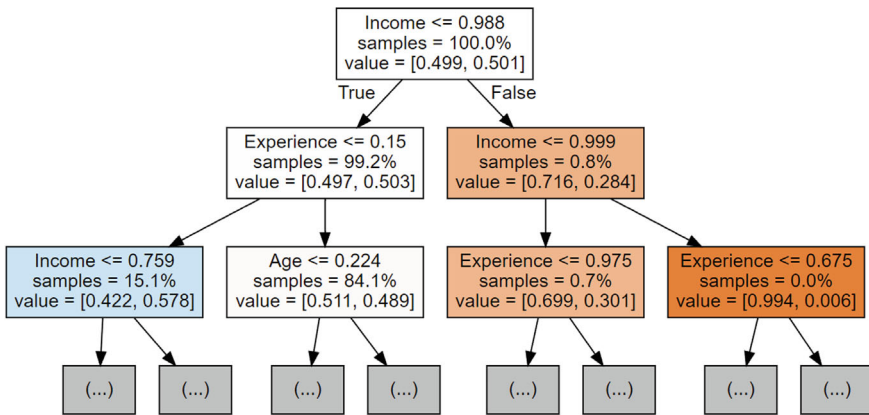


Fig. 15.16 Decision tree 3 of Random forest

algorithm a powerful tool in machine learning with numerous applications across various domains.

Advantage of using Random Forest are reduced overfitting compared to individual decision trees, improved accuracy, robust to noise and outliers. Disadvantage include less interpretable than individual decision trees, higher computational complexity.

15.5 Results and Discussion

A Confusion Matrix serves as a valuable tool in machine learning, for assessing the effectiveness of a classification model. It provides a breakdown of true positives, true negatives, false positives, and false negatives, offering deep insight into the

Fig. 15.17 Confusion matrix for binary classification

		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

model’s performance and the potential for enhancing predictive accuracy. Essentially, a Confusion Matrix takes the form of an $N \times N$ matrix, where N represents the total number of target classes within the classification problem. The matrix serves as a means to compare the model’s predictions with the actual target values, offering a comprehensive perspective on the classification model’s overall performance and the types of errors it may be prone to. In cases of binary classification, such as a “yes” or “no” scenario, a 2×2 matrix is employed, as shown in Fig. 15.17, with 4 values:

We can notice that the target variable encompasses of two distinct values: Positive and Negative. The columns of the Confusion Matrix correspond to the actual values of the target variable, while the rows correspond to the predicted values of the target variable. In this context, let’s elaborate on the significance of TP (True Positives), FP (False Positives), FN (False Negatives), and TN (True Negatives), which are pivotal terms within a Confusion Matrix.

True Positive (TP)

This occurs when the model’s prediction matches with the actual value or class. In the context of binary classification, it means that the actual value was positive, and the model correctly predicted it as positive.

True Negative (TN)

This happens when the model’s prediction aligns with the actual value or class. Specifically, in binary classification, it signifies that the actual value was negative, and the model accurately predicted it as negative.

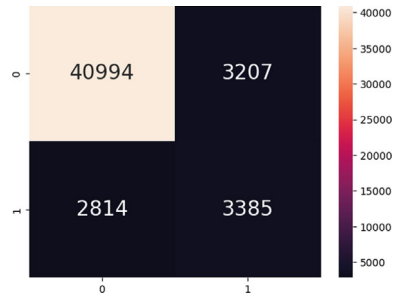
False Positive (FP)—Type I Error

FP takes place when the model falsely predicts a positive value. In binary classification, this translates that the actual value being negative, but the model incorrectly predicting it as positive. FP is also referred to as Type I Error, indicating an erroneous positive prediction.

False Negative (FN)—Type II Error

FN occurs when the model erroneously predicts a negative value. In binary classification, this means that the actual value was positive, but the model incorrectly predicted it as negative. FN is known as Type II Error, signifying an incorrect negative prediction. It can be observed from Fig. 15.18 i.e. the Confusion Matrix of Decision Tree which indicates that there are 40,994 TP values, 3385 TN values, 3207 FP values and 2814 FN values.

Fig. 15.18 Confusion matrix of decision tree



It can be observed from Fig. 15.19 i.e. the Confusion Matrix of Random Forest which indicates that there are 39,785 TP values, 4839 TN values, 4416 FP values and 1360 FN values.

The accuracy of the model can be calculated using the formula as mentioned in Eq. (15.5).

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN}) \tag{15.5}$$

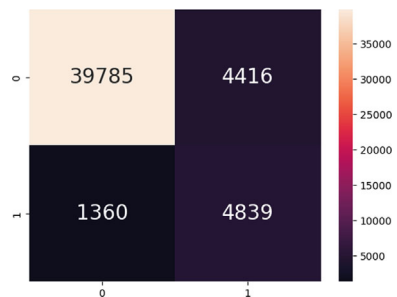
In addition to this, it is important to calculate the precision and recall as shown in Eq. (15.6), and Eq. (15.7). Precision tells us, how many of the correctly predicted cases actually turned out to be positive.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \tag{15.6}$$

This would determine whether our model is reliable or not. Recall tells us, how many of the actual positive cases we are able to predict correctly with our model. And here’s how we can calculate Recall:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \tag{15.7}$$

Fig. 15.19 Confusion matrix of random forest



In practice, when we try to increase the precision of our model, the recall goes down, and vice-versa. The F1-score captures both the trends in a single value as shown in Eq. (15.8).

$$F1\text{-score} = 2/(1/Recall + 1/Precision) \tag{15.8}$$

The F1-score is a statistical metric that serves as the harmonic mean of precision and recall. It offers a consolidated perspective on these two critical performance metrics and reaches its maximum value when precision equals recall. However, when the interpretability of the F1-score is lacking, it implies that it becomes challenging to discern whether our classifier is prioritizing precision or recall. In such cases of poor interpretability, it becomes unclear whether the classifier is favoring precision by minimizing the false positives or recall by minimizing the false negatives. This ambiguity can make it difficult to make informed decisions about the classifier’s performance, as we cannot precisely determine whether it’s striking the right balance between precision and recall or leaning more towards one of these metrics. Therefore, a clear understanding of the context and goals of the classification problem is essential to choose the appropriate evaluation metric and interpret the F1-score effectively. Figure 15.20, shows the classification report of Decision Tree classifier with an accuracy of **88%**. Figure 15.21, shows the classification report of Random Forest classifier with an accuracy of **89%**.

Decision Trees (DT) and Random Forest (RF) are popular algorithms available in machine learning, which are especially used for classification tasks. While Decision Trees offer simplicity and interpretability, Random Forest enhances predictive performance through ensemble learning.

Upon comparing the accuracy metrics, Random Forest generally outperforms Decision Trees due to its ensemble nature. Precision, recall, and F1-score metrics further validate the superiority of Random Forest in capturing both positive and negative cases effectively. Random Forest’s ensemble approach mitigates overfitting and variance, contributing to its superior performance. The aggregation of multiple decision trees leads to improved generalization and robustness in handling complex datasets and diverse patterns.

The Classification report :

	precision	recall	f1-score	support
0.0	0.94	0.93	0.93	44201
1.0	0.51	0.55	0.53	6199
accuracy			0.88	50400
macro avg	0.73	0.74	0.73	50400
weighted avg	0.88	0.88	0.88	50400

Fig. 15.20 Classification report of decision tree classifier

```

The Classification report :
              precision    recall  f1-score   support

     0.0         0.97      0.90      0.93     44201
     1.0         0.52      0.78      0.63     6199

 accuracy                   0.89     50400
 macro avg              0.74      0.84      0.78     50400
 weighted avg           0.91      0.89      0.89     50400

```

Fig. 15.21 Classification report of random forest classifier

Finally, in this work comparative analysis has been done by comparing the performance of Decision Tree classifier with Random Forest using a receiver operating characteristic curve (ROC curve), that shows how well a classification model performs, which helps us to see how the model makes decisions at different levels of certainty. The receiver operating characteristic (ROC) curve is a graphical representation of how well a classification model performs, providing insights into the model's decision-making process across different confidence levels. It is a valuable evaluation metric primarily used in case of binary classification problems. The ROC curve is essentially a probability curve that plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold values. Its main purpose is to effectively differentiate the 'signal' from the 'noise' in classification outcomes. In simpler terms, the ROC curve illustrates how well a classification model distinguishes between positive and negative cases across a range of decision thresholds.

The Area under the Curve (AUC) is a numerical measure derived from the receiver operating characteristic (ROC) curve as shown in Fig. 15.22. It quantifies the classifier's ability to discriminate between the two classes, serving as a concise summary of the ROC curve's performance. A higher AUC value indicates better overall model performance in binary classification, with a value of 1 indicating perfect discrimination, and a value of 0.5 indicating no discrimination (equivalent to random guessing).

AUC values range from 0 to 1, with higher values indicating superior model performance in distinguishing between positive and negative classes. An AUC value of 1, signifies that the classifier can flawlessly differentiate between the two classes. Conversely, an AUC value of 0 implies that it cannot make this distinction, effectively predicting all positives as negatives and vice versa. When AUC falls within the range of 0.5 to 1, it demonstrates the classifier's capacity to effectively discern positive and negative class values. Typically, it detects more True Positives and True Negatives than False Positives and False Negatives. An AUC value of 0.5 suggests that the classifier performs no better than random guessing, indicating it either predicts randomly or assigns a constant class to all data points.

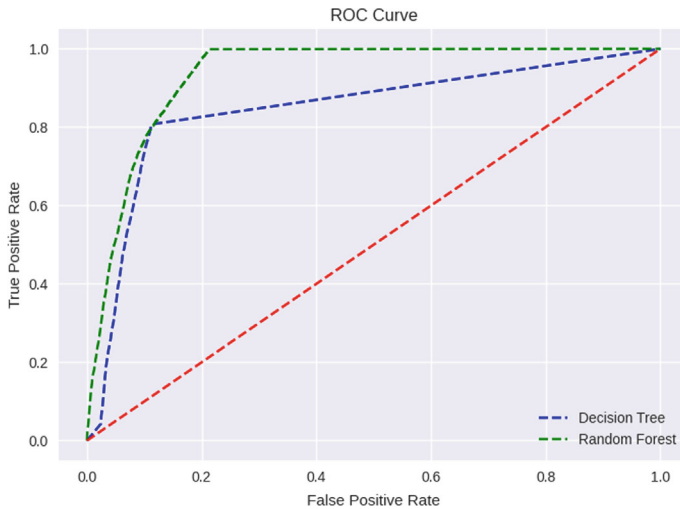


Fig. 15.22 ROC curve

Optimizing the threshold in classification models allows for customization based on specific requirements, and trade-offs between False Positives and False Negatives. Adjusting the threshold enables fine-tuning the model's behavior to align with the desired objectives and risk tolerance. By exploring different threshold values, stakeholders can evaluate the trade-offs between precision and recall according to the application context. For instance, in loan default prediction, prioritizing recall may lead to more conservative lending decisions, while emphasizing precision could reduce the risk of misclassifying non-defaulters.

15.6 Concluding Remarks

In a ROC curve, the choice of threshold is critical, as it determines the balance between False Positives (FP) and False Negatives (FN). The threshold can be adjusted to optimize the model's performance based on the specific problem requirements and the desired trade-off between these two types of errors. It can be observed from Fig. 15.22, i.e. ROC curve plot shows that Random Forest (RF) performs better than the Decision Tree (DT).

Finally, the comparative analysis has been done, which are based on the prediction results of both the algorithms i.e. Decision Tree (DT) classifier and Random Forest (RF), along with its accuracy and error rate. Based on the accuracy of the proposed developed system, it may be integrated in the banking industry in Indian subcontinent, as the dataset used for training and testing the model, contains the data from the Indian banking sector.

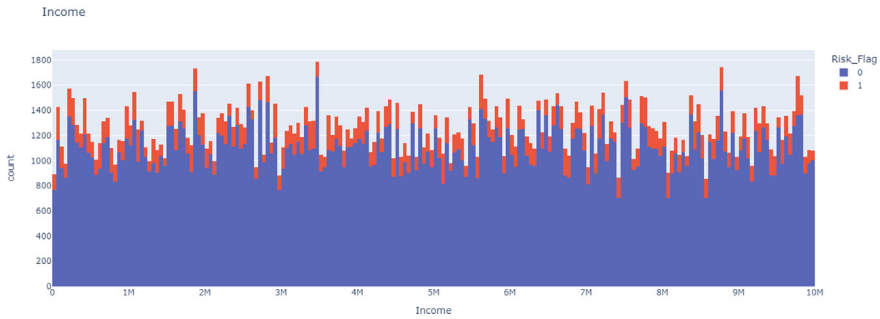


Fig. 15.23 Income versus Risk_Flag

Figure 15.23 shows the plot of Income versus Risk_Flag for the dataset used in this work. It can be observed from the data that the customers who defaulted in Loan payments, out of which top 5 defaulters are the customers who belongs to high income group. Financial institution's must be very conscious with customers of high-income group before lending any loans, and may use our developed system for predicting which customer may default on a loan before lending.

Examining the relationship between income levels and loan default risk reveals insights into customer behavior and financial risk dynamics. Factors such as income stability, debt levels, and spending patterns influence the likelihood of loan default, requiring nuanced risk assessment strategies. Financial institutions can refine risk assessment strategies by considering income-related variables alongside other demographic and financial indicators. By segmenting customers based on income tiers, lenders can tailor risk scoring models and lending criteria to mitigate default risks effectively.

Integrating predictive models into the banking industry offers opportunities for enhancing risk management practices and improving decision-making processes. By leveraging advanced analytics, financial institutions can optimize loan underwriting, credit risk assessment, and portfolio management strategies. The predictive model can support various applications within financial institutions, including loan approval processes, credit scoring, and portfolio optimization. By incorporating predictive analytics into operational workflows, banks can streamline processes, minimize risks, and enhance profitability.

Future research can focus on advancing predictive modeling techniques, exploring innovative features, and incorporating external data sources to enhance model accuracy and predictive power. Investigating novel algorithms and methodologies can address existing limitations and drive innovation in credit risk management. Recognizing the limitations inherent in the study, such as data biases, model assumptions, and scalability concerns are crucial for ensuring the reliability and applicability of the predictive models in real-world settings. Addressing these limitations through rigorous validation and sensitivity analysis strengthens the credibility and robustness of predictive models.

Finally, based on the comparison results and the analysis of advantages and disadvantages, it offers clear recommendations for the use of decision tree and random forest methods in identifying loan defaulters. By considering the factors such as dataset characteristics, computational resources, interpretability requirements, and the desired level of predictive accuracy are important factors in this context. Providing guidance on when to prefer one method over the other, or when to use them in combination for improved performance.

By addressing these aspects comprehensively, the analysis provides valuable insights into the performance, implications, and future directions of predictive modeling in the banking industry.

References

1. Aslam, U., Tariq Aziz, H.I., Asim, S., Kadhar, B.N.: An empirical study on loan default prediction models. *J. Comput. Theor. Nanosci.* **16**(8), 3483–3488 (2019). <https://doi.org/10.1166/jctn.2019.8312>
2. Anand, M., Velu, A., Whig, P.: Prediction of loan behaviour with machine learning models for secure banking. *J. Comput. Sci. Eng. (JCSE)*. **3**(1), 1–13 (2022). <https://doi.org/10.36596/jcse.v3i1.237>
3. Murthy, P.S., Shekar, G.S., Rohith, P., Reddy, G.V.V.: Loan approval prediction system using machine learning. *J. Innov. Inf. Technol.* **4**(1), 21–24 (2020)
4. Chen, H.: Prediction and analysis of financial default loan behavior based on machine learning model. *Comput. Intell. Neurosci.* (2022). <https://doi.org/10.1155/2022/7907210>
5. Patel, B., Patil, H., Hembram, J., Jaswal, S.: Loan default forecasting using data mining. In: International Conference for Emerging Technology (INCET), pp. 1–4. IEEE (2020). <https://doi.org/10.1109/INCET49848.2020.9154100>
6. Lee, T.S., Chiu, C.C., Chou, Y.C., Lu, C.J.: Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Comput. Stat. Data Anal.* **50**(4), 1113–1130 (2006). <https://doi.org/10.1016/j.csda.2004.11.006>
7. Ince, H., Aktan, B.: A comparison of data mining techniques for credit scoring in banking: a managerial perspective. *J. Bus. Econ. Manag.* **10**(3), 233–240 (2009). <https://doi.org/10.3846/1611-1699.2009.10.233-240>
8. Reddy, M. J., Kavitha, B.: Neural networks for prediction of loan default using attribute relevance analysis. In: International Conference on Signal Acquisition and Processing, pp. 274–277. IEEE (2010). <https://doi.org/10.1109/ICSAP.2010.10>
9. Odeh, O., Koduru, P., Featherstone, A., Das, S., Welch, S. M.: A multi-objective approach for the prediction of loan defaults. *Exp. Syst. Appl.* **38**(7), 8850–8857 (2011). <https://doi.org/10.1016/j.eswa.2011.01.096>
10. Zhou, L., Wang, H.: Loan default prediction on large imbalanced data using random forests. *TELKOMNIKA Indonesian J. Electric. Eng.* **10**(6), 1519–1525 (2012)
11. Kemalbay, G., Korkmazoğlu, Ö.B.: Categorical principal component logistic regression: a case study for housing loan approval. *Procedia Soc. Behav. Sci.* **109**, 730–736 (2014). <https://doi.org/10.1016/j.sbspro.2013.12.537>
12. Gahlaut, A., Singh, P. K.: Prediction analysis of risky credit using data mining classification models. In: 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–7. IEEE (2017). <https://doi.org/10.1109/ICCCNT.2017.8203982>
13. Tariq, H.I., Sohail, A., Aslam, U., Batcha, N.K.: Loan default prediction model using sample, explore, modify, model, and assess (SEMMA). *J. Comput. Theor. Nanosci.* **16**(8), 3489–3503 (2019). <https://doi.org/10.1166/jctn.2019.8313>

14. Obare, D.M., Njoroge, G.G., Muraya, M.M.: Analysis of individual loan defaults using logit under supervised machine learning approach. *Asian J. Prob. Stat.* **3**(4), 1–12 (2019)
15. Fati, S. M.: Machine learning-based prediction model for loan status approval. *J. Hunan Univ. Nat. Sci.* **48**(10) (2021)
16. Sujatha, C. N., Gudipalli, A., Pushyami, B., Karthik, N., Sanjana, B. N.: Loan prediction using machine learning and its deployment on web application. In: *Innovations in Power and Advanced Computing Technologies (i-PACT)*, pp. 1–7. IEEE (2021). <https://doi.org/10.1109/i-PACT52855.2021.9696448>
17. Doko, F., Kalajdziski, S., Mishkovski, I.: Credit risk model based on central bank credit registry data. *J. Risk Financ. Manag.* **14**(3), 138 (2021). <https://doi.org/10.3390/jrfm14030138>
18. Kokate, S., Chetty, M. S. R.: Credit risk assessment of loan defaulters in commercial banks using voting classifier ensemble learner machine learning model. *Int. J. Saf. Secur. Eng.* **11**(5), 565–572 (2021). <https://doi.org/10.18280/ijssse.110508>
19. Adebisi, M.O., Adeoye, O.O., Ogundokun, R.O., Okesola, J.O., Adebisi, A.A.: Secured loan prediction system using artificial neural network. *J. Eng. Sci. Technol.* **17**(2), 0854–0873 (2022)
20. Puneeth, B. R., Ashwitha, K., Kumar, A., Rao, B., Saliya, P., Supravi, A. P.: An approach to predict loan eligibility using machine learning. In: *International Conference on Artificial Intelligence and Data Engineering (AIDE)*, pp. 3–28. IEEE (2022). <https://doi.org/10.1109/AIDE57180.2022.10059881>
21. Luo, C., Wu, D., Wu, D.: A deep learning approach for credit scoring using credit default swaps. *Eng. Appl. Artif. Intell.* **65**, 465–470 (2017). <https://doi.org/10.1016/j.engappai.2016.12.002>

Chapter 16

Classification of Upper Body Fits Using Fit Models



Juan Carlos Leyva López , Otto Alvarado Guerra, Itzel Juárez Sánchez, and Raúl Oramas Bustillos

Abstract The fit model defines the body measurements a garment retail company has determined and expresses the proportional relationships between body parts that are basic to achieving the company fit. Fit refers to how clothing conforms or differs on a person's body. A company's fit is an essential factor that sets its products apart from those of other companies. This implies that, usually, garments of the same size but from a different brand have different fits. This chapter applies a fuzzy set and linear regression-based classification method to find the upper body fits of a simulated target population for developing women's swimwear design for one of Mexico's largest apparel retailers. Six upper body parts were classified from very loose to very tight, including the neutral level concerning size M. In this work, we set up a supervised learning model to classify body measurements according to expert perceptions of women's upper body fits. We define a classification function for each body position. Considering the vagueness and imprecision of expert perceptions, we classify data of each classification function into five levels using fuzzy techniques. The results will help develop a body-sizing system for garment design adapted to the Mexican population.

Keywords Anthropometry · Upper body fit classification · Swimwear design · Fuzzy techniques · Linear regression · Sensory evaluation

J. C. L. López (✉) · O. A. Guerra · I. J. Sánchez · R. O. Bustillos
Department of Economic and Management Sciences, Universidad Autónoma de Occidente,
Sinaloa, México
e-mail: juan.leyva@uadeo.mx

R. O. Bustillos
e-mail: raul.oramas@uadeo.mx

16.1 Introduction

Manufacturers use fit models to base their apparel sizing designations [1]. It is worth noting that the visual evidence showed a significant difference in fit, highlighting the inconsistency in size designation and body dimensions among manufacturers. This further emphasizes the challenge of defining the perfect size height.

In order to create a sizing system and select a fit model, a company uses proportional relationships among body measurements, or key dimensions, to achieve the desired fit [2, 3]. The fit model represents the body dimensions necessary to achieve proportional clothing. Fit refers to garment conformity or deviation from the body. A clothing brand can distinguish its products from its competitors by establishing its own company fit. As a result, it is unlikely that clothes of the same size from different companies will fit the same way. Tamburrino [4] noted that body dimensions vary considerably among manufacturers, industry divisions, and countries from which fit model measurement specifications and size are designated in women's wear.

The fit of a garment on a person's body is an important aspect to consider in clothing design [5–7]. "Fit" refers to how well the clothes conform to the wearer's body [8]. Analyzing body fit has become particularly crucial in meeting the needs of a specific population in garment design and mass customization [9]. Accurately classifying the population is essential in designing and producing ready-to-wear products such as swimwear.

The evaluation of how well a garment fits the body can be conducted using various methods, including live, scanned, or parametric fit models [10, 11]. It's essential that the assessment of garment fit for both real and virtual garments is the same [12]. In a virtual environment, the body is represented as a parametric and scanned body model, which is based on body measurements and silhouettes of a live human being [13–15]. A parametric human body is a digital model based on user-specified body size inputs, and it's usually integrated into CAD systems for garments such as Gerber, Lectra, and Optitex [16]. As a result, clothing engineering researchers are increasingly focusing on developing virtual prototyping for garments [17, 18].

Currently, body measurements are used to obtain information about body sizes and fit. During an anthropometric survey, various important body measurements can be taken on individuals [19]. However, different body positions can have varying morphological features [20], making the existing classification criteria imprecise and inadequate for achieving a specific body fit [21].

The current classifications of body types primarily use traditional statistical methods [22–24]. However, these methods have certain limitations. A statistical analysis employing body measurements has been developed to address these limitations. Using statistical analysis to categorize body shapes makes it possible to identify the anthropometric characteristics of each body shape and group them into several categories with interrelationships. Discriminant analysis can be used to determine classification criteria [25, 26]. Statistical analysis methods have also been used to establish an appropriate size system to fit most customers [27] and describe body dimensions for product design purposes.

A significant advantage of statistical analysis is the objective classification of body shape based on anthropometric measurements. Based on statistical analysis, objective measures can uncover anthropometric differences between body types. However, in garment design, designers tend to use descriptive terms like loose, neutral, tight, etc., to describe body fits. Unfortunately, classical methods often fail to produce satisfactory results as they are ineffective in processing human perception. Fuzzy techniques are appropriate for dealing with imprecise and vague perceptions described by conventional linguistic terms used [28–30].

This chapter examines the size charts of one of Mexico’s largest retailers of clothing, home goods, jewelry, and beauty products. The company caters to the family market and provides size charts based on standard body measurements such as XS, S, M, L, and XL. Their ranges appeal to all genders and ages, so their market has maintained their size range. This size chart, constructed for contemporary women’s standard sizing, is based on the international sizing system XS-XXL. The size charts reproduce the larger average size of modern Mexican women. It also suggests the shape of the Mexican body. This kind of chart is utilized for various types of clothing. These size charts for young contemporary women reflect the anthropometric body of a young modern woman today. In this work, we are interested in the medium size M.

The study aimed to develop an objective body-fit classification process integrating three-dimensional anthropometric measurements, expert visual assessments, and statistical analysis. The proposed method introduced linguistic variables that represent the degree of fitness of a garment on a woman’s body of size M, which are analyzed using a fuzzy method. The study effectively integrated the visual assessments of three-dimensional simulated data of women’s upper bodies by an expert panel into statistical analysis techniques. The body fits were accurately identified by utilizing linear regression (LR). A novel and efficient method for classifying women’s upper body fits was proposed in this study, which provides a proper fit for designing mass-customized swimwear for Mexican women users.

The chapter is organized into several sections. Section 16.2 describes the materials and methods used in this study, while Sect. 16.3 focuses on human body fit modeling. In Sect. 16.4, we discuss the proposed body fit classification method. Section 16.5 presents an illustrative example of how the proposed model can be used to classify simulated human bodies. Section 16.6 presents future trends. Section 16.7 includes a summary of our findings and suggestions for future research.

16.2 Materials and Methods

This section discusses the selection of upper body measurements for garment design. Then, we set up the key dimensions of the upper body. Next, we set up the characteristic indices of the upper body fit. Afterward, we set up the sensory evaluation by experts of body fit.

16.2.1 Subjects

Three-dimensional body-simulated data of 150 women were employed to classify upper body fit parts. The simulated subjects were selected by using simple stratified random sampling regarding fit.

16.2.2 Selection of Upper Body Measurements

In this section, we present two data acquisition methods, including upper human body measurements and sensory evaluation on body fits, performed by expert company staff according to their professional knowledge. For this purpose, an experiment was designed. The acquired data were mathematically formalized to set up a model in the following sections.

This paper focuses on women's swimwear design; for this reason, we only need the measurement related to women's upper body positions. However, the general principle can be easily applied to different body positions and types of garments.

As shown in Table 16.1, nine body dimensions were chosen based on a literature review and a discussion with a group of company experts; the stature dimension is fixed. The women's bodies were simulated based on the definitions of body measurements in the International Organization for Standardization (ISO) 8559 [31] for garment construction and corresponding anthropometric dimensions. The length dimensions of the bodies were also considered. Finally, a view of the simulated bodies was obtained using the Clo3D software.








16.2.2.1 Recognition of Key Dimensions of the Upper Body

3D anthropometric data has many upper body data points, but only key dimensions are relevant for a particular population and garment [32]. The upper body measurements include both vertical and horizontal dimensions. The vertical dimensions consist of stature (S), front waist length (FWL), back waist length (BWL), neck shoulder point to breast point (NB), and body rise (BR). The horizontal dimensions include chest girth (CG), waist girth (WG), hip girth (HG), abdomen girth (AG), and back width (BW).

In this study, we identified ten key upper body dimensions crucial for swimwear design. These dimensions are measured in centimeters and represent a quantitative measuring vector for a specific human body. It can be defined as $MEASURE = (S, FWL, BWL, NB, BR, CG, WG, HG, AG, BW)$. Their descriptions are listed in Table 16.1.



Only key measurements are significant in garment design for achieving a specific body fit and garment. For swimwear design, selecting key dimensions will effectively classify the upper body fits of different consumers.

Table 16.1 Key dimensions to effectively classify the upper body fits of different consumers

Measurement	Abbr	Description	Diagram
Front waist length	FWL	The distance from the neck shoulder point, over the nipple, then vertically straight to the front waist	
Back waist length	BWL	The distance from the 7th cervical vertebra, following the contour of the spinal column, to the waist	
Neck shoulder point to breast point	NB	The distance from the neck shoulder point to the breast point	
Chest girth	CG	The maximum horizontal girth measured during normal breathing with the subject standing upright, and the tape measure passed over the shoulder blades, under the armpits (axillae), and across the nipples	
Waist girth	WG	The girth of the natural waistline between the top of the hip bones and the lower ribs	
Back width	BW	The horizontal distance across the back measured half-way between the upper and lower scye levels	
Hip girth	HG	Circumference of the hip at its widest point	

(continued)

Table 16.1 (continued)

Measurement	Abbr	Description	Diagram
Abdomen girth	AG	Keeping the tape measure straight and parallel to the ground with the subject standing, the circumference is taken, passing through the protruding hip bones around the fattest part of the tummy	
Body rise	BR	The vertical distance measured using the measuring stand, between the waist level and the crotch level	

16.2.3 The Characteristic Indices of the Upper Body Fit

When designing garments, the ratios and differences between body measurements are more important than the actual measurements. This is especially true when classifying body fits for people of different heights [22]. Based on the garment size standard and expert analysis, we have selected 13 upper body indices as follows.

1. **Waist fit index (*wf*):** The index *wf* describes the entire waist fit.

$$wf_1 = \frac{WG}{S} \tag{16.1}$$

wf can reflect the fat or thinness of the waist, and the higher the *wf* value, the plumper the waist looks.

2. **Hip fit indices (*hf*):** Hip fit can be described by the combination of a difference *hf*₁, and the ratios *hf*₂, and *hf*₃.

$$hf_1 = HG - WG, hf_2 = \frac{HG}{S}, hf_3 = \frac{BR}{S} \tag{16.2}$$

*hf*₁ can describe the flat level of the hip, and the higher *hf*₁ value is, the more raised the hip is. *hf*₂ can describe the fat or thin level of the hip, and the higher *hf*₂ value is the plumper hip looks.

3. **Abdomen fit indices (*af*):** Abdomen fit can be described by a difference *af*₁, and a ratio. *af*₂.

$$af_1 = HG - WG, af_2 = \frac{HG}{S}, af_3 = \frac{BR}{S} \tag{16.3}$$

af_1 can describe the flat level of the abdomen. The higher af_1 the value is, the more convex the abdomen is. af_2 can describe large or small levels of the abdomen; the higher af_2 value is, the plumper abdomen looks.

4. **Chest fit indices (cf):** Three ratios describe bust fit.

$$cf_1 = \frac{CG}{WG}, \quad cf_2 = \frac{CG}{S}, \quad cf_3 = \frac{NB}{S} \quad (16.4)$$

5. **Back fit indices (bf):** The indices bf_1 bf_2 describe the entire back fit.

$$bf_1 = BW - WG, \quad bf_2 = \frac{BW}{S} \quad (16.5)$$

The index bf_1 reflects the harmony of the upper body. The index bf_2 can reflect the small or significant level of the back; the higher the bf_2 value, the wider the back looks.

6. **The trunk fit indices (tf):** Two ratios tf_1 , tf_2 are used to describe the proportion of body fit.

$$tf_1 = \frac{FWL}{S}, \quad tf_2 = \frac{BWL}{S} \quad (16.6)$$

tf_1 and tf_2 can describe the trunk proportion; the higher the value, the longer the trunk seems.

These indices are calculated using body measurements and accurately represent how well a garment fits. They also allow for further studies, such as body fit classification.

Since each index approximately satisfies the normal distribution, it can be normalized by the following z-score normalization method.

$$x' = \frac{(x - \bar{X})}{\sigma}$$

where x is the initial data, x' is the normalized data, \bar{X} is the mean of data, and σ is the standard deviation.

Initial symbols denote all the normalized indices and let the set of body indices be assigned as follows:

$$SI = \{wf_1, hf_1, hf_2, hf_3, af_1, af_2, cf_1, cf_2, cf_3, bf_1, bf_2, tf_1, tf_2\}$$

The set can effectively describe the human upper body fit and facilitate the performing of further studies such as upper body fit classification.

16.2.4 Identification of Experimental Samples

It can be challenging to form accurate classes due to the limited amount of data available and the complex and time-consuming measuring process. In this experiment, 150 simulated women were selected and measured randomly.

It is essential to ensure that the results of the simulated experimental samples can be extrapolated to the population by checking their validity and representativeness, given that the number of samples is much smaller than the target population. According to the general features of human body data, the practical method is normal distribution examination.

A group of data x_1, x_2, \dots, x_n , x_i , $i = 1, 2, \dots, n$, is an outlier if it satisfies:

$$|x_i - \bar{X}| > 3\sigma$$

where \bar{X} is the mean of data and σ is the standard deviation. From the index values calculated from the initial body measurements, we found seven outliers.

With the index values, we determined $Q-Q$ plots ($Q-Q$ plot is a scatter plot on which the quantiles of the standard normal distribution are taken as horizontal coordinates, and the sample values are taken as vertical coordinates). The sample data related to the index values roughly satisfy the normal distribution if all the points on the $Q-Q$ plot approximately lie on a line whose slope is the standard deviation and intercept is the mean value, whereas if all points lie on a line, the data have an exact normal distribution. The $Q-Q$ plots of the index values of the sample data are shown in Fig. 16.1.

The results of $Q-Q$ plots show that these data roughly satisfy normal distributions except for a few kurtosis deviations. Ultimately, 143 sets of characteristic indices will be used as experimental samples.

16.2.5 Sensory Evaluation of Body Fits

In this work, we acquire garment experts' knowledge by using the sensory evaluation technique. Visual perception is regarded as a key human perception in garment design.

Sensory evaluation was applied to obtain the designers' subjective experience of garment design. The study categorized body positions into five sensory classes based on the visual perception of designers on body fit. These classes are "Very Loose (VL)," "Loose (L)," "Neutral (N)," "Tight (T)," and "Very Tight (VT)." The evaluation scores were then expressed utilizing a linguistic level of {VL, L, N, T, VT}.

In our study, the sensory experiment aimed to acquire design knowledge by evaluating simulated human bodies and obtaining the relations between body fits.

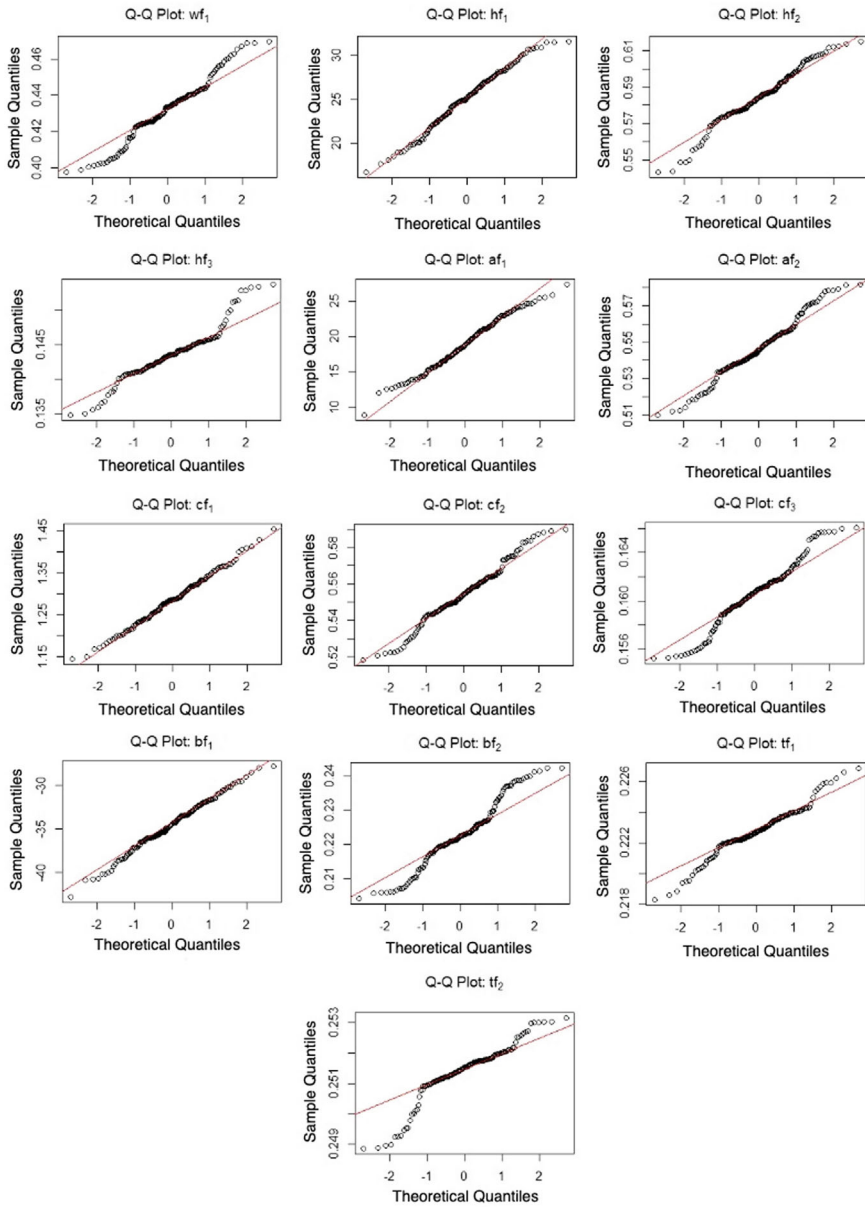


Fig. 16.1 Q-Q plots of index values of the sample data

Table 16.2 The descriptors and signs describing upper body fits

Table position	Abbr	Very loose	Loose	Neutral	Tight	Very tight
Trunk fit	<i>TF</i>	<i>TF^(VL)</i>	<i>TF^(L)</i>	<i>TF^(N)</i>	<i>TF^(T)</i>	<i>TF^(VT)</i>
Back fit	<i>BF</i>	<i>BF^(VL)</i>	<i>BF^(L)</i>	<i>BF^(N)</i>	<i>BF^(T)</i>	<i>BF^(VT)</i>
Chest fit	<i>CF</i>	<i>CF^(VL)</i>	<i>CF^(L)</i>	<i>CF^(N)</i>	<i>CF^(T)</i>	<i>CF^(VT)</i>
Waist fit	<i>WF</i>	<i>WF^(VL)</i>	<i>WF^(L)</i>	<i>WF^(N)</i>	<i>WF^(T)</i>	<i>WF^(VT)</i>
Hip fit	<i>HF</i>	<i>HF^(VL)</i>	<i>HF^(L)</i>	<i>HF^(N)</i>	<i>HF^(T)</i>	<i>HF^(VT)</i>
Abdomen fit	<i>AF</i>	<i>AF^(VL)</i>	<i>AF^(L)</i>	<i>AF^(N)</i>	<i>AF^(T)</i>	<i>AF^(VT)</i>

16.2.5.1 Description of Related Concepts

Before the experiment, we need to give basic descriptions of related concepts. As we are interested in the design of medium size women’s swimwear only, six upper body positions are considered (described in the experiment), and each body position can be evaluated using five scores: “Very Loose (*VL*),” “Loose (*L*),” “Neutral (*N*),” “Tight (*T*),” and “Very Tight (*VT*).” Their combinations lead to 15,625 upper body fits, described using the descriptors and signs of Table 16.2.

16.2.5.2 Sensory Panel and Training

A panel of experts with fashion design backgrounds carried out the sensory experiment since we needed to set up a decision support system. The panel is composed of five experts. Before the evaluation session, all panelists were invited to follow instructions on the experiment’s primary purpose, evaluation techniques and procedures, and interpretation of related concepts. The training session organized for the panelists took about one hour. This session was training body fit recognition by observing virtual pictures of 3D body generating, which helped panelists form unified recognition criteria.

The purpose of the present experiment was to classify the M-size body that effectively fits different simulated women according to the knowledge of expert panelists in garment design. Experts visualized how well a size M swimsuit fits a woman’s body cataloged to wear size M clothes, in each of the six parts that the upper part of the human body is divided, using the linguistic assessment scale {*VL*, *L*, *N*, *T*, *VT*}.

The trainer showed the panelist two 3D-modeled bodies; the first was with the company’s ideal body measurements (the fit model) for the medium size, and the second took the measures of 143 different women’s bodies within the medium size range. Both models were shown dressed in simple, modern swimsuits. When comparing the model with ideal measurements for the swimsuit (first) and the model with changing measures (second), the expert evaluated and determined the label that best described it according to the expert criteria. The specific labels to classify the parts of the bodies were very loose, loose, neutral, tight, and very tight.

Based on how the garment's fabric puckers or stretches on each part of the body, the expert determined if the swimsuit was "very loose," "loose," "neutral," "tight," or "very tight" on the body part the expert was viewing. Then, the expert checked the box corresponding to the part of the body visualized with the determined label.

The following text provides a comprehensive outline of the experimental design. For our sensory experiment, we have opted for discriminative tests due to their appropriateness.

The discriminative testing of sensory experimental design for garment design involves several key steps to ensure accurate and reliable results. Here is a detailed description of the steps involved in this process:

1. **Selection of Panelists:** The decision-maker, with the analyst's support, chose a panel of five experts with backgrounds in fashion design or garment fitting. The panelists had a keen eye for detail and a good understanding of garment construction and fit.
2. **Training Session:** The analysts conduct a training session for the panelists before the evaluation session. The training covered the primary purpose of the experiment, evaluation techniques, procedures, and interpretation of related concepts. This session helped panelists form unified recognition criteria for evaluating body fits.
3. **Presentation of Stimuli:** The analyst showed the panelists visual stimuli, such as 3D-modeled bodies wearing swimwear, to evaluate body fit. They presented two models simultaneously: one with ideal body measurements for the specific M-size (fit model measurements) and the other with measurements from various individuals within the same size range.
4. **Evaluation Process:** The analysts asked panelists to compare the two models and determine how well the swimwear fits each body. They used a linguistic assessment scale (very loose, loose, neutral, tight, very tight) to classify the fit of different body parts, such as trunk fit, back fit, chest/bust fit, waist fit, hip fit, and abdomen fit.
5. **Data Collection:** Each panelist's evaluations for the different body positions and fit descriptors were recorded. The analyst ensured that the evaluations were consistent and based on the predefined criteria discussed during the training session.
6. **Analysis of Results:** The sensory evaluation data were analyzed to identify patterns and trends in how panelists perceive body fits. The analysts looked for areas of agreement or disagreement among panelists to understand the variability in sensory perceptions.
7. **Integration with Body Measurements:** The analysts combined the sensory evaluation results with body fit measurements to create a comprehensive classification of body fits. This integration helped to develop a more accurate and design-oriented overall classification of body fits for swimwear design.
8. **Validation and Iteration:** The analysts validated the results obtained from the sensory experimental design by comparing them with other expert opinions.

16.2.5.3 Evaluation of Sensory Experiment of Different Body Position Fits

Five panelists were invited to the experiment, and 3D virtual body fit pictures were randomly generated for the 143 subjects selected for body fit measurements. As the evaluation procedure is very time-consuming and easy to tire people, we suggest that each panelist evaluate only one of six key upper body positions.

During the experiment, each body position was classified into five sensory classes based on the visual perception of 143 3D human samples. Figure 16.2 shows typical example images of various sensory classes.

In this way, each specific body fit belongs to one of the five sensory classes for each body position. These evaluation results (sensory body fit classification) were combined with the body fit measurements-based classification to perform a more accurate and significant design-oriented overall classification of body fits.

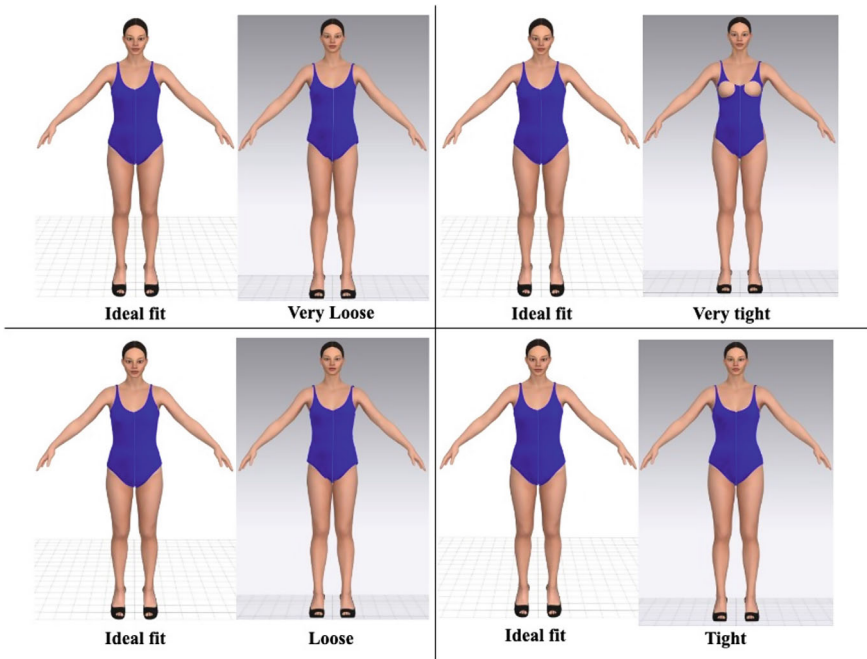


Fig. 16.2 Typical example images of various sensory classes

16.3 Modeling of the Human Body Fits

In this section, we will create a body fit classification model using body fit measurements most sensitive to the expert's perception.

16.3.1 Classification Model of Upper Body Fits

Note that the waist classification function depends only on the characteristic index of the waist fit. Nevertheless, not all indices can express their fit in the upper body position. Consequently, we should find the most appropriate classification function to model the classification of body fits.

Any upper body fit can be expressed by a 6-dimensional body fit vector $bf = (HF, WF, AF, TF, BF, CF)$. There are 15,625 potential medium-size body fits. In practice, analyzing the features of human bodies reveals that medium-sized body fits of a particular population are only a portion of them.

In this section, we set up a learning model to classify medium-size body fit measurements according to expert perceptions of upper human body fit.

In the context of classifying upper body fits in women's swimwear design, some multivariable methods can be utilized, including multiple linear regression (MLR), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NN). MLR is a valuable tool for modeling and classifying upper body fits in garment design, providing a systematic and data-driven approach to understanding and categorizing body fits.

When comparing MLR with SVM, RF, and NN for classifying upper body fits in women's swimwear design, we found several key differences and considerations:

- (i) Linear regression models provide straightforward interpretations of the relationships between input variables (body measurements) and the output (expert perceptions of body fit) whereas SVM, RF, and NN are often considered "black box" models, making it challenging to interpret how the model arrives at its predictions. Understanding the decision-making process and the importance of individual input variables may be more complex with these methods.
- (ii) Linear regression assumes a linear relationship between input variables and the output. While it is simple and easy to implement, it is unnecessary to use complex methods capable of capturing nonlinear relationships.
- (iii) Finally, MLR is computationally efficient and relatively simple to implement, making it suitable for datasets with a moderate number of input variables, whereas SVM, RF, and NN can be more computationally intensive, especially for large datasets or complex models. Training and optimizing these models may require more computational resources and time.

In this work, we use an MLR approach for the establishment of relationships between multiple body measurements (such as hip fit, waist fit, abdomen fit, trunk fit, back fit, and chest fit) and expert perceptions of body fit.

First, we accurately acquired the data on various dimensions of 143 simulated medium-sized human bodies (see Sect. 16.2.1). These measured body dimensions are related to six upper body positions (input data). Next, these 143 virtual human models dressed in swimsuits were evaluated by several experts according to their perceptions and experiences (output data). The classification model aims to extract rules that describe the relationship between input and output data. We define a classification function for each body position. Considering the vagueness and imprecision of human perceptions, we classify data of each classification function into five levels (scores) using fuzzy techniques. The different combinations of levels for all the six classification functions (six body positions) constitute all the possible upper body fits.

The experimental results have shown that the classification model is entirely acceptable to experts in terms of their perception.

16.3.1.1 The Classification of Different Upper-Body Positions

In this section, we model the classification for each upper body position. For simplicity, we propose to define a unidimensional classification function and classify body fits with this function.

The objective of the model is to model the parts of the body with a set of indices that reflect the appropriate fit of each body based on data and measurements of a group of randomly generated bodies. The coefficients required for the proposed equations corresponding to each body part and additional data necessary to evaluate how good each linear regression model is are shown.

16.3.1.2 Identification of the Classification Function of Each Body Position

As we know, the linear relation is the most straightforward relation between variables. Therefore, we create a linear link between the corresponding body fit measurement and the expert's perceptions of the body fit by using linear regression analysis for each body position. To apply the regression equation, the linguistic evaluation scores on body fit (from "very loose" to "very tight") are transformed into numbers 1–5.

We set up a linear regression Eq. (16.7) for the hip position by taking the three normalized hip fit indices hf_1 , hf_2 , hf_3 as independent variables and the human perception of hip fit (HF) as the dependent variable.

$$HF = k_0 + k_1hf_1 + k_2hf_2 + k_3hf_3 + \varepsilon \quad (16.7)$$

where ε is the residual error.

Assuming that the confidence level is $\gamma = 95\%$. After introducing all the sample data to Eq. (16.7), we have:

$$HF = -31.54735 + 0.04463hf_1 + 63.26092hf_2 - 23.81068hf_3 \quad (16.8)$$

where the F -statistic is $F = 157.1$, the residual standard error $RSE = 0.5623$, the adjusted R squared value is $R^2 = 0.7674$ and the p -value $p < 2.2E-16$. We can find that the linear regression form of all the input variables hf_1 , hf_2 , hf_3 and the dependent variable HF are validated because the p -value is smaller than the significance level α . Figure 16.3a shows the residual errors of all the data.

Like the analysis on hip position, we can obtain the linear regression form of waist position, denoted as WF .

$$WF = -19.4974 + 51.922wf_1 \quad (16.9)$$

The F -statistic is $F = 802.7$, the residual standard error $RSE = 0.3495$, the adjusted R squared value is $R^2 = 0.8495$, and the p -value $p < 2.2E-16$. In this case, the linear regression form of the input variable wf_1 and the dependent variable WF are also validated because the p -value is smaller than the significance level α . The residual plot is as Fig. 16.3b.

The linear regression form of the abdomen fit denoted as AF satisfies:

$$AF = -22.482386 + 0.003493af_1 + 46.581719af_2 \quad (16.10)$$

The F -statistic is $F = 311.3$, the residual standard error $RSE = 0.3506$, the adjusted R squared value is $R^2 = 0.8138$, and the p -value $p < 2.2E-16$. The linear regression form of the input variables af_1 , af_2 and the dependent variable AF are validated because the p -value is smaller than the significance level α . The residual plot is as Fig. 16.3c.

The linear regression form of the trunk fit, denoted as TF , satisfies:

$$TF = -216.14 + 317.61tf_1 + 589.78tf_2 \quad (16.11)$$

The F -statistic is $F = 321$, the residual standard error $RSE = 0.3711$, the adjusted R squared value is $R^2 = 0.8184$, and the p -value $p < 2.2E-16$. The linear regression form of the input variables tf_1 , tf_2 and the dependent variable TF are validated because the p -value is smaller than the significance level α . The residual plot is as Fig. 16.3d.

The linear regression form of the back fit, denoted as BF , satisfies:

$$BF = -21.89860 - 0.01273bf_1 + 109.8619bf_2 \quad (16.12)$$

The F -statistic is $F = 611.2$, the residual standard error $RSE = 0.3307$, the adjusted R squared value is $R^2 = 0.8958$, and the p -value $p < 2.2E-16$. The linear regression form of the input variables bf_1 , bf_2 and the dependent variable BF are validated because the p -value is smaller than the significance level α . The residual plot is as Fig. 16.3e.

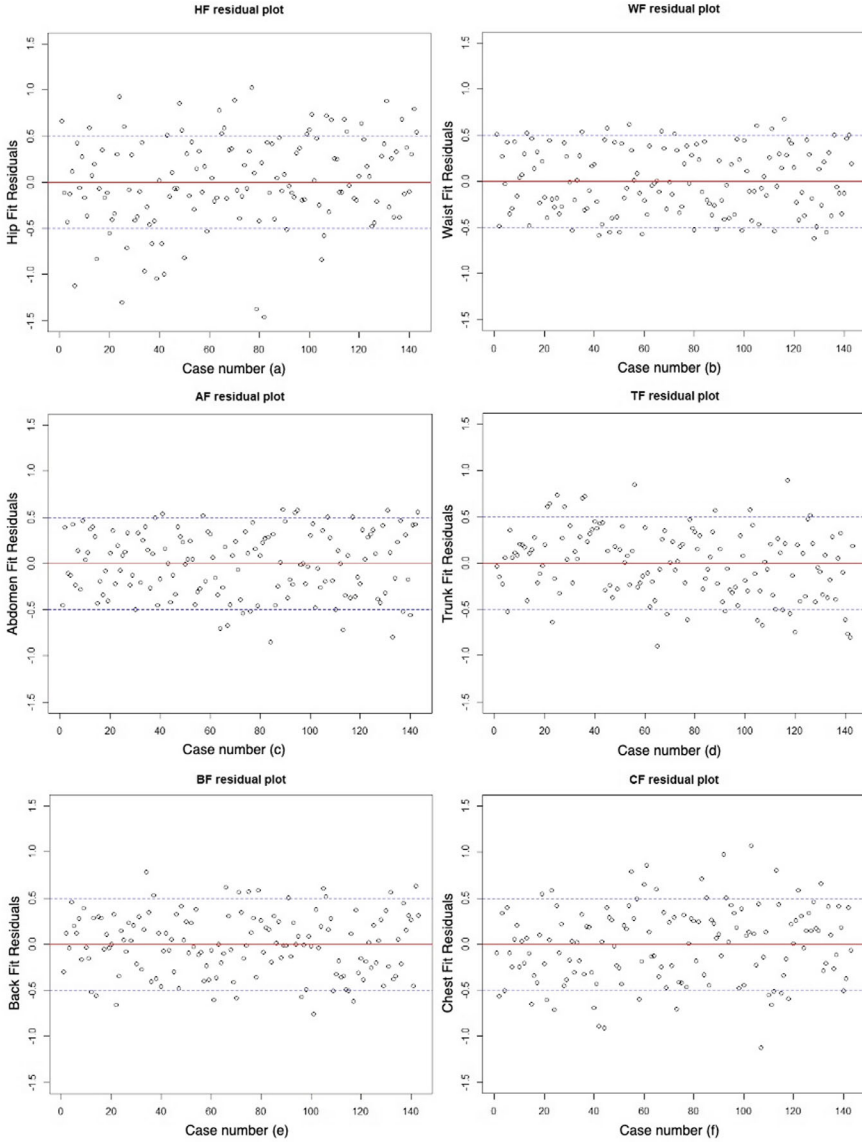


Fig. 16.3 The residual plot of the six fit functions

The linear regression form of the chest fit, denoted as CF , satisfies:

$$CF = -45.1312 + 0.7148cf_1 + 45.7030cf_2 + 136.1133cf_3 \tag{16.13}$$

The F -statistic is $F = 242.5$, the residual standard error $RSE = 0.4147$, the adjusted R squared value is $R^2 = 0.8361$, and the p -value $p < 2.2E-16$. The linear regression form of the input variables cf_1 , cf_2 , cf_3 and the dependent variable CF are validated since the p -value is smaller than the significance level α . The residual plot is as Fig. 16.3f.

After the application of these linear regression models, it is concluded that the models are reasonable, considering that they all show an explanatory capacity (R -squared) of 70% or more, the p -value is the same in all parts of the body, and is relatively close to 0; therefore, the probability that these results are due to randomness is very low. The R -squared value is a metric used to determine the quality of the best-fit line or the goodness of fit. A higher R -squared value indicates a better regression model, as it means that the model can explain most of the variation in actual values from the mean value.

Section 16.2.3 defines all the symbols of body indices used in these Equations. From these results, we can conclude that the body fit at each position can be modeled by a classification function generated from a linear combination of its corresponding body indices. Therefore, one 6-dimensional body fit vector $bf = (HF, WF, AF, TF, BF, CF)$ can express each upper body fit. Because the upper body fits explained by linguistic terms are more expressive in garment design, the previously defined classification functions are transformed into fuzzy sets.

16.3.1.3 Fuzzification of the Classification Function Values

Let the j th classification function value of the i th human body b_i in the human bodies set B be x_{ij} ($i = 1, 2, \dots, n$; $j = 1, 2, \dots, 6$).

In our study, the classification functions take values from the set of evaluation scores $\{VL$ (*Very Loose*), L (*Loose*), N (*Neutral*), T (*Tight*), and VT (*Very Tight*) $\}$.

We designate the following five numerical values to get five fuzzy sets stated by five evaluation levels (scores).

$$X_1^{(j)} = \min_{1 \leq i \leq n} \{x_{ij}\} \quad (16.14)$$

$$X_5^{(j)} = \max_{1 \leq i \leq n} \{x_{ij}\} \quad (16.15)$$

$$X_3^{(j)} = \text{median}_{1 \leq i \leq n} \{x_{ij}\} \quad (16.16)$$

$$hf_1 = HG - WG, hf_2 = \frac{HG}{S}, hf_3 = \frac{BR}{S} \quad (16.17)$$

$$X_4^{(j)} = \frac{X_3^{(j)} + X_5^{(j)}}{2} \quad (16.18)$$

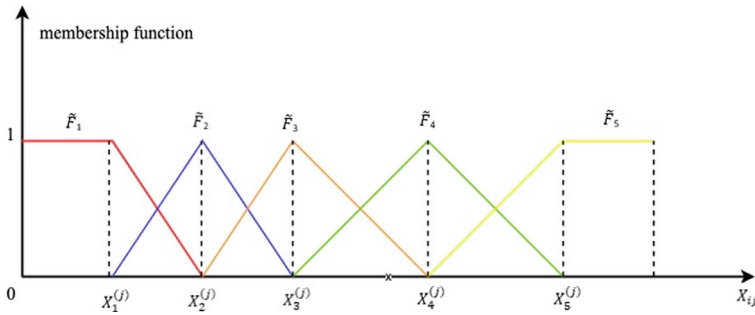


Fig. 16.4 Fuzzy membership functions of the upper body fit data

Based on these values, we can express the fuzzy set in the following form:

$$\begin{aligned} \tilde{F}_1 &= \text{Trapezoid}(0, 0, X_1^{(j)}, X_2^{(j)}), \\ \tilde{F}_2 &= \text{Triangle}(X_1^{(j)}, X_2^{(j)}, X_3^{(j)}), \\ \tilde{F}_3 &= \text{Triangle}(X_2^{(j)}, X_3^{(j)}, X_4^{(j)}), \\ \tilde{F}_4 &= \text{Triangle}(X_3^{(j)}, X_4^{(j)}, X_5^{(j)}), \\ \tilde{F}_5 &= \text{Trapezoid}(X_4^{(j)}, X_5^{(j)}, \infty, \infty). \end{aligned}$$

We select the median value rather than the mean value because the extreme values cannot influence it and don't always change with the sample change.

Each classification index value x_{ij} ($i = 1, 2, \dots, n; j = 1, 2, \dots, 6$) can be expressed by a vector based on the five fuzzy sets $\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \tilde{F}_4, \tilde{F}_5$, each having a triangle or trapezoidal membership function (Fig. 16.4).

The vector of membership degrees of x_{ij} , also called fuzzy distribution, is denoted as, $(\mu_{ij}^{(1)}, \mu_{ij}^{(2)}, \mu_{ij}^{(3)}, \mu_{ij}^{(4)}, \mu_{ij}^{(5)})$, where $\mu_{ij}^{(k)}$ is the membership degree of the function value x_{ij} ($i = 1, 2, \dots, n; j = 1, 2, \dots, 6$) to the fuzzy set \tilde{F}_k ($k = 1, 2, \dots, 5$).

According to the Maximal Membership Principle, \bar{x}_{ij} is affected by \tilde{F}_{k^*} if $\mu_{ij}^{(k^*)} = \bigvee_{k=1}^n \{\mu_{ij}^{(k)}\}$. The five fuzzy sets $\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \tilde{F}_4, \tilde{F}_5$ constitute the standard model base on the body fits.

16.3.2 The Classification of the Upper Body Fits

According to the above discussion, the fit of each body position concerning a garment of a predetermined size can be classified using the five fuzzy sets $\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \tilde{F}_4, \tilde{F}_5$.

From the simulated human body data obtained in experiments (body dimensions) and (body fit evaluation), we get the five key numerical values of the membership functions for each classification function as follows (Table 16.3).

Table 16.3 The five key numerical values of each classification function

	HF	WF	AF	TF	BF	CF
X ₁	0.2181	1.1389	1.3215	0.2714	1.0152	0.561
X ₂	1.6670	2.0675	2.1615	1.6051	1.9858	1.777
X ₃	3.1159	2.9961	3.0015	2.9387	2.9564	2.993
X ₄	4.2463	3.9414	3.8553	3.8951	4.0291	4.1199
X ₅	5.3767	4.8867	4.7091	4.8514	5.1018	5.2460

For example, for the waist fit (*WF*):

$$X_1^{(2)} = 1.1389, X_2^{(2)} = 2.0675, X_3^{(2)} = 2.9961, X_4^{(2)} = 3.9414, X_5^{(2)} = 4.8867$$

The following five values can be used to express the fuzzy sets:

$$\begin{aligned}\tilde{F}_1 &= \text{Trapezoid}(0, 0, 1.1389, 2.0675), \\ \tilde{F}_2 &= \text{Triangle}(1.1389, 2.0675, 2.9961), \\ \tilde{F}_3 &= \text{Triangle}(2.0675, 2.9961, 3.9414), \\ \tilde{F}_4 &= \text{Triangle}(2.9961, 3.9414, 4.8867), \\ \tilde{F}_5 &= \text{Trapezoid}(3.9414, 4.8867, \infty, \infty).\end{aligned}$$

According to the proposed fuzzification method, each classification function can be transformed into a fuzzy value. Then, we set up a discrete information system (data table) in which the overall six classification functions are taken as conditional attributes for modeling the upper body fits. Using the equivalent classification method of the information system, all the body fits are separated into 143 classes according to the fits of 6 upper body positions. The generated body fits in the whole population are somewhat distinct. The computational details can be shown in the illustrative example of Sect. 16.5.

16.3.3 Proof of the Model

Our study categorizes body fits for garment design, meeting designers' perception criteria. Consequently, the model's validity is established by whether the model output is consistent or compatible with the designers' perception criteria.

Figures 16.5, 16.6, 16.7, 16.8, 16.9 and 16.10 display the distribution graphs of the model output and perception criterion for the different body parts. The two distributions appear quite similar, but a more precise analysis is needed to confirm this. The intuitive analysis is not accurate enough, so further validation using a quantized method is required.

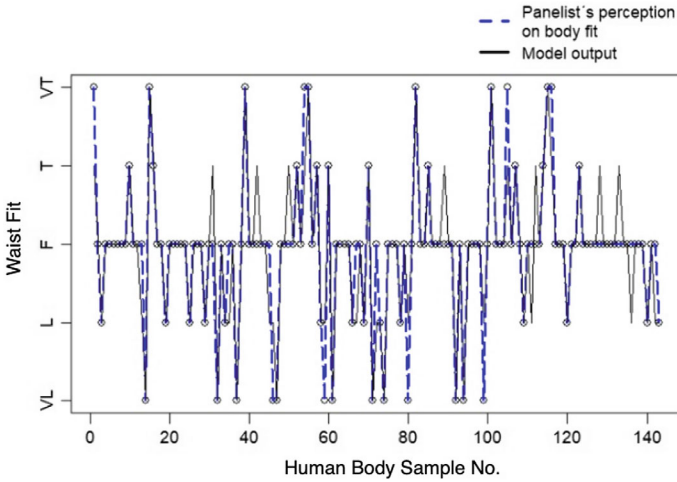


Fig. 16.5 Comparison between results of the model and human perceptions of waist fit

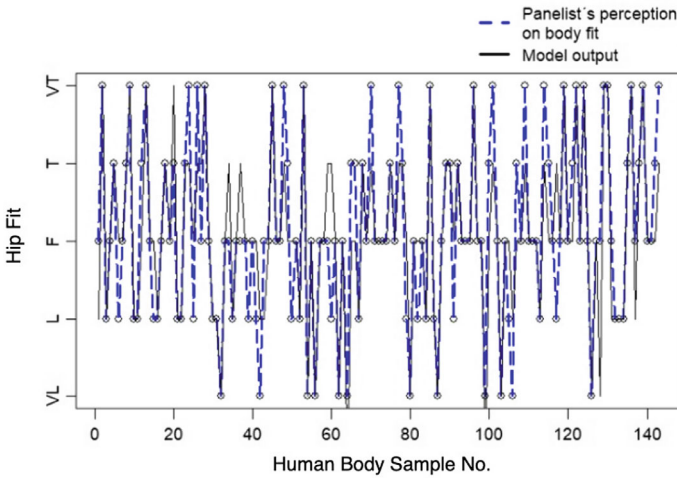


Fig. 16.6 Comparison between results of the model and human perceptions of hip fit

On analyzing the proposed model, it is observed that all the error results are less than 0.1. This indicates that the model's output is quite close to the perception criterion, on average. Based on this analysis, it can be inferred that the model shows a robust ability to categorize body fits.

This validation is performed on the population of 143 simulated human models created in Sect. 16.2.1. The results of the classification model and the panelist's evaluation of fits of various body positions are shown in Figs. 16.5, 16.6, 16.7, 16.8, 16.9 and 16.10.

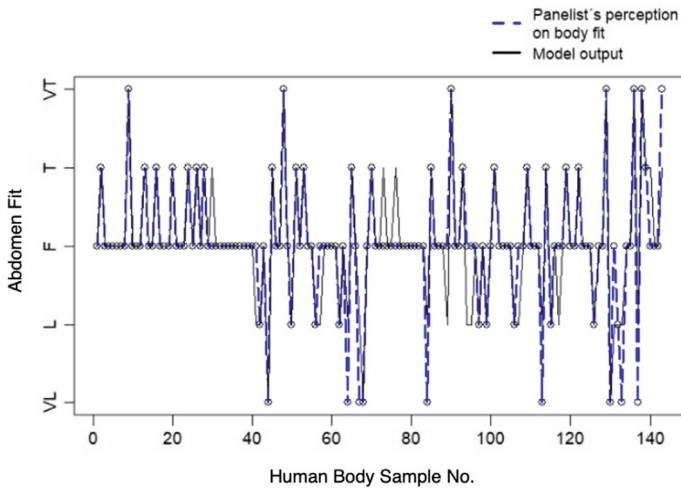


Fig. 16.7 Comparison between results of the model and human perceptions of abdomen fit

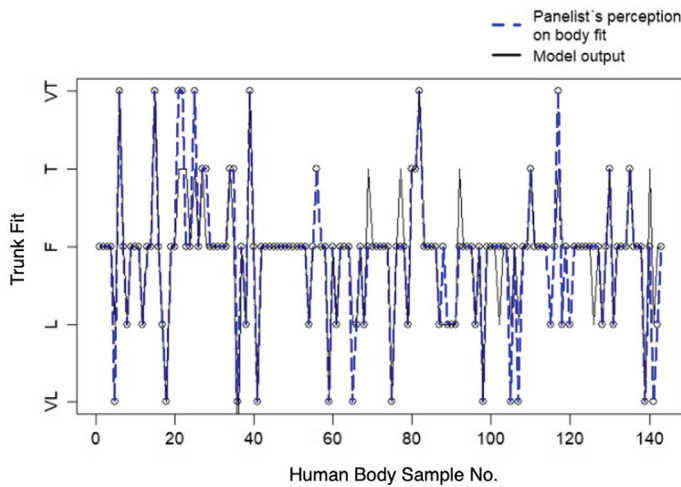


Fig. 16.8 Comparison between results of the model and human perceptions of trunk fit

The difference between the two results is expressed by the *Model_Error* formula, whose definition is as follows.

Let the set of all results delivered by the model be $X = \{x_1, x_2, \dots, x_n\}$, and the set of panelists' perceptions be $Y = \{y_1, y_2, \dots, y_n\}$. The criterion of *Model_Error* is defined as

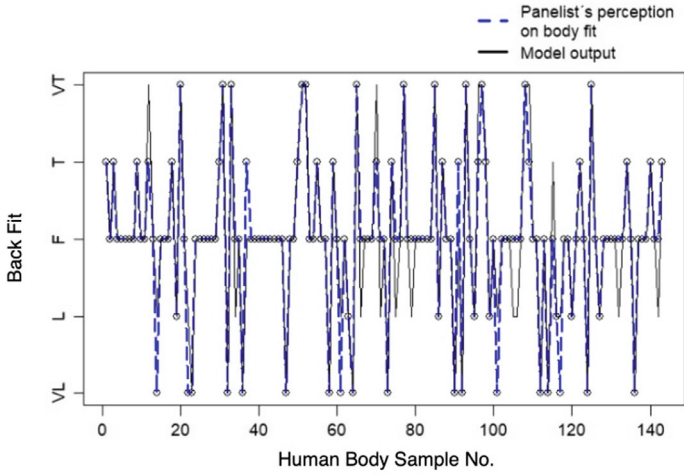


Fig. 16.9 Comparison between results of the model and human perceptions of back fit

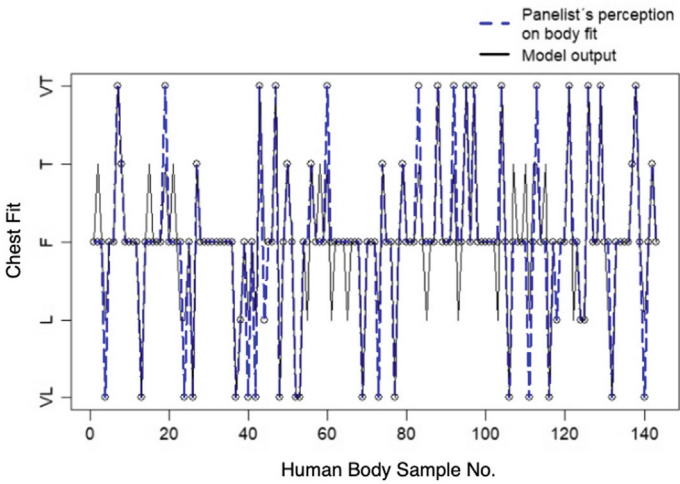


Fig. 16.10 Comparison between results of the model and human perceptions of chest fit

$$Model_Error(X, Y) = \sqrt{\frac{\sum_{i=1}^n (x_i - y_i)^2}{n}}$$

in this equation, all the results $x_i, y_i \in \{1, 2, \dots, 5\}$.

The values of the *Model_Error* for all the body positions are listed in Table 16.4.

Table 16.4 The values of the *Model_Error* for each body position

	Waist fit	Hip fit	Abdomen fit	Trunk fit	Back fit	Chest fit
Model error	0.0836	0.4181	0.0836	0.2508	0.0836	0.2508

All these errors are no more than 0.4, meaning that the error of human data is lower than 0.4 level of 5 ($< 10\%$) on average. Therefore, we can believe that the proposed model is acceptable.

16.4 Discussion

Various garment size systems are used to create clothing in a wide range of sizes for different body types. However, these sizes are still based on the average body measurements of a specific population. When it comes to water activities, the garment must fit the body perfectly and provide support to particular body parts. However, achieving a high level of fit for clothing models of different sizes and body types using only traditional clothing size systems and 3D construction methods can be challenging.

Swimwear customers often request customized clothing more than non-swimwear customers. A cost-effective mass-customized production system can provide suitable swimwear for customers.

In this study, the upper body fits of swimwear women users were investigated for a better-fit design using six upper body positions, five vertical key dimensions, and five horizontal key dimensions collected from 143 simulated swimwear users. The anthropometric measurements can be confidently utilized as reference data to design standard suit patterns for the identified body fits. The current study is limited by the range of stature for Mexican women users. In this study, the defined height for a Mexican woman was 1.65 m.

A group of experts visually categorized the body shapes of 143 computer-generated female models into six distinct types. However, it is important to note that the accuracy of visual classification may be influenced by various factors, including the number of people on the panel, their level of experience, and the volume of simulated data that needs to be classified. Therefore, it is recommended that objective criteria be used to differentiate one body fit from another.

This study aimed to develop a standardized method for categorizing women's upper body fits using three-dimensional body-simulated data. It introduced linguistic variables that describe the body fit and found important linguistic variables for each body type using objective measurements from three-dimensional simulated body data.

In the first part of this work, we acquire garment experts' knowledge using the sensory evaluation technique. In our study, the sensory experiments aimed to develop design knowledge by evaluating simulated upper human bodies. Expert participants were presented with two 3D simulated models generated simultaneously; the first had the ideal body measurements for a size medium (the fit model measurements), and the second took measures of 143 different simulated women's bodies within the size medium range. Both models were shown dressed in a simple modern style swimsuit and compared the model with ideal measurements for the swimsuit (first) and the model with changing measures (second). The expert evaluated and determined the label that best described it according to the expert criteria. The specific labels to classify the parts of the bodies were very loose, loose, neutral, tight, and very tight.

In the work's second part, we set up a supervised learning model to classify body measurements according to expert perceptions of upper human body fits. We define a classification function for each body position. Considering the vagueness and imprecision of human perceptions, we classify data of each classification function into five levels (scores) using fuzzy techniques. The different combinations of levels for all the six classification functions (six body positions) constituted all the possible upper body fits.

The body fit at each position was modeled by a classification function generated from a linear combination of its corresponding body indices. Therefore, one 6-dimensional body fit vector $bf = (HF, WF, AF, TF, BF, CF)$ expressed each upper body fit. Because the upper body fits described by linguistic terms are more meaningful in garment design, the previously defined classification functions were transformed into fuzzy sets. The fit of each body position was classified by using five fuzzy sets. Then, we set up a discrete information system (data table) in which the overall six classification functions are taken as conditional attributes for modeling the upper body fits.

16.5 An Illustrative Example

To validate the proposed supervised learning model to classify body measurements according to expert perceptions of women's upper body fits, we applied it to the design of women's swimwear for a specific body fit. The aim is to validate whether the model can offer a suitable classification for a particular body fit.

We give an example of swimwear design for two specific simulated human bodies (body_1 and body_2). To do so, we randomly selected two human bodies from the target population, and their 3D body images are presented in Fig. 16.11. The measuring vector for the 3D body image is as follows.

We obtain the nine key dimensions for the simulated human bodies, body 1, and body_2 (Fig. 16.11), and these two measure vectors are as in Tables 16.5 and 16.6.

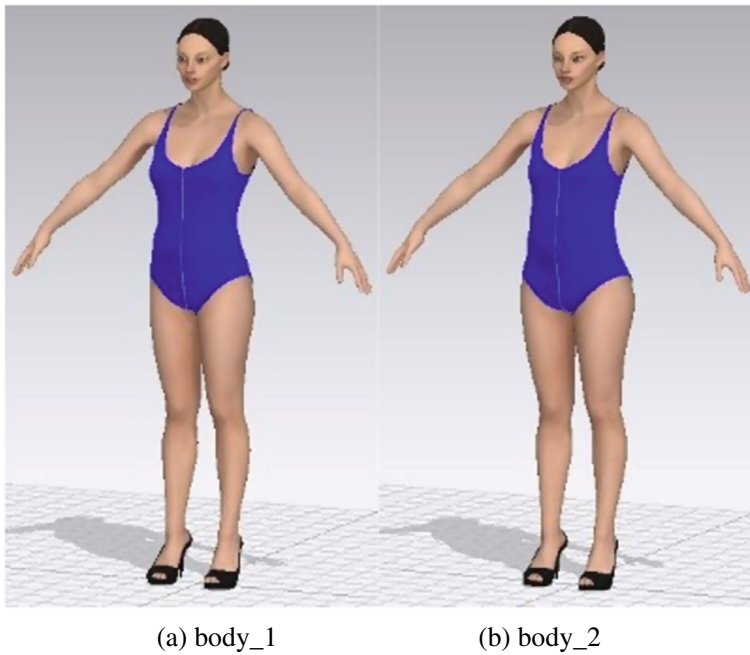


Fig. 16.11 3D pictures of two simulated bodies

Table 16.5 Measure body 1

FWL	BWL	NB	CG	WG	BW	HG	BR	AG
(36.7367,	41.5764,	26.6756,	92.9498,	65.7411,	38.4858,	84.0637,	23.9646,	93.7376)

Table 16.6 Measure body 2

FWL	BWL	NB	CG	WG	BW	HG	BR	AG
(36.9934,	41.4367,	26.1070,	86.8896,	70.9254,	36.6578,	85.7236,	23.4260,	100.7047)

All units are in “cm”

We compute these two simulated human bodies’ upper body index values using Eqs. (16.1)–(16.6) of Sect. 16.2.3.

Using these body index values and Eqs. (16.8)–(16.13) of Sect. 16.3.1.2, we obtain the body fit vectors of these two human bodies (body_1 and body_2).

Body_1:

	HF	WF	AF	TF	BF	CF
bf_body_1 =	(2.0006,	2.0020,	2.0041,	3,	4.0002,	3.0003)

Body₂:

	HF	WF	AF	TF	BF	CF
bf_body ₂ =	(3.0005,	3.0021,	4.0043,	3,	3.0002,	2.0002)

According to the proposed classification algorithm (Sect. 16.3.2), the classification functions can be fuzzified as.

	HF	WF	AF	TF	BF	CF
Fuzzyfy(bf_body ₁) =	(F ₂ ,	F ₂ ,	F ₂ ,	F ₃ ,	F ₄ ,	F ₃)

	HF	WF	AF	TF	BF	CF
Fuzzyfy(bf_body ₂) =	(F ₃ ,	F ₃ ,	F ₄ ,	F ₃ ,	F ₃ ,	F ₂)

Therefore, two body fits can be identified as follows.

The upper body fit of body₁ is ($HF^{(L)}$, $WF^{(L)}$, $AF^{(L)}$, $TF^{(F)}$, $BF^{(T)}$, $CF^{(F)}$), meaning that the M-size garment is loose on the hip, waist, and abdomen, fits on the trunk and chest, and is tight on the back of the body.

The upper body fit of body₂ is ($HF^{(F)}$, $WF^{(F)}$, $AF^{(T)}$, $TF^{(F)}$, $BF^{(F)}$, $CF^{(L)}$), meaning that the M-size garment fits on the hip, waist, trunk, and back, is tight on the abdomen, and is loose on the chest of the body.

These results show we can easily compare two body fits even if their statures differ. For example, The M-size garment fits better on body₂ than on body₁ at the hip, waist, and back.

16.6 Future Trends

This chapter explains a commonly used method (Multiple Linear Regression) for analyzing the relationship between multiple body fit measurements, such as hip fit, waist fit, abdomen fit, trunk fit, back fit, and chest fit, and expert perceptions of body fit. This method aims to highlight the significance of knowing the distribution of various body fit dimensions, which is crucial when designing clothing. The variations in body fit dimensions are carefully analyzed to identify correlations.

The method used in this study to classify the upper body fits of women’s swimwear design involved establishing a correlation between multiple body fit measurements and expert perceptions of body fit, based on linear relationships. It is important to note that not all anthropometric body fit dimensions and the expert’s perceptions of the body fit necessarily have linear relationships for each body position. For instance, waist fit measurement does not necessarily increase in proportion to the expert’s perception of waist fit; however, it has been observed that when waist fit measurement increases, the expert’s perception also increases.

The study has provided preliminary insights into challenges and techniques for classifying upper body fits. Future research on the classification of upper body fit should be undertaken to ensure better-fitting clothing for a successful business. Clothing manufacturers and retailers must focus on understanding consumers' different body shapes, fits, and sizes to improve consumer relationship management. It is crucial to classify upper body fits to establish predictable patterns for a specific population.

Advanced techniques, such as machine learning and neural networks, can be applied to better understand the relationships and patterns of body dimensions. Machine learning techniques can analyze non-linear relationships of body fit dimensions for a specific population and make predictions to support meaningful decisions [33]. One advantage of using machine learning techniques is that they do not require prior knowledge about the data since they are non-linear statistical data modeling tools compared to traditional linear approaches.

The technique explained in this chapter has illustrated the process of analyzing the variations of body fit dimensions in a simulated sample population. This method has been utilized to expose the linear relationships between multiple body fit measurements and expert perceptions of body fit. However, this discovery is just a preliminary step for further research, which should employ advanced machine-learning techniques to ascertain the cause-and-effect relationships between relevant variables.

16.7 Conclusions and Future Research

This chapter proposes a classification method using fuzzy set and linear regression to identify upper body fits for developing women's swimwear designs for one of Mexico's largest apparel retailers. Six M-size upper body parts were classified from very loose to very tight.

The proposed method for classifying upper body fits can reduce the time and effort required while ensuring reliability and accuracy. This method is beneficial when dealing with large amounts of anthropometric data. By focusing on the properties of the human body, the proposed method successfully separates and identifies upper body fits.

This study proposes an objective and efficient method for classifying body fits, which can be considered the primary contribution. After validation in practice, this method is expected to provide insights into 3D body image data for industrial practitioners. Mass customization will enable them to improve garment fitness and enhance customer satisfaction.

Further studies are required to investigate the effectiveness of anthropometric size labels. For example, if the range of size options is expanded hierarchically considering the body fit, current size labels would be limited. In this sense, the obtained results could effectively help to set up a newly studied retailer's body-sizing system for garment design adjusted to a Mexican target population and realize the concept of mass customization by developing customized garment styles using

hierarchically extended garment sizes where at the first level, there are the nodes XS, S, M, L, XL, XXL and below each of them the nodes “Very Loose (VL),” “Loose (L),” “Neutral (N),” “Tight (T),” and “Very Tight (VT).”

References

1. Fellingham, C.: Whose body is this for, anyway? *Secrets Perfect Fit Glamour* 159–160 (1991)
2. Bougourd, J.: Sizing systems, fit models and target markets. In: Ashdown (ed.) *Sizing in Clothing*, pp. 108–151. Woodhead Publishing (2007). <https://doi.org/10.1533/9781845692582.108>
3. Bougourd, J.: Ageing populations: 3D scanning for apparel size and shape. In: McCann, J., Bryson, D. (eds.) *Textile-Led Design for the Active Ageing Population*, pp. 139–169. Woodhead Publishing, Cambridge (2015). <https://doi.org/10.1016/B978-0-85709-538-1.00010-9>
4. Tamburrino, N.: *Sized to Sell. *Bobbin** vol. 33, pp. 69–74 (1992)
5. Faust, M.E., Carrier, S.: 3D Body scanning’s contribution to the use of apparel as an identity construction tool. In: Faust, M.-E., Carrier, S. (eds.) *Digital Human Modeling*, Vol. 5620, pp. 1–25. Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02809-0_3
6. Liu, K., Wu, H., Zhu, C., et al.: An evaluation of garment fit to improve customer body fit of fashion design clothing. *Int. J. Adv. Manuf. Technol.* **120**, 2685–2699 (2022). <https://doi.org/10.1007/s00170-022-08965-z>
7. Liu, K., Zhu, C., Tao, X., et al.: A novel evaluation technique for human body perception of clothing fit. *Multimed Tools Appl* **82**, 21057–21069 (2023). <https://doi.org/10.1007/s11042-023-14530-x>
8. Chrames, C., Boardman, R., McCormick, H., Vignali, G.: Investigating the impact of body shape on garment fit. *J. Fash. Mark. Manag.* **27**(5), 741–759 (2023). <https://doi.org/10.1108/JFMM-03-2022-0049>
9. Wang, Z., Tao, X., Zeng, X., Xing, Y., Xu, Y., Xu, Z., Bruniaux, P., Wang, J.: An interactive personalized garment design recommendation system using intelligent techniques. *Appl. Sci.* **12**, 4654 (2022). <https://doi.org/10.3390/app12094654>
10. Jevšnik, S., Pilar, T., Stjepanović, Z., Rudolf, A.: Virtual prototyping of garments and their fit to the body. In: Katalinić, B. (ed.) *DAAAM International Scientific Book 2012*, pp. 601–618. DAAAM International (2012). <https://doi.org/10.2507/daaam.scibook.2012.50>
11. Jevšnik, S., Kalaoğlu, F., Hanife Eryuruk, S., Bizjak, M., Stjepanović, Z.: Evaluation of a garment fit model using AHP. *Fibres & Text. East. Eururo.* **23**(2), 116–122 (2015)
12. Bye, E., McKinney, E.: Fit analysis using live and 3D scan models. *Int. J. Cloth. Sci. Technol.* **22**, 88–100 (2010). <https://doi.org/10.1108/09556221011018586>
13. Mahnic, M., Petrak, S.: Investigation of the fit of computer-based parametric garment prototypes. *J. Fiber Bioeng. Inform.* **6**(1), 51–61 (2019). <https://doi.org/10.3993/jfbi03201305>
14. Liu, K., Zeng, X., Wang, J., Tao, X., Xu, J., Jiang, X., Ren, J., Kamalha, E., Agrawal, T.K., Bruniaux, P.: Parametric design of garment flat based on body dimension. *Int. J. Ind. Ergon.* **65**, 46–59 (2018). <https://doi.org/10.1016/j.ergon.2018.01.013>
15. Liu, K., Zhu, C., Tao, X., Bruniaux, P., Zeng, X.: Parametric design of garment pattern based on body dimensions. *Int. J. Ind. Ergon.* **72**, 212–221 (2019). <https://doi.org/10.1016/j.ergon.2019.05.012>
16. Chaudhary, S., Kumar, P., Johri, P.: Maximizing performance of apparel manufacturing industry through CAD adoption. *Int. J. Eng. Bus. Manag.* **12** (2020). <https://doi.org/10.1177/1847979020975528>
17. Tanja, P., Stjepanović, Z., Simona, J.: Evaluation of fitting virtual 3D skirt prototypes to body. *Tekstilec* **56**, 47–62 (2013). <https://doi.org/10.14502/Tekstilec2013.56.47-62>

18. Jevšnik, S., Stjepanovič, Z., Rudolf, A.: 3D Virtual prototyping of garments: approaches, developments and challenges. *J. Fiber Bioeng. & Inform.* **10**(1), 51–63 (2017). <https://doi.org/10.3993/jfbim00253>
19. Zakaria, N.: Body shape analysis and identification of key dimensions for apparel sizing systems. In: Gupta, D., Zakaria, N. (eds.) *Anthropometry, Apparel Sizing and Design*, pp. 95–119. Woodhead Publishing (2014). <https://doi.org/10.1533/9780857096890.1.95>
20. Sabina, Olaru, F., Filipescu, E., Niculescu, C., Filipescu, E.: Morphological assessment of human body for clothing patterns design. *Indust Text.* **64**, 254–259 (2013)
21. Zakaria, N., Ruznan, W.S.: Developing apparel sizing system using anthropometric data: body size and shape analysis, key dimensions, and data segmentation. In: Zakaria, N., Gupta, D. (eds.) *Anthropometry, Apparel Sizing and Design (Second Edition)*, pp. 91–121. Woodhead Publishing (2020). <https://doi.org/10.1016/B978-0-08-102604-5.00004-4>
22. Dong, M., Hong, Y., Zhang, J., Liu, K., Wagner, M., Jiang, H.: A body measurements and sensory evaluation-based classification of lower body shapes for developing customized pants design. *Industria Textila* **69**(2), 111–117 (2018). <https://doi.org/10.35530/it.069.02.1381>
23. Takabu, H.: Analysis and classification of human body shape aimed applying to clothing design. *J. Jpn. Home Econ.* **59**(1), 687–697 (2008). <https://doi.org/10.11428/jhej.59.687>
24. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B. and Seidel, H.-P.: A statistical model of human pose and body shape. *Comput. Graph. Forum* **28**(2), 337–346 (2009). <https://doi.org/10.1111/j.1467-8659.2009.01373.x>
25. Yu, M., Kim, D.E.: Body shape classification of Korean middle-aged women using 3D anthropometry. *Fash Text.* **7**(35) (2020). <https://doi.org/10.1186/s40691-020-00223-8>
26. Mahnic Naglic, M., Petrak, S.: A method for body posture classification of three-dimensional body models in the sagittal plane. *Text. Res. J.* **89**(2), 133–149 (2019). <https://doi.org/10.1177/0040517517741155>
27. Balach, M., Lesiakowska-Jablonska, M., Frydrych, I.: Anthropometry and size groups in the clothing industry. *Autex Res. J.* **20** (2019). <https://doi.org/10.2478/aut-2019-0001>
28. Wang, L.C., Zeng, X.Y., Koehl, L., Chen, Y.: Intelligent fashion recommender system: fuzzy logic in personalized garment design. *IEEE Trans. Hum. Mach. Syst.* **45**(1), 95–109 (2015). <https://doi.org/10.1109/THMS.2014.2364398>
29. Liu, K., Wang, J., Tao, X., Zeng, X., Bruniaux, P., Kamalha, E.: Fuzzy classification of young women's lower body based on anthropometric measurement. *Int. J. Ind. Ergon.* **55**, 60–68 (2016). <https://doi.org/10.1016/j.ergon.2016.07.008>
30. Cisneros, L., Rivera, G., Florencia, R., Sánchez-Solís, J.P.: Fuzzy optimisation for business analytics: a bibliometric analysis. *J. Intell. & Fuzzy Syst.* **44**(2), 2615–2630 (2023). <https://doi.org/10.3233/JIFS-221573>
31. International Organization for Standardization: *Size Designation of Clothes—Part 1: Anthropometric Definitions for Body Measurement (ISO No. 8559–1)*. Geneva, Switzerland: International Organization for Standardization (2014)
32. Dong, M., Zeng, X., Koehl, L.: Dynamic fuzzy clustering of lower body shapes for developing personalized pants design. In: *Conference on Uncertainty Modelling in Knowledge Engineering and Decision Making*, pp. 898–904 (2016). https://doi.org/10.1142/9789813146976_0139
33. Ochoa, A., Ponce, J., Ornelas, F., Jaramillo, R., Zatarain, R., Barrón, M., Gómez, C., Martínez, J. and Elias, A.: *New Implementations of Data Mining in a Plethora of Human Activities*. InTech (2011). <https://doi.org/10.5772/13454>

Chapter 17

Regression Models for Estimating the Stress Concentration Factor of Rectangular Plates



J. Alfredo Ramírez Monares and Rogelio Florencia Juárez

Abstract Estimating Stress Concentration Factors (SCF) guarantees resistance and durability criteria in structures and design components. Failure to correctly identify the SCFs could lead to premature material failure. In this chapter, eight regression models were used to predict the SCF. The regression models were multiple linear regression, random sample consensus, ridge regression, LASSO regression, elastic net, random forest regression, support vector regression, and polynomial regression. The models were trained on a dataset resulting from a two-dimensional Finite Element Analysis from the Finite Element Method for different values of the parameters: large, width, and circular hole radius in a tensile plate. Least squares polynomial equations were fitted to these design points. The performance of the models was compared using the MSE, RMSE, MAE, MAPE, and R2 metrics. The random forest regression performed the best.

Keywords Stress concentration factor · Rectangular plates · Polynomial curve fitting · Artificial intelligence · Regression models · Random sample consensus · Ridge regression · LASSO regression · Elastic Net · Random forest regression · Support vector regression · Polynomial regression

17.1 Introduction

The distribution of elastic stress across the section of a member may be nominally uniform or may vary in some regular manner. When the variation is abrupt so that within a very short distance, the intensity of stress increases greatly, the condition is described as stress concentration [10]. A sudden change in geometry can cause a

J. A. R. Monares (✉) · R. F. Juárez
Universidad Autónoma de Ciudad Juárez, Av. Plutarco Elías Calles 1210 Fovissste Chamizal
Ciudad Juárez, Chihuahua, Mexico
e-mail: jose.ramirez@uacj.mx

R. F. Juárez
e-mail: rogelio.florencia@uacj.mx

localized increase in stress in a specific area. The region of geometric change is characterized by significantly higher stress values. It is usually due to local irregularities of form, such as small holes, screw threads, scratches, and similar stress raisers.

The analysis of stress concentrations began in 1898 with Ernst Gustav Kirsch's. Linear elastic solution for stresses around a hole in an infinite plate was presented in [2]. A factor-of-three stress concentration at the hole under uniaxial loading is present in Kirsch's solution.

"Streamlining" is a common term for the analogy of flow. The process involves the fundamental analysis of a system's behavior using flow components. The stress distribution in an axially loaded plate resembles the stress distribution in a channel due to fluid flow in a channel. The abrupt decrease in the channel's cross-sectional area causes the flow velocity to increase to maintain the flow rate, leading to the convergence of the streamlines and the overall path narrowing. A similar situation occurs when a stress-loaded plate is stressed. Furthermore, it transpires that the equations governing fluid flow and stressed systems are remarkably similar, and in some instances, even identical.

Understanding stress concentration near holes is crucial for ensuring the reliable design of structural components. Typically, high-strength materials are employed to design structural parts with high mechanical performance to minimize the stress concentration factor. This necessitates a deeper comprehension and accurate modeling of the behavior of these structures [9].

Pilkey [5] and Young [10] have presented data regarding the stress concentration factor, taking into account a diverse range of dimensional ratio configurations. Many sources of information rely exclusively on a two-dimensional solution to elasticity theory.

Assuming that the material remains elastic, the maximum stress in a circular hole in an infinite plate under tension is three times the applied stress [6].

Troyani et al. [7] have determined the theoretical in-plane stress concentration factors for short rectangular plates with centered circular holes subjected to uniform stress using the Finite Element Method. The thickness and Poisson's ratio affect the stress concentration factor.

On the other hand, there are works related to Stress Concentration Factors estimation and Artificial Intelligence. For example, in [8], presents a numerical model for predicting the Stress Concentration Factor with 7 parameters as inputs in a fatigue application. In [1] there is a similar estimation in welded joints.

It presents a study in [4] about the optimization of the ellipse dimensions in order to find a minimum SCF based on results obtained by FE and Curve Fitting. In the present work, Curve Fitting is applied to adjust the data of SCF in a circular hole, and the aim is not to find a minimum SCF, if not a representation of the input and output parameters, their relation, and a comparison with the AI algorithms.

In [3] it is presented a numerical model based on an Artificial Neural Network for the Stress Concentration Factor Estimation in a plate with a V-shaped notch. The stress concentration factor is obtained according to the strength of the upper limit safety factor value. In the present work, a similar model is obtained, but for plates in tensile stress with a circular hole at the center.

The difficulty in calculating the SCF lies in the complexity of the geometric, material, load, and interaction factors involved. Analytical methods have limitations, such as they are only applicable to standard geometries and loading conditions and are not easily adapted to complex geometries or non-standard loading conditions. Numerical methods can demand significant computational resources, be slow for large, complex problems, and require advanced knowledge to set up and analyze models correctly. Regression algorithms to estimate the SCF could offer significant advantages in handling complex geometries, reducing calculation time, providing flexibility, and enabling the ability to predict SCF under new conditions.

In this chapter, regression models are proposed to estimate the SCFs in rectangular plates. The models were trained on a dataset resulting from a two-dimensional Finite Element Analysis (FEA) from the Finite Element Method (FEM) for different values of the parameters: large, width, and circular hole radius in a tensile plate. Least squares polynomial equations were fitted to these design points.

This chapter is structured as follows: In Sect. 17.2, the stress concentration factor in a rectangular plate with a central hole will be described, how it was modeled by the finite element method will be shown, and the results will be presented for different combinations of the design parameters. Section 17.3 describes the regression models used to estimate the stress concentration factors. Section 17.4 presents the methodology used. Section 17.5 shows the results obtained by the regression models. Lastly, Sect. 17.6 presents the conclusions and future work.

17.2 Stress Concentration Factor

The Stress Concentration Factor (SCF) is crucial in mechanical and structural engineering. It describes how stresses increase around geometric discontinuities, such as holes, notches, abrupt changes in section, or inclusions in a material. It is defined as the relationship between the maximum stress in the area of the discontinuity and the nominal reference stress that would occur in a uniform section without the discontinuity.

Pilkey [5] states that the SCF is a theoretical value determined by Eq. (17.1).

$$k_t = \frac{\sigma_{max}}{\sigma_{nom}} \quad (17.1)$$

where σ_{max} represents the maximum stress anticipated in the component under the applied load, while σ_{nom} denotes the nominal stress, also referred to as the reference stress.

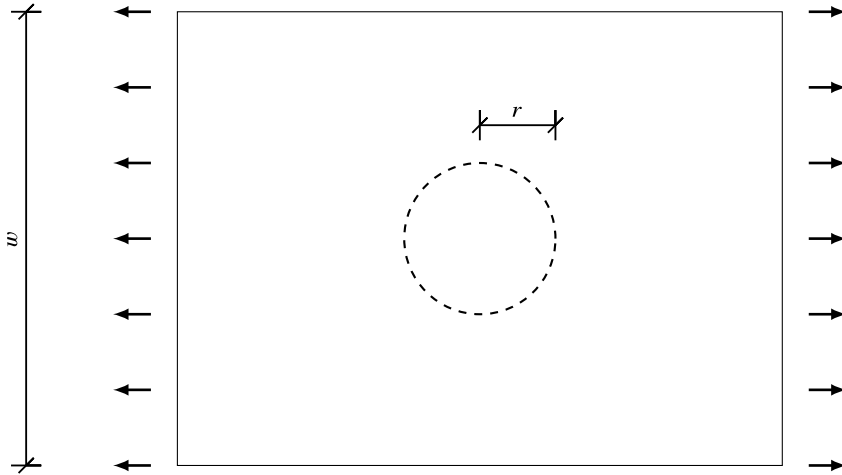


Fig. 17.1 Rectangular plate with central hole

17.2.1 Tensile Plate with Circular Hole in the Center

Evaluating a tensile plate with a circular hole in the center is a classic problem in materials mechanics and stress analysis. This type of geometric discontinuity generates a stress concentration around the hole, which can be critical to the design and structural integrity of the component. Analyzing a tensile plate with a circular hole in the center is essential to guaranteeing the structural integrity and safety of mechanical and structural components.

The model comprises a traditional rectangular plate characterized by a height w , length L , uniform thickness t , a centrally positioned aperture with radius r , and an applied tensile force P , as depicted in Fig. 17.1. The dotted lines in the figure illustrate the aperture responsible for stress concentration. The geometry of the aperture is solely defined by its radius r . In this scenario, the nominal stress is determined by Eq. (17.2).

$$\sigma_{nom} = \frac{P}{t(w - 2r)} \quad (17.2)$$

Normalizing the dimensions of the plate is necessary, so the input parameters considered herein are the ratios w/r and L/t .

17.2.2 Finite Element Model Results

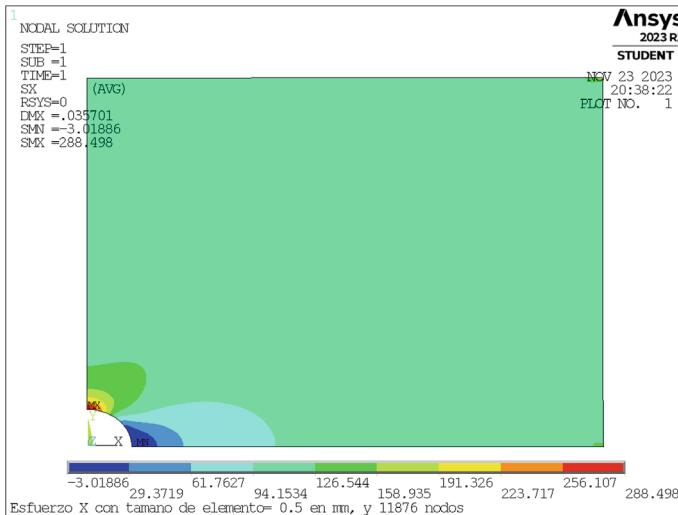


Fig. 17.2 Stress contour plot of a quarter of the whole plate

Maximum stresses for various plate dimensions were calculated using the Finite Element Method (FEM). The plate exhibits both horizontal and vertical symmetry, allowing for the modeling of only one-fourth of the entire plate with its full thickness, as illustrated in Fig. 17.2. Several mesh discretizations were examined, achieving satisfactory convergence with a structured mesh of 12,500 Kirchhoff plate finite elements. It was implemented a 2D model with different dimensions corresponding to every case here analyzed. The traction force $P = 10$ kN was included as a distributed load in the boundary conditions of the models. Table 17.7, in the Appendix, has a total of 48 dimension combinations; they are described in the 2 leftmost columns. With the P load and the mentioned dimensions, the σ_{nom} was obtained, according to Eq. (17.2). A .mac file was edited and loaded in ANSYS APDL software for every dimension combination. Thus, maximum stress is achieved. Then, for each combination of parameters, the stress concentration factor k_t is determined according to Eq. (17.2). These factors are generated from the FEM results and are in the third column of the Table 17.7.

Figure 17.3 is a graph of the k_t factor as a function of the L/t parameter. There are 9 different lines corresponding to 9 different w/r parameters. It is clear the SCF dependence of both parameters, L/t and w/r .

The code to create one of the finite element models was developed in a script file with a .mac extension in ANSYS APDL.¹

¹ <https://drive.google.com/file/d/1Tg-JQO6-H6TfjgpkMaZnZ0bxPDSdNpGn/view?usp=sharing>.

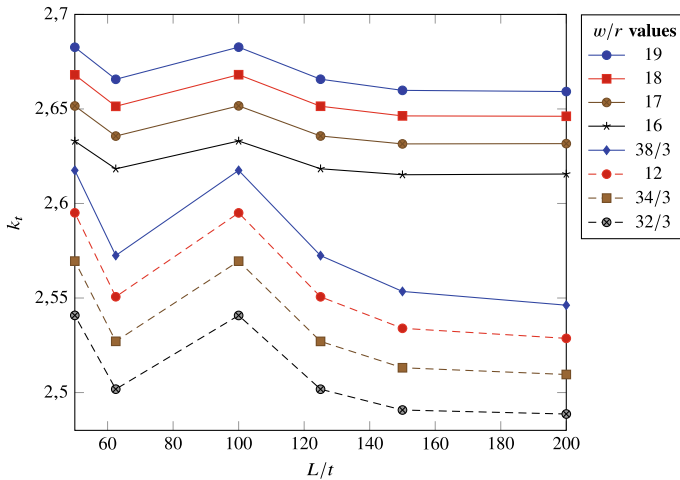


Fig. 17.3 Stress concentration factors k_t for different parameters w/r in the rectangular plate

Table 17.1 Constants c_i

i	1	2	3	4	5
c_i	2.22128	0.0414192	0.000920944	0.00130236	0.000136818

17.2.3 Stress Concentration Factors by Polynomial Curve Fitting

This paper uses SCF values obtained from FE data in the previous section to interpolate polynomials to estimate SCF k_t as a result of the independent variables, the parameters L/t and w/r . The SCF equations are obtained by fitting the finite element results in the second-, third-, and fourth-order polynomial equations. In this way, Eq. (17.3) is obtained.

$$\begin{aligned}
 k_t = & c_1 + c_2 \left(\frac{w}{r}\right) - c_3 \left(\frac{w}{r}\right)^2 - c_4 \frac{L}{t} + c_5 \frac{w}{r} \frac{L}{t} \\
 & - c_6 \left(\frac{w}{r}\right)^2 \frac{L}{t} + c_7 \left(\frac{L}{t}\right)^2 - c_8 \frac{w}{r} \left(\frac{L}{t}\right)^2 \\
 & + c_9 \left(\frac{w}{r}\right)^2 \left(\frac{L}{t}\right)^2
 \end{aligned} \tag{17.3}$$

where the constants c_i are shown in the Tables 17.1 and 17.2.

Table 17.2 Constants c_i

i	6	7	8	9
c_i	3.92567×10^{-6}	1.73889×10^{-6}	3.40912×10^{-7}	1.34912×10^{-8}

Using this Eq. (17.3), the fourth column of Table 17.7 is obtained for every parameter variation. The last column in the Table 17.7 represents the relative percentage error between the Curve Fitting estimation and the FE one.

17.3 Machine Learning Regression Models

Section 17.3.1 describes the regression models used to estimate the SCF. Section 17.3.2 describes the metrics used to evaluate the performance of the regression models.

17.3.1 Regression Models

The following subsections describe the models utilized: *Multiple Linear Regression*, *Random Sample Consensus*, *Ridge Regression*, *LASSO Regression*, *Elastic Net*, *Random Forest Regression*, *Support Vector Regression*, and *Polynomial Regression*.

We selected these algorithms because they use different regression techniques. Multiple linear regression is based on linear regression. Random Sample Consensus is based on parameter estimation in a data set with outliers using an iterative approach. Ridge, LASSO, and Elastic Net are based on regularized regression. Random Forest Regression is tree-based. Support Vector Regression is based on support vector machines. Polynomial regression is based on non-linear regression.

17.3.1.1 Linear Regression Model

Linear regression is a widely used statistical technique in data analysis. It mathematically models the relationship between a dependent variable (response variable) y and an independent variable (predictor variable) x through a linear equation.

Multiple linear regression is a generalization of simple linear regression where the value of the dependent variable y is determined from a set of independent variables x_1, x_2, \dots, x_n . Multiple linear models can be expressed using Eq. (17.4).

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_n \cdot x_n + \epsilon, \quad (17.4)$$

Where y is the dependent variable being predicted, x_1, x_2, \dots, x_n are the independent variables, $\beta_0, \beta_1, \dots, \beta_n$ are the model coefficients representing the effect of each independent variable on the dependent variable, and ϵ represents the residual or error derived from the difference between the observed and estimated values by the model, capturing the variability not explained by the model.

Linear regression models allow for the estimation of coefficients $\beta_0, \beta_1, \dots, \beta_n$ that minimize the sum of the squares of the residuals, thus minimizing the difference between real observations and model predictions, enabling predictions on new data. Although linear regression is widely used, its ability to model nonlinear relationships between variables is limited.

17.3.1.2 Random Sample Consensus Model

The *Random Sample Consensus* model is a robust technique that can be used in machine learning to estimate the parameters of a linear regression model. It is commonly used for both linear and non-linear regression problems.

RANSAC is a non-deterministic algorithm that produces a reasonable result only with a certain probability, which increases as more iterations are allowed. Through these iterations, the regression model is fitted to a random subset of the data, known as *inliers*, and then evaluates the quality of how many data points fit well to the model (*inliers*) and how many are considered *outliers*.

This process is repeated a specified number of times or until a predefined convergence criterion is reached. Finally, the model that produced the best fit according to a consensus criterion is selected. RANSAC is considered a robust model that can be particularly useful in datasets containing large noise or outliers.

17.3.1.3 Ridge Regression Model

Ridge, also known as L2 regularization or Tikhonov regularization, is a regression technique used to address the issue of multicollinearity and overfitting in linear regression models.

This technique adds an L2 regularization term to the cost function that allows controlling multicollinearity among predictor variables, i.e., when some of these variables are highly correlated with each other. In addition, it prevents the model coefficients from becoming too large, which occurs when there is a correlation between independent variables. This regularization term is only added to the cost function during training. Model performance evaluation is conducted considering only the unregularized cost function.

The equation for Ridge regression is similar to that of linear regression but includes the L2 penalty term, which is proportional to the sum of the squares of the model coefficients multiplied by a regularization parameter α . Introducing this penalty term causes the coefficients to be small but not exactly zero, helping to reduce multicollinearity and prevent overfitting. The cost function is defined in Eq. (17.5).

$$J(\theta) = \sum_{i=1}^m (y_i - h_{\theta}(x_i))^2 + \alpha \sum_{j=1}^n \theta_j^2 \quad (17.5)$$

Where $J(\theta)$ is the cost function, m is the number of observations, n is the number of features, y_i is the dependent variable in the i -th observation, $h_{\theta}(x_i)$ is the model prediction for the i -th observation, θ_j represents the coefficients of the model and, α controls the degree of penalty, the higher this value is, the coefficients will be smaller, resulting in more robustness to collinearity. A value of 0 causes the penalty term to have no effect, producing the same coefficients estimated by least squares.

17.3.1.4 LASSO Regression Model

LASSO (Least Absolute Shrinkage and Selection Operator) regression, also known as L1 regularization regression, is a linear regression technique that uses regularization to improve generalization and perform automatic feature selection.

Like Ridge regression, LASSO incorporates a regularization term in the cost function. However, instead of using the L2 norm to penalize the magnitude of coefficients, it uses the L1 norm. This helps control multicollinearity. LASSO also allows for automatic variable selection by forcing some coefficients to be 0, meaning certain independent variables are excluded from the model, eliminating those irrelevant to prediction.

The cost function used in LASSO regression is the sum of squared prediction errors plus the penalty term, which is proportional to the sum of the absolute values of the model coefficients multiplied by a regularization parameter α .

L1 regularization tends to force some coefficients to 0, leading to automatic feature selection by removing those that are not relevant for prediction. The cost function in LASSO regression is defined in Eq. 17.6.

$$J(\theta) = \sum_{i=1}^m (y_i - h_{\theta}(x_i))^2 + \alpha \sum_{j=1}^n |\theta_j| \quad (17.6)$$

Where $J(\theta)$ is the cost function, m is the number of observations, n is the number of features, y_i is the dependent variable at the i -th observation, $h_{\theta}(x_i)$ is the model prediction for the i -th observation, θ_j are the model coefficients, α is the regularization parameter which controls the strength of the regularization. A higher value implies stronger regularization. A value of 0 causes the penalty term to have no effect.

17.3.1.5 Elastic Net Model

The *Elastic Net* regression model is an extension of Ridge and LASSO regression that combines the L1 and L2 penalty terms into a single cost function. This combina-

tion allows for the simultaneous leveraging of the advantages of both regularization approaches to address the issues of multicollinearity, overfitting, and variable selection in regression models.

The cost function used in Elastic Net regression is a combination of the sum of squared prediction errors, a penalty term proportional to the sum of the absolute values of the model coefficients (L1 regularization), and a penalty term proportional to the sum of the squares of the model coefficients (L2 regularization). An additional parameter λ is used to control the relative contribution of each penalty term. The cost function in Elastic Net regression is defined by Eq. (17.7).

$$J(\theta) = \sum_{i=1}^m (y_i - h_{\theta}(x_i))^2 + \lambda_1 \sum_{j=1}^n |\theta_j| + \lambda_2 \sum_{j=1}^n \theta_j^2 \quad (17.7)$$

Where $J(\theta)$ is the cost function, m is the number of observations, n is the number of features, y_i is the dependent variable at the i -th observation, $h_{\theta}(x_i)$ is the model prediction for the i -th observation, θ_j are the model coefficients, λ_1 and λ_2 are the regularization parameters that control the strength of the L1 and L2 regularization respectively.

17.3.1.6 Random Forest Regression Model

Random Forest is a machine-learning technique for classification and regression problems. In regression, it allows building a predictive model for a quantitative response variable based on quantitative or qualitative variables.

It combines multiple regression trees to increase predictive capability. Each tree is built on a random sample with the replacement of the training data and on a random subset of features, which helps to increase diversity among the trees and reduce overfitting, as each tree captures different aspects and relationships in the data.

The prediction from each tree is obtained and averaged to produce a final estimation, which reduces the model's variance and provides more stable and accurate predictions.

Random Forest is known for its ability to handle large datasets with high dimensionality and non-linearities in the data. It is also robust against overfitting, requires little or no hyperparameter tuning, and is easily interpretable.

17.3.1.7 Support Vector Regression Model

The *Support Vector Regression* (SVR) model is a supervised learning technique used to solve regression problems on data that can be linear or non-linear.

In non-linear data, a *kernel* function is used to transform the input data from linearly inseparable low-dimensional features to a higher-dimensional space where the relationship between predictor variables and the response variable may be more

linear. Let $F = \{\varphi(x) \mid x \in X\}$ where F contains the transformations of elements from the dataset X . Each x in X is a set of features, and $\varphi(x)$ is a mapping function acting on x to produce a new set of features in F . If $x = \{x_1, x_2, \dots, x_n\}$ is a feature vector with n dimensions, then $\varphi(x) = \{\varphi_1(x), \varphi_2(x), \dots, \varphi_n(x)\}$ represents the result of applying the mapping function φ to each of the dimensions of x . Each $\varphi_i(x)$ can be a non-linear transformation of the corresponding x_i , which may allow capturing more complex relationships among the original features.

SVR aims to identify a hyperplane that minimizes the distance between the samples and the hyperplane, in other words, minimizing the difference between the model predictions and the real values, considering an error tolerance defined by a parameter ϵ . The regression function, defined in Eq. (17.8), is used to predict the value of the response variable for a new data point x . The loss function, defined in Eq. (17.9), penalizes deviations of the function above or below the error tolerance but does not penalize deviations within the error tolerance. Additionally, it may also include regularization terms that penalize the complexity of the model, which can help prevent overfitting by preventing the model from fitting too closely to the specific details of the training data.

$$f(x) = \langle w, \varphi(x) \rangle + b \quad (17.8)$$

Where w is the weight vector, $\varphi(x)$ is the mapping function to a potentially nonlinear higher-dimensional feature space, and b is the bias.

$$Loss = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (17.9)$$

Where $\frac{1}{2} \|w\|^2$ is a penalty on the magnitude of the model weights to favor simpler solutions, ξ_i represents the slack of training sample i on the positive side of the regression hyperplane, ξ_i^* represents the slack of training sample i on the negative side of the regression hyperplane, C is a regularization parameter that controls the trade-off between the error penalty and the model's flexibility. A low value allows for a larger margin and more violations of the margin rule, resulting in a smoother model and less prone to overfitting. A high value strongly penalizes margin violations, seeking to minimize the sum of the regression errors, resulting in a model that fits better to the training data.

17.3.1.8 Polynomial Regression Model

Polynomial regression is an extension of the linear regression model that allows capturing non-linear relationships between the independent variables and the dependent variable. Instead of fitting a straight line to the data, polynomial regression fits a polynomial of degree n , where n is an integer greater than or equal to 1. The general equation for a polynomial regression model is defined in Eq. (17.10).

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_n \cdot x_n + \epsilon \quad (17.10)$$

Where y is the dependent variable being predicted, x is the independent variable, $\beta_0, \beta_1, \beta_2, \dots, \beta_n$ are the coefficients of the model estimated during the fitting process, and ϵ is the error term, capturing the variability unexplained by the model.

The *polynomial features* technique captures non-linear relationships between predictor and response variables. This technique extends the feature space by introducing polynomial combinations of the original features. The idea is to transform the original features into a set of polynomial features, allowing the capture of non-linear relationships that a simple linear model would not detect. This involves generating new features by all combinations of the original features up to a specified degree. These new features are then used to fit a regression model.

Given a dataset $X = [x_1, x_2, \dots, x_n]$, where x_i is an observation of the original feature, the polynomial transformation of degree 2 would be *Polynomial Features* (X) = $[1, x_1, x_1^2, x_2, x_2^2, \dots, x_n, x_n^2]$ where each element of the new list represents a polynomial feature, including the constant term 1, the linear terms x_i , and the quadratic terms x_i^2 . This process can be generalized for polynomial transformations of higher degrees.

Polynomial features can increase the model's capacity to fit the data and improve its predictive ability, especially in cases where the relationship between predictor variables and the response variable is non-linear. However, it is important to note that increasing the dimensionality of the feature space can lead to an increase in the model's complexity and overfitting.

17.3.2 Evaluation Metrics

To determine the performance of the models, we used the metrics *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), and the *Coefficient of Determination* R^2 , commonly used in scientific literature. Table 17.3 presents each of the metrics and their description.

17.4 Methodology

This section presents the proposed methodology for estimating Stress Concentration Factors using machine learning algorithms. Figure 17.4 shows the proposed architecture for the regression models implemented.

To estimate the Stress Concentration Factors, we trained eight regression models, described in Sect. 17.3. The models were trained on a dataset resulting from the FEM, described in Sect. 17.2.2. The dataset consists of 48 observations, 2 predictor attributes (L/t and w/r), and 1 target attribute (SCF). We used *Python* 3.9.13 and

Table 17.3 Metrics used for the evaluation of the regression models

Metric	Description
$MSE = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$	It calculates the average of the squares of the differences between the model's predictions and the actual values. The lower the MSE, the better the model's fit to the data
$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2}$	The Root Mean Squared Error (RMSE) measures the magnitude of errors on the same scale as the dependent variable. Similar to MSE, a value close to 0 is preferred
$MAE = \frac{1}{N} \sum_{i=1}^N Y_i - \hat{Y}_i $	It calculates the average of the absolute differences between the model's predictions and the actual values. The lower the MAE, the better the model fits the data
$MAPE = \frac{100}{N} \sum_{i=1}^N \frac{ Y_i - \hat{Y}_i }{ Y_i }$	It measures the accuracy of a model's predictions in percentage by calculating the average absolute difference between the predicted values and the actual values of the dependent variable, normalized by the actual values. A value close to 0 indicates high accuracy in the model's predictions relative to the actual values
$R^2 = 1 - \frac{\sigma_r^2}{\sigma^2} = 1 - \frac{\sum_{i=1}^N (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}$	It measures the proportion of variance in the dependent variable explained by the independent variables. It ranges between 0 and 1, where 1 indicates a perfect fit of the model to the data and 0 indicates that the model does not explain the variability of the data better than a horizontal line

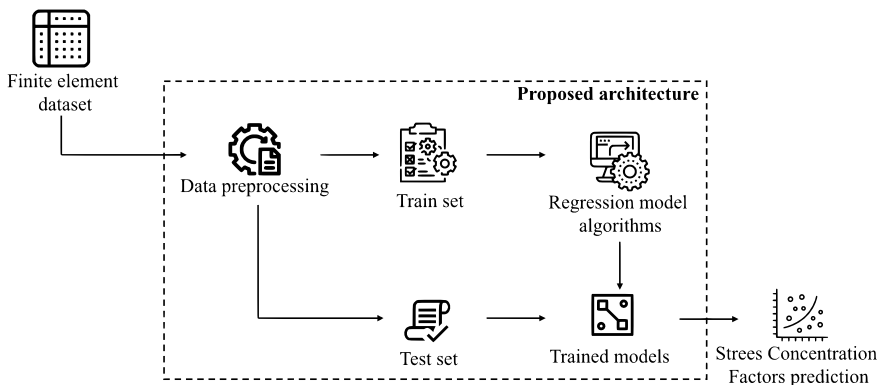


Fig. 17.4 Proposed architecture for the regression models

the libraries *SciKitLearn* 1.3.0, *scipy* 1.11.1, *pandas* 2.0.3, and *numpy* 1.24.3 to implement the models.

In Algorithm 1, the steps followed to train the regression models are presented.

The algorithm takes as input the dataset, which is stored in a dataframe, df , in Line 1.

In Lines 2 and 3, the target attribute is separated from the dataframe, and its values are stored in y .

After removing the target attribute, the values of the attributes L_t and w_r are scaled between 0 and 1 in Lines 4 and 5. To scale the data, the *MinMaxScaler* function from the *SciKitLearn* library is used, and the scaled values are stored in $df_{Normalized}$. Typically, data scaling is applied only to the predictor variables to ensure that the features have a uniform scale, which can improve the performance of certain machine learning algorithms, especially those sensitive to feature scaling.

In Lines 6 and 7, the *BoxCox* function from the *scipy* library is used to transform the data stored in $df_{Normalized}$ to stabilize the variance and make the data follow a normal distribution. The transformed data is stored in $df_{Transformed}$.

In Line 8, the values of the attributes L_t and w_r are stored in X .

Lines 9–17 create the regression models and store them in the list *regressors*.

Line 18 implements stratified cross-validation, which divides the dataset into 5 folds; 4 are used to train the regression models, and 1 is used to evaluate the models' performance. The folds are rotated in each iteration to train and evaluate the models on different parts of the dataset, providing evidence of their robustness. The observations in each fold are determined by the indices specified in the lists $train_{index}$ and $test_{index}$.

In Line 19, the observations used to train the models in each fold are defined (X_{train}), and in Line 20, the observations of the target attribute (y_{train}) are defined. Similarly, Lines 21 and 22 define the observations used to evaluate the models (X_{test} and y_{test}).

In each iteration of the loop in Line 23, each regressor (*regressor*) stored in the list *regressors* is accessed. In Line 24, each regressor is trained on the sets X_{train} and y_{train} .

In Line 25, predictions are made on the set X_{test} and stored in *predictions*.

In Lines 26–31, performance metrics are calculated and visualized by comparing *predictions* with the actual values of y_{test} . The metrics used to evaluate the models' performance include Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, and the Coefficient of Determination R^2 .

Finally, predictions are returned in Line 34.

Algorithm 1: Algorithm developed for implementing regression models

Input: *Filename* (Dataset)

Output: SCF predictions

```

1 df ← OpenDataset( Filename )
2 y ← df['SCF'].values
3 df ← df.drop['SCF']
4 scaler ← MinMaxScaler()
5 dfNormalized ← scaler.fit_transform(df.values)
6 dfTransformed['L/t'] ← BoxCox( dfNormalized['L/t'] )
7 dfTransformed['w/r'] ← BoxCox( dfNormalized['w/r'] )
8 X ← dfTransformed['L/t', 'w/r']
9 regressors ← []
10 regressors.append( LinearRegression() )
11 regressors.append( RANSACRegressor() )
12 regressors.append( Ridge() )
13 regressors.append( Lasso() )
14 regressors.append( ElasticNet() )
15 regressors.append( RandomForestRegressor() )
16 regressors.append( SVR() )
17 regressors.append( PolynomialFeatures() )
18 foreach trainindex, testindex in StratifiedKfold(folds=5,
   data=X, target=y) do
19   Xtrain ← dfTransformed[trainindex]
20   ytrain ← dfTransformed[trainindex]
21   Xtest ← dfTransformed[testindex]
22   ytest ← dfTransformed[testindex]
23   foreach regressor in regressors do
24     regressor.fit(Xtrain, ytrain)
25     predictions ← regressor.predict(Xtest)
26     metrics[MAE] ← mean_absolute_error(ytest, predictions)
27     metrics[MAPE] ←
       mean_absolute_percentage_error(ytest, predictions)
28     metrics[MSE] ← mean_squared_error(ytest, predictions)
29     metrics[RMSE] ← sqrt(mean_squared_error(ytest,
       predictions))
30     metrics[R2] ← r2_score(ytest, predictions)
31     print(metrics)
32   end
33 end
34 return predictions

```

Table 17.4 Hyperparameters of the regression models

Regression models	Hyperparameters
Linear regression	fit_intercept = True positive = False
RANSAC	estimator='LinearRegression' loss = 'absolute_error' max_trials=150 min_samples=None stop_probability=0.8
Ridge	alpha = 0.1 fit_intercept = True positive = False solver = 'saga'
Lasso	alpha = 0.0001 fit_intercept = True positive = True selection = 'cyclic'
Elastic Net	alpha = 0.0001 fit_intercept = True l1_ratio = 0.1 positive = False selection = 'random'
Random forest	n_estimators = 50 bootstrap = True criterion = 'squared_error'
SVR	kernel = 'rbf' C = 1000000 epsilon = 0.01
Polynomial features	degree = 2 interaction_only = True include_bias = True order = 'C'

The hyperparameters of the regression models were adjusted using the *Grid-SearchCV* function from the *SciKitLearn* library. This function explores all possible combinations of the specified hyperparameter values and selects the combination that yields the best performance. The determined hyperparameters for each model are presented in Table 17.4.

Table 17.5 Average results obtained by the regression models in the cross-validation

Regression model	MSE	RMSE	MAE	MAPE	R ²
Linear regression	0.00	0.02	0.01	0.01	0.88
RANSAC	0.00	0.02	0.01	0.00	0.89
Ridge	0.00	0.02	0.01	0.01	0.88
Lasso	0.00	0.02	0.02	0.01	0.83
Elastic Net	0.00	0.02	0.01	0.01	0.88
Random Forest	0.00	0.01	0.01	0.00	0.93
SVR	0.00	0.01	0.01	0.00	0.92
Polynomial features	0.00	0.02	0.01	0.00	0.89

17.5 Results

This section describes the evaluation of the eight regression models conducted to identify the model that best fits the data in this study case.

A cross-validation technique was used in the evaluation. This technique divides the dataset into k different subsets, called *folds*. The models are trained on $k - 1$ folds, and the remaining fold is used to evaluate the performance of the models, applying the metrics shown in Table 17.3. This process iterates k times, and the models are evaluated on a different fold in each iteration. Cross-validation allows for a more robust evaluation of the model's performance by providing a more precise estimation of prediction error on unseen data. In the evaluation, we used a 5-fold cross-validation. Table 17.5 presents the average results of the models obtained using the metrics in each of the 5 folds, highlighting in bold the best values obtained in each metric, and Table 17.6 presents the details of each fold.

As observed, all models exhibited similar performance, with values close to 0 in the metrics *MSE*, *RMSE*, *MAE*, and *MAPE*, which quantify the error between the model predictions and the real values. This indicates that the models provide accurate and consistent predictions relative to the real values.

However, *SVR* and *Random Forest* showed similar performance based on the results obtained in the R² metric, which quantifies the variability in the dependent variable explained by the model.

Based on the results presented in Table 17.5, the best-performing model that fit the data in this case study was *Random Forest*. Figure 17.5 shows the real values and the predictions to visually evaluate the quality of the predictions made by the model. The predictions should be close to the diagonal line since the closer they are, the more accurate they are.

Finally, the QQ-Plot function was used to analyze the residuals of the model. Residual analysis allows us to assess the quality of the model fit and identify patterns that the model may have missed, such as non-linearities or non-linear relationships between variables. Additionally, it enables the comparison of the distribution of a data sample with a theoretical distribution, such as the normal distribution. The closer

Table 17.6 Results obtained by the regression models in each of the 5 folds

Regression models	Metrics	1	2	3	4	5	Average
Linear regression	MSE	0.00	0.00	0.00	0.00	0.00	0.00
	RMSE	0.02	0.02	0.02	0.02	0.01	0.02
	MAE	0.02	0.01	0.01	0.02	0.01	0.01
	MAPE	0.01	0.01	0.00	0.01	0.00	0.01
	R2	0.72	0.91	0.91	0.91	0.97	0.88
RANSAC	MSE	0.00	0.00	0.00	0.00	0.00	0.00
	RMSE	0.02	0.02	0.02	0.01	0.01	0.02
	MAE	0.02	0.01	0.01	0.01	0.01	0.01
	MAPE	0.01	0.00	0.00	0.00	0.00	0.00
	R2	0.72	0.92	0.90	0.95	0.96	0.89
Ridge	MSE	0.00	0.00	0.00	0.00	0.00	0.00
	RMSE	0.02	0.02	0.02	0.02	0.01	0.02
	MAE	0.02	0.01	0.01	0.02	0.01	0.01
	MAPE	0.01	0.01	0.00	0.01	0.00	0.01
	R2	0.73	0.91	0.91	0.90	0.97	0.88
Lasso	MSE	0.00	0.00	0.00	0.00	0.00	0.00
	RMSE	0.03	0.02	0.02	0.02	0.02	0.02
	MAE	0.02	0.02	0.01	0.02	0.01	0.02
	MAPE	0.01	0.01	0.00	0.01	0.01	0.01
	R2	0.56	0.88	0.89	0.88	0.94	0.83
Elastic Net	MSE	0.00	0.00	0.00	0.00	0.00	0.00
	RMSE	0.02	0.02	0.02	0.02	0.01	0.02
	MAE	0.02	0.01	0.01	0.02	0.01	0.01
	MAPE	0.01	0.01	0.00	0.01	0.00	0.01
	R2	0.72	0.91	0.91	0.90	0.97	0.88
Random forest	MSE	0.00	0.00	0.00	0.00	0.00	0.00
	RMSE	0.02	0.01	0.01	0.01	0.00	0.01
	MAE	0.02	0.00	0.01	0.01	0.00	0.01
	MAPE	0.01	0.00	0.00	0.00	0.00	0.00
	R2	0.70	0.99	0.99	0.98	0.99	0.93
SVR	MSE	0.00	0.00	0.00	0.00	0.00	0.00
	RMSE	0.02	0.01	0.01	0.01	0.02	0.01
	MAE	0.01	0.01	0.01	0.01	0.02	0.01
	MAPE	0.00	0.00	0.00	0.00	0.01	0.00
	R2	0.83	0.97	0.94	0.97	0.89	0.92
Polynomial features	MSE	0.00	0.00	0.00	0.00	0.00	0.00
	RMSE	0.02	0.02	0.01	0.02	0.01	0.02
	MAE	0.01	0.01	0.01	0.02	0.01	0.01
	MAPE	0.01	0.00	0.00	0.01	0.00	0.00
	R2	0.76	0.93	0.91	0.89	0.97	0.89

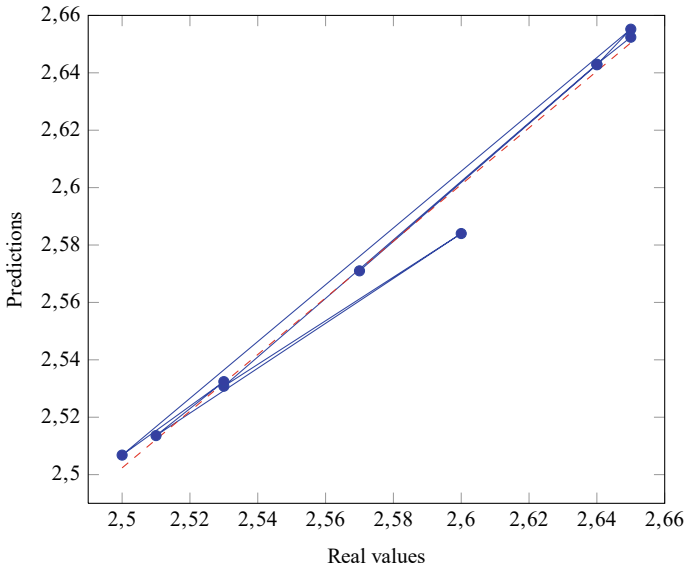


Fig. 17.5 Real values versus predictions (random forest regression)

the points are to the diagonal line, the better the data fits a normal distribution. Conversely, points far from the diagonal line could indicate the presence of outliers or a lack of model fit to the data. Since Random Forest does not have a traditional regression function like linear regression, the residuals were calculated by the difference between the actual values and the predictions made by the model. Figure 17.6 shows the QQ-Plot results. As can be seen, the points are close to the diagonal line except for one point, which could be considered an outlier.

17.6 Conclusions

This chapter presented eight regression models to estimate the Stress Concentration Factors (SCF) of rectangular plates. The models implemented were Multiple Linear Regression, Random Sample Consensus, Ridge Regression, LASSO Regression, Elastic Net, Random Forest Regression, Support Vector Regression, and Polynomial Regression.

The models were trained on a dataset resulting from a two-dimensional Finite Element Analysis from the Finite Element Method for different values of the parameters: large L/t , width w/r , and circular hole radius in a tensile plate. Least squares polynomial equations were fitted to these design points. It is important to mention that the implementation of these models is limited to the prediction of the SCF with input data within the interval with which the models learned. These are input parameter values $50 \leq L/t \leq 200$ and $32/3 \leq w/r \leq 19$.

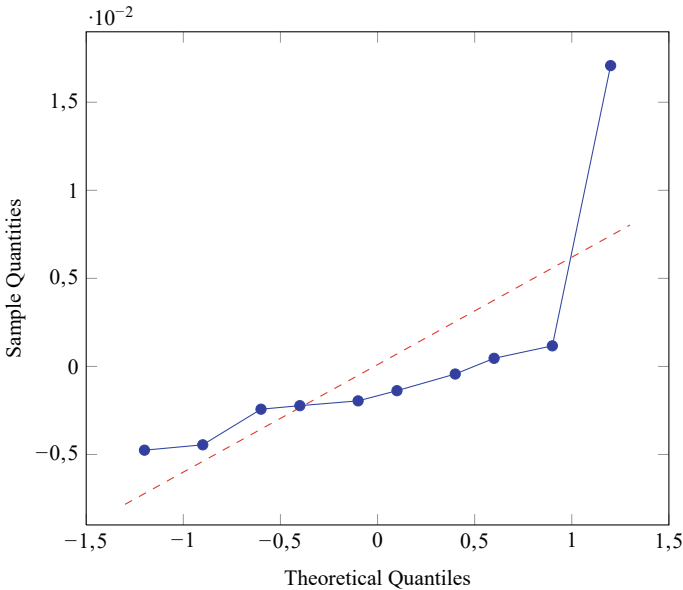


Fig. 17.6 Random forest QQ-plot

The equations for estimating the tensile stress concentration factor k_t were presented. A .mac file of the finite element model of the plate with a circular hole and a structured mesh was created. This file was executed in ANSYS APDL for all combinations of the parameters w/r and L/t here analyzed. With this, Von Mises' maximum stresses were obtained, and in turn, the stress concentration factors through the Eq. (17.3).

The curve fitting technique was used to obtain a polynomial that estimated the SCF as a polynomial function of the input parameters w/r and L/t . The relative percentage errors were estimated for all stress concentration factors obtained by curve fitting with respect to those obtained by finite elements, finding a maximum error of 1.658 %.

The metrics used to evaluate the performance of the regression models were Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, and the Coefficient of Determination R^2 , commonly used in scientific literature. Despite all eight models achieving good results, the presented outcomes indicate that Random Forest performed the best, followed by Support Vector Regression.

These regression algorithms acceptably described the SCF in the hole plate. This shows its effectiveness for its potential use in other mechanical components, such as beams and shafts, as well as stress concentrators with different geometries.

Table 17.7 Stress concentration factors for different parameters dimensions; large L , height w , radius r and thickness t

L/t	w/r	FE			Curv fit	
		k_t	k_t	k_t	k_t	%
200	19	2.66	2.66	2.66	0.080	
150	19	2.66	2.66	2.66	0.035	
125	19	2.67	2.66	2.66	0.108	
100	19	2.68	2.67	2.67	0.655	
200	18	2.65	2.65	2.65	0.040	
150	18	2.65	2.65	2.65	0.181	
125	18	2.65	2.65	2.65	0.100	
100	18	2.67	2.66	2.66	0.415	
200	17	2.63	2.63	2.63	0.034	
150	17	2.63	2.64	2.64	0.287	
125	17	2.64	2.64	2.64	0.274	
100	17	2.65	2.65	2.65	0.191	
200	16	2.62	2.61	2.61	0.057	
150	16	2.62	2.62	2.62	0.359	
125	16	2.62	2.63	2.63	0.420	
100	16	2.63	2.63	2.63	0.028	
200	38/3	2.55	2.54	2.54	0.175	
150	38/3	2.55	2.56	2.56	0.214	
125	38/3	2.57	2.57	2.57	0.220	
100	38/3	2.62	2.57	2.57	1.658	
200	12	2.53	2.52	2.52	0.178	
150	12	2.53	2.54	2.54	0.341	
125	12	2.55	2.55	2.55	0.017	
100	12	2.60	2.56	2.56	1.387	
200	34/3	2.51	2.51	2.51	0.160	
150	34/3	2.51	2.53	2.53	0.481	
125	34/3	2.53	2.53	2.53	0.286	
100	34/3	2.57	2.54	2.54	1.036	
200	32/3	2.49	2.49	2.49	0.111	
150	32/3	2.49	2.51	2.51	0.640	
125	32/3	2.50	2.52	2.52	0.587	
100	32/3	2.54	2.53	2.53	0.593	
62.5	19	2.67	2.67	2.67	0.117	
50	19	2.68	2.67	2.67	0.467	
62.5	18	2.65	2.66	2.66	0.376	
50	18	2.67	2.66	2.66	0.197	
62.5	17	2.64	2.65	2.65	0.609	
50	17	2.65	2.65	2.65	0.065	
62.5	16	2.62	2.64	2.64	0.821	
50	16	2.63	2.64	2.64	0.329	
62.5	38/3	2.57	2.58	2.58	0.451	
50	38/3	2.62	2.59	2.59	1.160	
62.5	12	2.55	2.57	2.57	0.757	
50	12	2.60	2.57	2.57	0.835	
62.5	34/3	2.53	2.55	2.55	1.099	
50	34/3	2.57	2.56	2.56	0.424	
62.5	32/3	2.50	2.54	2.54	1.479	
50	32/3	2.54	2.54	2.54	0.083	

References

1. Braun, M., Kellner, L.: Comparison of machine learning and stress concentration factors-based fatigue failure prediction in small-scale butt-welded joints. *Fatigue Fract. Eng. Mat. Struct.* **45**(11), 3403–3417 (2022). <https://doi.org/10.1111/ffe.13800>
2. Kirsch, C.: Die theorie der elastizität und die bedürfnisse der festigkeitslehre. *Zeitschrift des Vereines Deutscher Ingenieure* **42**, 797–807 (1898)
3. Mehmet, E., TOKTAŞ, İ., ÖZKAN, M.T.: Modeling of stress concentration factor using artificial neural networks for a flat tension bar with opposite v-shaped notches. *Politeknik Dergisi*, 1 (2023). <https://doi.org/10.2339/politeknik.1275466>
4. Monares, J.A.R., Soto, L.G.: Stress concentration reduction in an axially loaded rectangular bar with an elliptical hole. *CULCyT: Cultura Científica y Tecnológica* **20**(1), 14–21 (2023). <https://doi.org/10.20983/culcyt.2023.1.2.2>
5. Pilkey, W.D., Pilkey, D.F., Bi, Z.: *Peterson's Stress Concentration Factors*. John Wiley & Sons, Hoboken, New Jersey (2008). <https://doi.org/10.1002/9780470211106>
6. Stowell, E.Z.: *Stress and Strain Concentration at a Circular Hole in an Infinite Plate*. National Advisory Committee for Aeronautics Washington, DC, Washington, United States of America (1950)
7. Troyani, N., Gomes, C., Sterlacci, G.: Theoretical stress concentration factors for short rectangular plates with centered circular holes. *J. Mech. Des.* **124**(1), 126–128 (2002). <https://doi.org/10.1115/1.1412849>
8. Wang, B., Zhao, W., Du, Y., Zhang, G., Yang, Y.: Prediction of fatigue stress concentration factor using extreme learning machine. *Comp. Mat. Sci.* **125**, 136–145 (2016). <https://doi.org/10.1016/j.commatsci.2016.08.035>
9. Yang, Z., Kim, C.B., Cho, C., Beom, H.G.: The concentration of stress and strain in finite thickness elastic plate containing a circular hole. *Int. J. Solid. Struct.* **45**(3–4), 713–731 (2008). <https://doi.org/10.1016/j.ijsolstr.2007.08.030>
10. Young, W.C., Budynas, R.G., Sadegh, A.M.: *Roark's formulas for stress and strain*. McGraw-Hill Education, New York (2002)

Chapter 18

A Performance Analysis of Technical Indicators on the Indian Stock Market



Hetvi Waghela , Jaydip Sen , and Sneha Rakshit 

Abstract This chapter delves into three powerful and widely-used technical indicators, Bollinger Bands, Moving Average Convergence Divergence (MACD), and the Relative Strength Index (RSI), and makes a comparative study of the effectiveness of these indicators on the Indian stock market. For this study, stocks were chosen from 14 sectors listed on India's National Stock Exchange (NSE). The top stocks of each sector are identified based on their free-float market capitalization from the NSE's report published on July 1, 2022 (NSE Website). For each stock in 14 sectors, trading was done for one year from July 1, 2022, to June 30, 2023, with an initial capital of Indian Rupees (INR) 100,000 following the three technical indicators. The technical indicator that yielded the highest return is identified for each stock. A comparative analysis is made based on the overall performance of the three indicators for all 14 sectors.

Keywords Technical indicators · Bollinger bands · Moving average convergence divergence · Relative strength indicator · Return · Risk

18.1 Introduction

In the ever-changing realm of financial markets, investors and traders are looking for techniques and strategies that can offer them valuable perspectives on potential price changes and assist them in making well-founded choices. Technical analysis emerges as a leading method in this pursuit, presenting a structured method for analyzing past

H. Waghela · J. Sen (✉) · S. Rakshit
Praxis Business School, Kolkata, India
e-mail: jaydip.sen@acm.org

H. Waghela
e-mail: waghelah@acm.org

S. Rakshit
e-mail: srakshit149@gmail.com

price data and recognizing trends and patterns that could influence future market dynamics.

Technical analysis entails assessing and forecasting future price shifts of stocks and other financial assets by scrutinizing past price data and trading volumes. It operates under the assumption that past price movements and trading behaviors can offer clues about future price shifts. This methodology employs various tools and indicators to scrutinize charts, recognize trends, support, resistance levels, and potential trade entry and exit points.

A robust technical analysis involves the construction of charts, trend lines, indicators, chart patterns, volume analysis, and candlestick patterns. There are several well-known technical indicators such as Bollinger bands, moving average convergence divergence (MACD), relative strength index (RSI), stochastic oscillator, Fibonacci retracement, average directional index (ADX), on-balance volume (OBV), volume-weighted average price (VWAP), parabolic-SAR, etc. However, Bollinger Band, MACD, and RSI are the three most important technical indicators for the following reasons [1]. First, these three indicators together provide a comprehensive view of market conditions. Bollinger Bands are effective for identifying volatility and price action points, MACD is good for detecting trend and momentum, and RSI is for identifying relative strength and potential reversal point. Second, these indicators can be applied across different time frames and are useful for short-term and long-term trading strategies. Finally, these indicators have been widely tested and validated across different market conditions, providing a level of trust and reliability.

This chapter explores three influential and commonly used technical indicators: Bollinger Bands, Moving Average Convergence Divergence (MACD), and the Relative Strength Index (RSI). It conducts a comparative analysis to assess their effectiveness in the context of the Indian stock market. Stocks from 14 sectors listed on the National Stock Exchange (NSE) of India were selected for the study. According to the NSE's report, the top stocks from each sector were identified based on their free-float market capitalization as of July 1, 2022. Trading activities were conducted for one year, from July 1, 2022, to June 30, 2023, using an initial capital of Indian Rupees (INR) 100,000, following the signals provided by these three technical indicators. The technical indicator that resulted in the highest return for each stock was determined, and a comparative analysis was performed based on the collective performance of the three indicators across all 14 sectors.

This study offers three distinct contributions. First, it demonstrates the effective utilization of significant technical indicators—Bollinger Bands, Moving Average Convergence and Divergence (MACD), and Relative Strength Index (RSI)—in trading within the Indian stock market. Second, it introduces a comparative framework to comprehend the efficacy of these three indicators in generating investment returns. Third, the findings of this research offer profound insights into the present profitability of various sectors, serving as valuable guidance for investors operating within the Indian stock market.

The organization of the chapter unfolds as follows. Section 18.2 outlines various stock price prediction models in the academic literature, including technical indicators-based analysis and portfolio design approaches. Section 18.3 details the

research approach adopted in the present study. Section 18.4 provides a comprehensive set of results accompanied by a thorough analysis of the same. Section 18.5 discusses the current work's possible applications of AI and machine learning algorithms and models. Finally, Sect. 18.6 concludes the chapter.

18.2 Related Work

Several methods have been proposed in literature to tackle the complex task of accurate stock price prediction and optimizing stock combinations to enhance investment returns. Scholars have extensively utilized machine learning models in forecasting future stock prices [2–10].

The use of deep learning architectures and algorithms has boosted the accuracy of prediction models [3, 5–7, 9, 11–22]. Various text mining techniques have also been successfully applied on social media platforms and the internet, improving accuracy in predicting stock prices [6, 20, 23–26].

Among various methods for forecasting stock prices, there is considerable interest in employing statistical and econometric approaches based on time series decomposition [3, 27–39].

In some studies, various types of the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model have been used to predict future volatility and assess stock portfolio risk [40].

In recent years, reinforcement learning has seen widespread adoption for accurately predicting stock prices and crafting reliable portfolios [41–50].

The conventional mean–variance optimization method is prominently acknowledged as the most widely accepted approach for portfolio optimization [51–56].

Several scholars have proposed alternative methods to the mean–variance approach for portfolio optimization. Notable among these techniques include (i) multi-objective optimization [57, 58], (ii) use of different ratios in the mean–variance optimization [59], (iii) Eigen portfolios derived from principal component analysis [51, 60], (iv) methods based on risk parity [60–64], and (v) approaches utilizing swarm intelligence [65, 66]. Additionally, the literature suggests the utilization of genetic algorithms [67], fuzzy sets [68], prospect theory [69], and quantum evolutionary algorithms [70].

Certain scholars have proposed pair-trading portfolios, which involve a pair of stocks, as an alternative to portfolios consisting of multiple stocks [71–75].

Seshu et al. introduced an approach capable of evaluating the performance of various automated trading strategies using different metrics [76]. Their proposed method utilizes predictions from two strategies: Bollinger Bands and Long Short-Term Memory (LSTM) networks. The LSTM strategy incorporates forecasts from 250 LSTM neural networks (with 5 models per company). In contrast, the Bollinger Bands strategy relies on close price, simple moving averages, and standard deviations for trading decisions. These strategies' performance was assessed through backtesting with historical and real-time data from stocks in the NIFTY50 index

[77]. The findings indicate that the custom strategies surpassed market benchmarks in 35.93% of tested periods, demonstrating higher returns than investing in the stock market index during the same periods.

Zheng et al. developed a hybrid predictive model that adheres to conservative risk hedging and split-position trading principles. This model integrates the Bollinger Bands strategy with a regression polynomial combination model to reconfigure position distribution [78]. It successfully captured the temporal fluctuations of gold and bitcoin, exhibiting high total returns, low transaction costs, and minimal maximum drawdown.

Lauguico et al. developed an algorithm that uses three fuzzy logic controllers to carry out a specific trading strategy [79]. This strategy incorporates technical indicators, such as candlestick parameters and Bollinger Bands (BB), to evaluate the strength of buy, hold, and sell signals. Stock price data, including opening and closing prices, are gathered from a particular company for BB calculations. These raw and computed values are then used as precise input parameters for the Fuzzy Inference System (FIS). Membership functions are categorized into very low, low, high, and very high levels based on typical input parameters used by traders. Fuzzy logic rules are established to generate signals indicating the strength of trade execution recommendations. Implemented using NI LabVIEW and MATLAB, the system demonstrated satisfactory results, achieving an accuracy of approximately 94.44%.

Au & Keung argue that while MACD is straightforward to interpret, it suffers from two notable drawbacks, the time-lagging problem and the issue of generating false signals, leading to delays in decision-making for buying or selling [80]. The authors introduce a novel approach called the volume square-weighted moving average convergence and divergence (VSWMACD) to improve the performance of MACD. The proposed methodology is subjected to various evaluation tools to validate the improvements. Testing is conducted on five datasets, each containing 200 stocks from the Hong Kong Stock Market. The results indicate that, compared to MACD, VSWMACD yields an approximately 15% increase in the average return and a reduction of around 5% in the average maximum drawdown.

Deac and Iancu suggest using a genetic algorithm (GA) to optimize two strategies: a crossover of MACD strategy and an ensemble strategy that combines MACD with RSI, specifically applied to Nvidia stock using daily data [81]. The GA determines the optimal parameter sets for MACD and MACD-RSI, which are significant technical indicators in trading. Their work provides a detailed description of the design of the GA, the metrics used, the overall architecture, and the indicators.

Chen et al. introduce a novel method to enhance the effectiveness of the widely used technical indicator, the RSI [82]. They utilize an algorithm of metaheuristics known as GNQTS to efficiently determine the optimum values of the RSI parameters. Their approach also includes a sliding window technique to dynamically adjust training periods, reducing the risk of overfitting. The study covers major indices and companies in the U.S. stock market. The results indicate that GNQTS effectively identifies optimal RSI parameters, resulting in higher profits than traditional RSI and buy-and-hold strategies.

Zatwarnicki et al. presents an algorithmic methodology for assessing the efficacy of signals produced by the RSI. Backtesting of the strategies was conducted using a model mirroring an authentic cryptocurrency exchange [83]. The results suggest that using RSI as a momentum gauge in the cryptocurrency market carries significant risks. Investigating alternative uses of RSI may offer traders an advantage in this market.

Overall, integrating advanced computational techniques and innovative methodologies continues to push the boundaries of stock price prediction and portfolio optimization, offering promising avenues for financial decision-making. A summary of all the approaches discussed in this section is presented in Table 18.1.

The current work explores three influential and commonly employed technical indicators: Bollinger Bands, MACD, and RSI. It presents a comparative analysis to evaluate the effectiveness of these indicators across 14 sectors within the Indian stock market. To the authors' knowledge, no prior studies have pursued this specific direction. Consequently, the findings from this research are anticipated to provide valuable insights for investors in the Indian stock market.

18.3 Methodology

This section presents the methodology employed in this study, specifically emphasizing the steps undertaken to identify the signal points for buy and sell decisions in trading based on the three technical indicators, Bollinger bands, MACD, and RSI. The methodology encompasses a sequence of eight steps, which are discussed below.

- (i) **Choice of the sectors for analysis:** Fourteen diverse sectors are initially chosen from those listed on the NSE to represent a cross-section of the Indian stock market. These selected sectors include banking, auto, consumer durables, financial services excluding banks, information technology (IT), fast-moving consumer goods (FMCG), metal, media, oil and gas, mid-small IT and telecom, private banks, pharma, realty, and PSU banks. The monthly reports from the NSE identify the ten stocks with the highest free-float capitalization in each sector. For this study, the report released on June 30, 2022, is utilized to select the ten stocks from each of the fourteen sectors and the 50 stocks from NIFTY 50 (NSE Website) [84].
- (ii) **Extraction of historical stock prices from the web:** Using the DataReader function from the pandas_datareader module in Python, the historical daily prices of the stocks are obtained from the Yahoo Finance website for the period from July 1, 2022, to June 30, 2023. The *close* values of the stocks are used in computing three technical indicators, Bollinger band (BB), moving average convergence divergence (MACD), and relative strength indicator (RSI).
- (iii) **Computation of the simple moving averages of prices:** The 20-day rolling simple moving average values for the *close* prices are computed for every stock in each sector. For the computation of the 20-day rolling average, the

Table 18.1 Summary of approaches to stock price prediction and portfolio optimization

Category	Approach	Description	References
Machine learning	Building various machine learning models	Utilization of historical data to identify patterns and predict future stock prices	[2–10]
Deep learning	LSTM, CNN, and other hybrid deep-learning models	Advanced architectures capturing complex patterns and dependencies in financial data	[3, 5–7, 9, 11–22]
Text mining	Analyzing social media and web data	Extracting sentiment and information from unstructured text to enhance stock price prediction	[6, 20, 23–26]
Econometric	Time series decomposition, GARCH models	Statistical and econometric methods for modeling and forecasting stock prices and volatility	[3, 27–40]
Reinforcement learning	Various RL algorithms	Learning optimal trading and portfolio design through trial and error to maximize returns	[41–50]
Classical optimization	Mean–variance optimization	Balancing expected return and risk in portfolio construction	[51–55]
Multi-objective optimization	Multi-objective optimization approaches	Balancing multiple objectives, such as return and risk, simultaneously	[56, 57]
Eigen portfolio	Principal component analysis	Constructing portfolios that capture the most significant market	[51, 60]
Risk parity-based methods	Risk parity-based portfolio optimization	Allocating assets to achieve equal risk contribution from each asset	[60–64]
Swarm intelligence	Swarm intelligence-based optimization	Leveraging evolutionary and bio-inspired algorithms for portfolio optimization	[65, 66]
Genetic algorithms	Genetic algorithm optimization	Using genetic algorithms to optimize trading strategies and portfolio parameters	[67, 81]
Advanced theoretical frameworks	Fuzzy sets, prospect theory, quantum evolutionary algorithms	Utilizing advanced mathematical and theoretical frameworks for decision-making under uncertainty	[68–70]
Pair-trading portfolios	Pair-trading strategies	Investing in two stocks to exploit relative price movements	[71–75]
Custom trading strategies	Combination of LSTM and Bollinger Bands	Integrating LSTM networks and Bollinger Bands for superior performance through backtesting	[76]

(continued)

Table 18.1 (continued)

Category	Approach	Description	References
Hybrid predictive models	Bollinger Bands and polynomial regression combination	Combining Bollinger Bands with regression models to capture market fluctuations	[78]
Fuzzy logic controllers	Fuzzy logic-based trading strategy	Using fuzzy logic to interpret technical indicators and generate trading signals with high accuracy	[79]
Volume-weighted moving averages	Volume square-weighted MACD	Addressing the drawbacks of traditional MACD to improve returns and reduce drawdowns	[80]
Optimizing technical indicators	Genetic algorithm optimization of MACD-RSI and GNQTS for RSI	Enhancing trading performance by fine-tuning parameters of technical indicators using genetic algorithms and other metaheuristic methods	[81, 82]
Algorithmic assessment of RSI	Backtesting RSI in a volatile market such as cryptocurrency	Evaluating the effectiveness and risks of using RSI as a momentum indicator	[83]

rolling function in Python is used with a parameter value of 20, and then the *mean* function is applied over the 20 observations.

(iv) **Computation of the Bollinger bands:** The concept of Bollinger bands was introduced by John Bollinger in 1980. Bollinger Bands consist of three lines: the middle band, the upper band, and the lower band. These bands are based on a *simple moving average* (SMA) and standard deviation. In the following, the computation of the bands is discussed.

(a) *The computation of the middle band:* To compute the middle band, the period N for the computation of the SMA is chosen first. The most common value of N is 20. Next, the SMA for each period is computed by summing up the stock's closing prices over N periods and dividing the sum by the number of periods. The computation of the middle band is shown in (18.1).

$$\text{Middle Band (SMA)} = \frac{\text{Sum of closing prices over } N \text{ periods}}{N} \quad (18.1)$$

(b) *The computation of the standard deviation:* In this step, the standard deviation of the closing prices for N observations is computed using (18.2).

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}} \quad (18.2)$$

In (18.2), X_i is the closing price for each day, and \bar{X} is the mean closing price over N periods.

- (c) *The computation of the upper and lower bands:* First, a factor K is chosen to compute the upper and the lower bands. The usual value of K is 2. The upper band is computed by adding K times the standard deviation to the middle band, as in (18.3). Similarly, the lower band is computed by subtracting K times the standard deviation from the middle band, as shown in (18.4).

$$\text{Upper Band} = \text{Middle Band} + (K * \text{Standard Deviation}) \quad (18.3)$$

$$\text{Lower Band} = \text{Middle Band} - (K * \text{Standard Deviation}) \quad (18.4)$$

The three Bollinger Bands are used extensively to analyze market conditions and identify trends and potential reversal points. In a strong uptrend, prices often touch or exceed the upper band, while in a strong downtrend, prices may touch or fall below the lower band. Traders use this information to identify the direction of the trend. Bollinger Bands expand and contract based on market volatility. Wide bands indicate volatility, while narrow bands suggest low volatility. Traders can use this information to gauge the market environment and adjust their strategies accordingly. Sudden price movements that cause the bands to expand can be interpreted as a volatility breakout. Traders might look for opportunities to enter trades in the direction of breakout. When prices touch or exceed the upper band, it may indicate an overbought condition, suggesting a potential reversal to the downside.

Conversely, when prices touch or fall below the lower band, it may signal an oversold condition and a potential reversal to the upside. Traders often look for divergence or convergence between price and the Bollinger Bands. For example, if prices are making new highs, but the upper band is expanding, it could signal a weakening trend.

- (v) *Computation of the Moving Average Convergence Divergence (MACD):* MACD is a popular momentum indicator used in technical analysis to identify potential trend reversals, generate trading signals, and assess the strength of a trend. The concept was first introduced by Gerald Appel in 1970. The MACD indicator is calculated using two *exponential moving averages* (EMAs) of an asset's price. The two main components of the MACD are (a) MACD line (the fast line), which is the difference between a short-term EMA (usually based on 12 periods) and a longer-term EMA (usually 26 periods), and (b) Signal line (the slow line): this is the 9-day EMA of the MACD line that is used to generate trading signals. The computation of the MACD lines involves the following steps.

- (a) *The computation of the Short-Term EMA (Exponential Moving Average):* To compute the short-term EMA, the number of days of the short-term EMA is first chosen. The usual length of this period is 12 days. The EMA for each day is computed using the closing prices of stocks using (18.5).

$$EMA_{short-term} = \left(CP * \frac{2}{STP + 1} \right) + \left(PEMA * \left(1 - \frac{2}{STP + 1} \right) \right) \quad (18.5)$$

In (18.5), CP, STP, and PEMA denote the *closing price*, *short-term period*, and *previous EMA*, respectively. The initial EMA is usually the SMA of the first day's closing prices.

- (b) *The computation of the MACD line (fast line):* The MACD line is computed by subtracting the *long-term EMA* (EMA_{LT}) from the *short-term EMA* (EMA_{ST}) as in (18.6).

$$MACD\ Line = EMA_{ST} - EMA_{LT} \quad (18.6)$$

- (c) *The computation of the Signal line (slow line):* To compute the *signal line*, first, the number of periods for the signal line is chosen. The most used value for the number of periods for the signal line is 9. The signal line is the EMA of the MACD line computed based on the signal line period as in (18.7).

$$Signal\ Line = EMA(MACD\ Line, Signal\ Line\ Period) \quad (18.7)$$

- (d) *The computation of the MACD histograms:* In the final step, the MACD histograms are computed by subtracting the signal line from the MACD line as in (18.8).

$$MACD\ Histogram = MACD\ Line - Signal\ Line \quad (18.8)$$

The resulting MACD histograms visually represent the difference between the MACD line and the signal line. When the MACD line crosses above the signal line, it generates a bullish signal, indicating a potential upward momentum. Conversely, when the MACD line crosses below the signal line, it generates a bearish signal, indicating potential downward momentum. Divergence occurs when the asset price and the MACD indicator move in opposite directions, while convergence occurs in the same direction. Divergence can be a sign of a potential reversal. When the MACD histograms rise above the zero line, it indicates a bullish momentum. When it is below the zero line and falling, it suggests a bearish momentum.

(vi) **Computation of the Relative Strength Index:** The *relative strength index* (RSI) is a momentum oscillator that measures the speed and change of price movements. The RSI is often used to help traders identify potential trend reversals and generate buy or sell signals. The computation of RSI involves the following steps.

- (a) *Choosing the period:* The most used period length is 14, representing 14 trading days or periods. However, traders can adjust this period based on their preferences and the timeframe they are analyzing.
- (b) *The computation of the average gain and average loss:* The average gain and average loss are computed over the selected period using (18.9) and (18.10). The gain or loss for each period is determined by comparing the current closing price with the previous closing price.

$$\text{Average Gain} = \frac{\text{Sum of Gains over } N \text{ periods}}{N} \quad (18.9)$$

$$\text{Average Loss} = \frac{\text{Sum of Losses over } N \text{ periods}}{N} \quad (18.10)$$

- (c) *The computation of the relative strength (RS):* The *relative strength* (RS) is computed by dividing the average gain by the average loss as in (18.11).

$$RS = \frac{\text{Average Gain}}{\text{Average Loss}} \quad (18.11)$$

- (d) *The computation of the RSI:* Finally, the RSI value is computed as in (18.12). The value of RSI lies between 0 and 100.

$$RSI = 100 - \left(\frac{100}{1 + RS} \right) \quad (18.12)$$

If the RSI is above 70, it is often considered overbought, indicating that the stock may be overvalued, and a trend reversal or corrective pull-back may be imminent. On the other hand, if the RSI is below 30, it is considered oversold, indicating that the asset may be undervalued, and a trend reversal or corrective upward movement may be on the horizon. The divergence between the RSI and the price movement can also provide signals. For example, if the price is making new highs, but the RSI is not confirming them, it may indicate a weakening momentum. Some traders use the level of 50 as a threshold. An RSI above 50 is considered bullish, while an RSI below 50 is considered bearish.

- (vii) **Graphical representation of the technical indicators:** For each stock, the Bolinger Bands, MACD histograms, and RSI plots are constructed, and the buy and sell signal points are identified.

- (viii) **Returns computation:** This step involves the computation of returns for 12 months from July 1, 2022, to June 30, 2023, for each stock, using the technical indicators. A comparative analysis of the three indicators is made based on the computed annual returns.

18.4 Experimental Results

This section presents details of the stocks chosen from the 14 sectors for the study and the results of the performance of the three technical indicators on the stocks. The 14 sectors examined in this study are banking, auto, consumer durables, financial services except banks, IT, FMCG, metal, media, oil and gas, mid-small IT and telecom, private banks, pharma, realty, and PSU banks. The Bollinger Bands, MACD, and RSI values for the top ten stocks in the 14 sectors are calculated and plotted using Python 3.9.8 and its libraries numpy, pandas, matplotlib, and yfinance. These programs were run on the Google Colab platform [85].

Section 4.1 identifies the top ten stocks from the sectors based on their free-float market capitalization (FMCC). Subsequently, Sect. 4.2 presents the detailed results of the performance of the three technical indicators on the stocks. A critical analysis of the results is done in Sect. 4.3.

18.4.1 The Selection of Stocks for Analysis

This section mentions the top 10 stocks from each of the above-mentioned 14 sectors. These stocks have the large FMCC in their respective sectors.

Auto sector: According to the report released by the NSE on June 30, 2022, the ten stocks with the largest FFMC are the following: Maruti Suzuki India (MARUTI), Mahindra & Mahindra (M&M), Tube Investment of India (TIINDIA), Tata Motors (TATAMOTORS), Bajaj Auto (BAJAJ-AUTO), TVS Motor Company (TVSMOTOR), Eicher Motors (EICHERMOT), Ashok Leyland (ASHOKLEY), Hero MotoCorp (HEROMOTOCO), and Bharat Forge (BHARATFORG) [84]. The ticker symbols for these stocks, unique identifiers on the stock exchange, are provided in parentheses.

Banking sector: The top ten banking sector stocks by their FFMC are: HDFC Bank (HDFCBANK), ICICI Bank (ICICIBANK), Kotak Mahindra Bank (KOTAK-BANK), State Bank of India (SBIN), IndusInd Bank (INDUSINDBK), Axis Bank (AXISBANK), AU Small Finance Bank (AUBANK), Bank of Baroda (BANKBARODA), IDFC First Bank (IDFCFIRSTB), and Federal Bank (FEDERALBNK) [84].

Financial Services Ex-Banks sector: The ten stocks with the highest FFMC are as follows: HDFC (HDFC), Bajaj Finance (BAJFINANCE), Bajaj Finserv (BAJAJFINSV), SBI Life Insurance Company (SBILIFE), HDFC Life Insurance Company (HDFCLIFE), Shriram Finance (SHRIRAMFIN), Cholamandalam Investment and Finance (CHOLAFIN), ICICI Lombard General Insurance Company (ICICIGI), (ix) Bajaj Holdings and Investment (BAJAJHLDNG), and Power Finance Corporation (PFC) [84].

Consumer Durables sector: The top ten stocks in this sector based on their FFMC are the following: Titan Company (TITAN), Crompton Greaves Consumer Electricals (CROMPTON), Havells India (HAVELLS), Dixon Technologies (DIXON), Voltas (VOLTAS), Bata India (BATAINDIA), Rajesh Exports (RAJESHEXPO), Kajaria Ceramics (KAJARIACER), Blue Star (BLUESTARCO), and Relaxo Footwears (RELAXO) [84].

FMCG sector: The ten stocks with the highest FFMC in this sector are: ITC (ITC), Nestle India (NESTLEIND), Hindustan Unilever (HINDUNILVR), Tata Consumer Products (TATACONSUM), Britannia Industries (BRITANNIA), Godrej Consumer Products (GODREJCP), Dabur India (DABUR), Varun Beverages (VBL), Marico (MARICO), and United Spirits (MCDOWELL-N) [84].

Information Technology (IT) sector: The top ten stocks in the IT sector based on their FFMC are as follows: Tata Consultancy Services (TCS), Infosys (INFY), Wipro (WIPRO), Tech Mahindra (TECHM), HCL Technologies (HCLTECH), LTIMindtree (LTIM), Persistent Systems (PERSISTENT), Coforge (COFORGE), Mphasis (MPHASIS), and L&T Technology Services (LTTS) [84].

Media sector: The top ten stocks with the highest FFMC in this sector are the following: PVR (PVRINOX), Zee Entertainment Enterprises (ZEEL), Sun TV Network (SUNTV), Nazara Technologies (NAZARA), TV18 Broadcast (TV18BRDCST), Dish TV India (DISHTV), Navneet Education (NAVNETEDUL), Network18 Media & Investments (NETWORK18), NDTV (NDTV), and Hathway Cable & Datacom (HATHWAY) [84]. However, NAZARA and PVRINOX were not included due to their listing on NSE after the portfolio formation began on July 1, 2019. Consequently, TV Today (TVTODAY) and Saregama India (SAREGAMA) replaced PVRINOX and NAZARA, respectively, as their market capitalization was higher.

Metal sector: The top ten stocks in this sector, ranked by their FFMC, are the following: Tata Steel (TATASTEEL), JSW Steel (JSWSTEEL), Adani Enterprises (ADANIENT), Hindalco Industries (HINDALCO), APL Apollo Tubes (APLAPOLLO), Vedanta (VEDL), Jindal Stainless (JSL), Jindal Steel and Power (JINDALSTEL), NMDC (NMDC), and Steel Authority of India (SAIL) [84].

Mid-Small IT and Telecom sector: This category encompasses mid and small-cap stocks in the IT & telecom sector. The top ten stocks with the highest FFMC are: Persistent Systems (PERSISTENT), Tata Elxsi (TATAELXSI), Coforge (COFORGE), Tata Communications (TATACOMM), KPIT Technologies

(KPITTECH), Mphasis (MPHASIS), L&T Technology Services (LTTS), Cyient (CYIENT), Oracle Financial Services Software (OFSS), and Sonata Software (SONATSOFTW) [84].

Oil and Gas sector: The top ten stocks in the oil and gas sector, based on their FFMC, are Reliance Industries (RELIANCE), Bharat Petroleum Corporation (BPCL), Oil and Natural Gas Corporation (ONGC), GAIL India (GAIL), Hindustan Petroleum Corporation (HINDPETRO), Indian Oil Corporation (IOC), Adani Total Gas (ATGL), Indraprastha Gas (IGL), Petronet LNG (PETRONET), Oil India (OIL) [84].

Pharma sector: The top ten stocks in this sector, based on their FFMC are the following: Dr. Reddy's Labs (DRREDDY), Sun Pharmaceuticals Industries (SUNPHARMA), Divi's Laboratories (DIVISLAB), Cipla (CIPLA), Aurobindo Pharma (AUROPHARMA), Lupin (LUPIN), Alkem Laboratories (ALKEM), Zydus Lifesciences (ZYDUSLIFE), Torrent Pharmaceuticals (TORNTPHARM), and Laurus Labs (LAURUSLABS) [84].

Private Banks sector: The top ten stocks in this sector based on their FFMC are: ICICI Bank (ICICIBANK), HDFC Bank (HDFCBANK), IndusInd Bank (INDUSINDBK), Kotak Mahindra Bank (KOTAKBANK), Axis Bank (AXISBANK), IDFC First Bank (IDFCFIRSTB), Federal Bank (FEDERALBNK), Bandhan Bank (BANDHANBNK), City Union Bank (CUB), and RBL Bank (RBLBANK), and [84].

PSU Banks sector: The top ten stocks with the largest FFMC in this sector are: State Bank of India (SBIN), Punjab National Bank (PNB), Bank of Baroda (BANKBARODA), Canara Bank (CANBK), Bank of India (BANKINDIA), Indian Bank (INDIANB), Union Bank of India (UNIONBANK), Indian Overseas Bank (IOB), Bank of Maharashtra (MAHABANK), and Central Bank of India (CENTRALBK) [84].

Realty sector: The top ten stocks in the realty sector based on their FFMC are the following: DLF (DLF), Godrej Properties (GODREJPROP), Macrotech Developers (LODHA), Oberoi Realty (OBEROIRLTY), Phoenix Mills (PHOENIXLTD), Brigade Enterprises (BRIGADE), Prestige Estate Projects (PRESTIGE), Indiabulls Real Estate (IBREALEST), Mahindra Lifespace Developers (MAHLIFE), and Sobha (SOBHA) [84]. The stock of LODHA was replaced by Sunteck Realty (SUNTECK) as the former was listed on the NSE only from 2021 in the month of April.

18.4.2 Performance Results

The results of the performance of the three technical indicators on the top 10 stocks of the chosen 14 sectors are presented in this section. The results are presented sector-wise. For each sector, as an illustration, the plots for the Bollinger Bands, MACD,

and RSI are shown for one of the stocks from the sector. However, the annual returns for all stocks are presented for three technical indicators for comparative analysis.

Auto Sector: Figures 18.1, 18.2, and 18.3 present the Bollinger Bands, MACD, and RSI plots, respectively, from July 1, 2022, to June 30, 2023, of Mahindra and Mahindra, the stock with the highest FFMC of the *auto* sector. Table 18.2 exhibits the annual returns of the three strategies for the 10 stocks of this sector.

Banking Sector: Figures 18.4, 18.5, and 18.6 present the Bollinger Bands, MACD, and RSI plots, respectively, of ICICI Bank, of the banking sector from July 1, 2022 to June 30, 2023. Table 18.3 exhibits the annual returns of the three strategies for the 10 sector stocks. The highest return for a given stock is shown in a bold font.

Financial Services Except Banks: Figures 18.7, 18.8, and 18.9 present the Bollinger Bands, MACD, and RSI plots, respectively, of Bajaj Finance from July 1, 2022, to June 30, 2023. Table 18.4 exhibits the annual returns of the three strategies for the 10 stocks of this sector. The highest return for a given stock is shown in a bold font.

Consumer Durable Sector: Figures 18.10, 18.11 and 18.12 present the Bollinger Bands, MACD, and RSI plots, respectively, of Titan Company from July 1, 2022, to June 30, 2023. Table 18.5 exhibits the annual returns of the three strategies for the 10 stocks of this sector. The highest return for a stock is shown in bold font.

FMCG Sector: Figures 18.13, 18.14 and 18.15 present the Bollinger Bands, MACD, and RSI plots, respectively, of Hindustan Unilever from July 1, 2022, to June 30, 2023. Table 18.6 exhibits the annual returns of the three strategies for the 10 stocks of the FMCG sector.

IT Sector: Figures 18.16, 18.17 and 18.18 present the Bollinger Bands, MACD, and RSI plots, respectively, of Coforge from July 1, 2022, to June 30, 2023. Table 18.7 exhibits the annual returns of the three strategies for the 10 stocks of the IT sector.

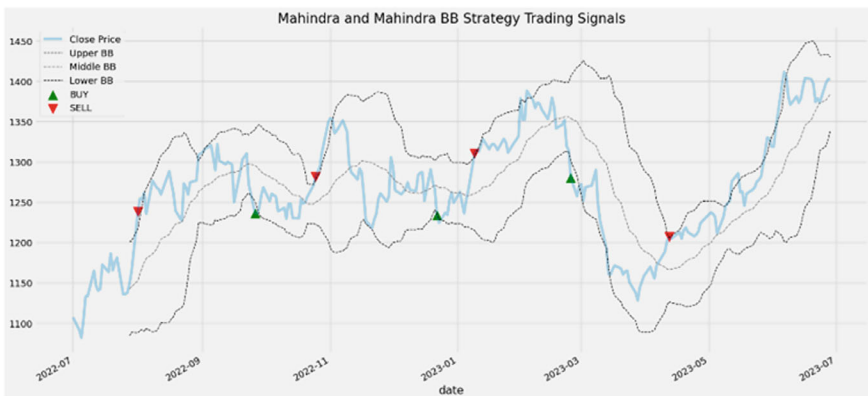


Fig. 18.1 The Bollinger Bands plot of Mahindra and Mahindra stock with the trading signal points identified

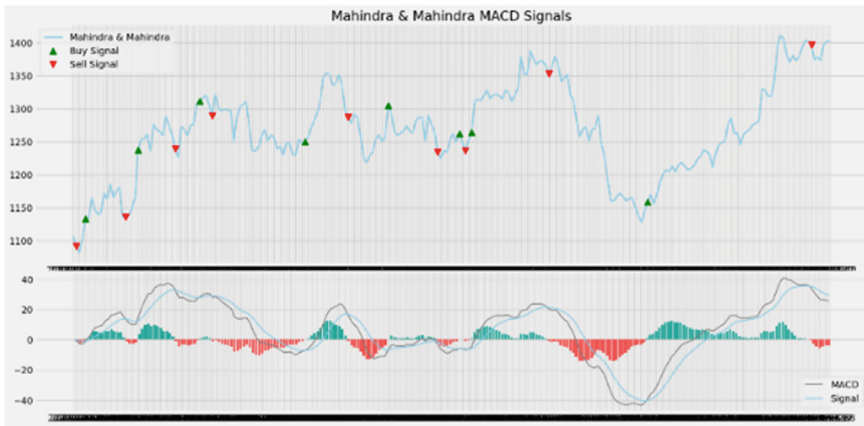


Fig. 18.2 The MACD plot of Mahindra and Mahindra stock with the trading signal points identified

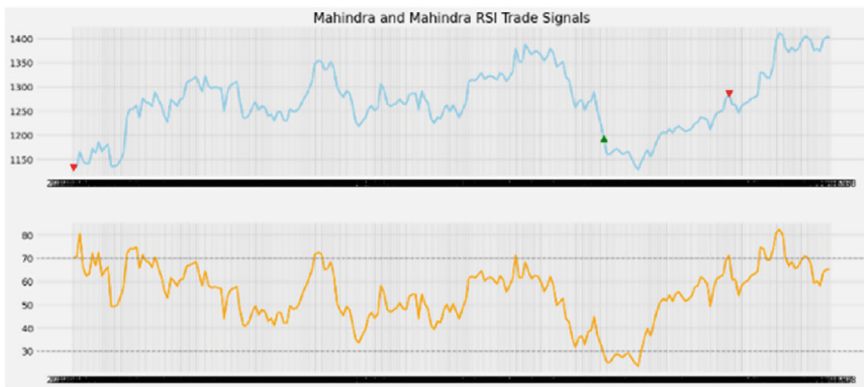


Fig. 18.3 The RSI plot of Mahindra and Mahindra stock with the trading signal points identified

Table 18.2 The Auto sector’s annual return (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
M&M	12.71	20.74	6.55
MARUTI	10.40	9.58	7.42
TATAMOTORS	21.09	29.60	9.50
BAJAJ-AUTO	13.43	30.30	5.19
EICHERMOT	10.54	19.04	13.53
HEEROMOTOCO	1.88	5.12	17.03
TIINIDA	35.13	41.47	0.00
TVSMOTOR	13.83	28.70	8.54
ASHOKLEY	12.25	-4.20	13.41
BHARATFORG	29.75	-5.04	6.65

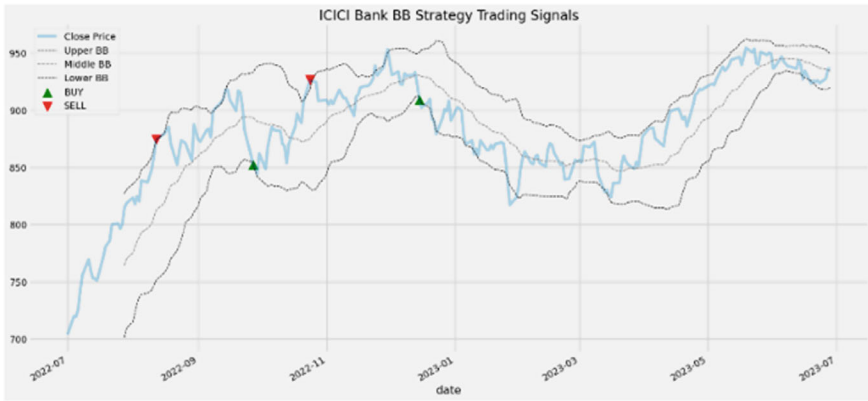


Fig. 18.4 The Bollinger Bands plot of ICICI Bank stock with the trading signal points identified

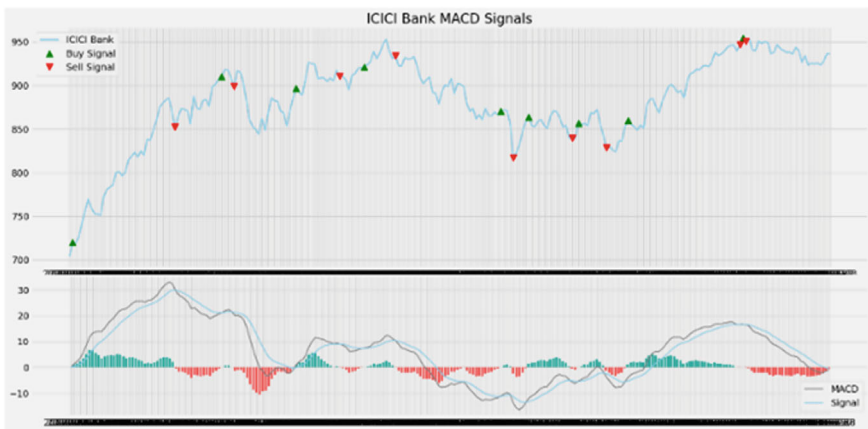


Fig. 18.5 The MACD plot of ICICI Bank stock with the trading signal points identified

Media Sector: Figures 18.19, 18.20, and 18.21 present the Bollinger Bands, MACD, and RSI plots, respectively, of Zee Entertainment Enterprises from July 1, 2022, to June 30, 2023. Table 18.8 exhibits the annual returns of the three strategies for the 10 media sector stocks. The highest return for a given stock is shown in a bold font.

Metal Sector: Figures 18.22, 18.23 and 18.24 present the Bollinger Bands, MACD, and RSI plots, respectively, of Hindalco Industries from July 1, 2022, to June 30, 2022. Table 18.9 exhibits the annual returns of the three strategies for the 10 stocks of this sector.

Mid-Small IT and Telecom Sector: Figures 18.25, 18.26 and 18.27 present the Bollinger Bands, MACD, and RSI plots, respectively, of L&T Technology Services

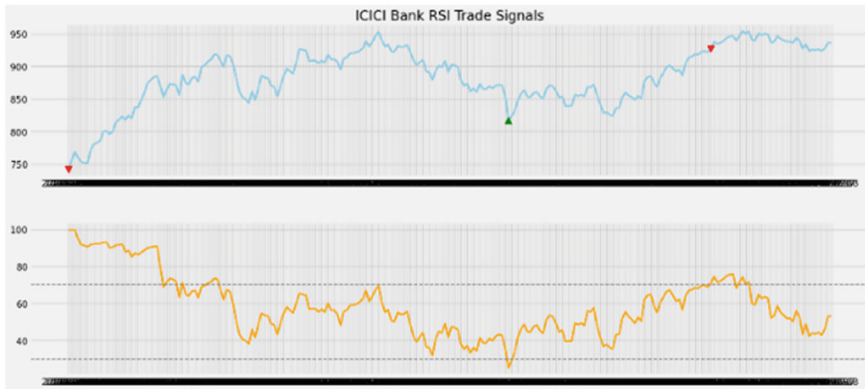


Fig. 18.6 The RSI plot of ICICI Bank stock with the trading signal points identified

Table 18.3 The Banking sector’s annual return (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
HDFCBANK	23.11	12.82	0.00
ICICIBANK	28.78	19.95	11.63
SBIN	6.46	22.04	6.75
AXISBANK	30.81	32.29	6.81
KOTAKBANK	11.22	3.16	3.22
INDUSINDBK	24.29	11.73	11.24
AUBANK	13.39	25.08	20.18
BANKBARODA	18.96	29.12	14.04
FEDERALBNK	42.35	2.29	0.00
IDFCFIRSTB	22.48	38.13	0.00

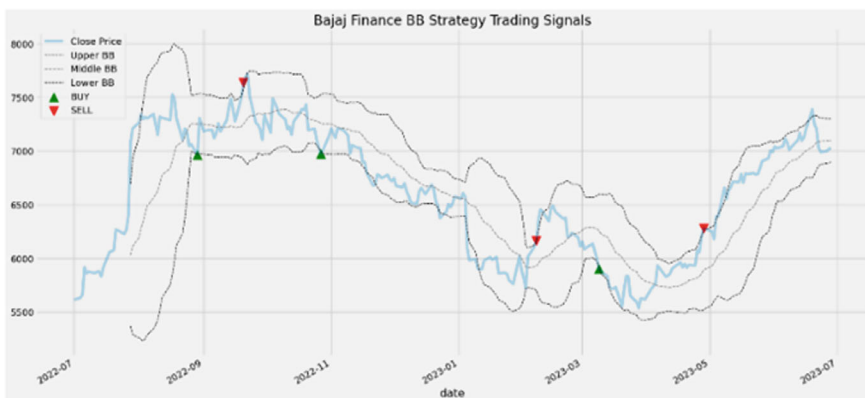


Fig. 18.7 The Bollinger Bands plot of Bajaj Finance stock with the trading signal points identified

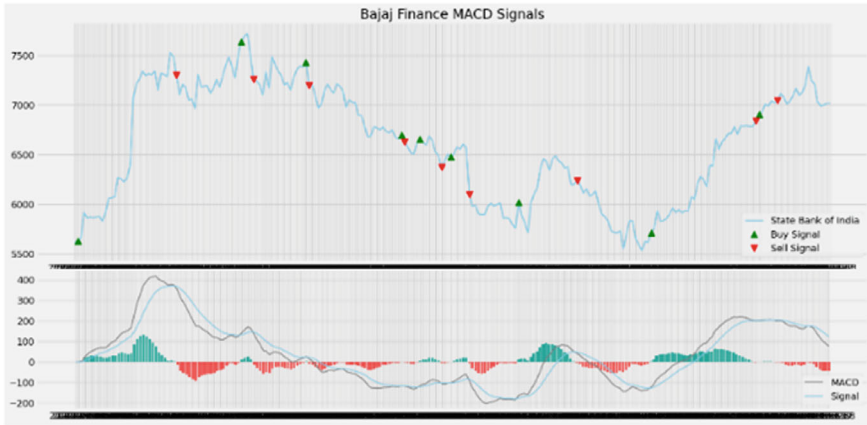


Fig. 18.8 The MACD plot of Bajaj Finance stock with the trading signal points identified

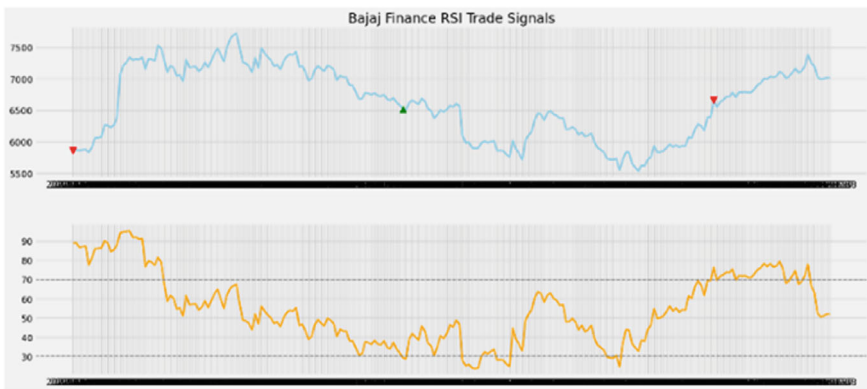


Fig. 18.9 The RSI plot of Bajaj Finance stock with the trading signal points identified

Table 18.4 The Financial Services Ex-Banks sector annual return (in percentage) of the BB, MACD, and RSI methods

Stock	BB	MACD	RSI
HDFC	23.11	20.54	0.00
BAJFINANCE	22.25	30.96	1.97
BAJAJFINSV	27.70	28.40	-4.47
HDFCLIFE	4.44	19.18	5.02
SBILIFE	14.55	11.59	8.37
SHRIRAMFIN	3.08	20.49	18.47
CHOLAFIN	21.44	57.03	0.00
ICICIGI	8.59	4.39	5.37
BAJAJHLDNG	12.24	44.19	8.77
PFC	19.72	56.69	4.21

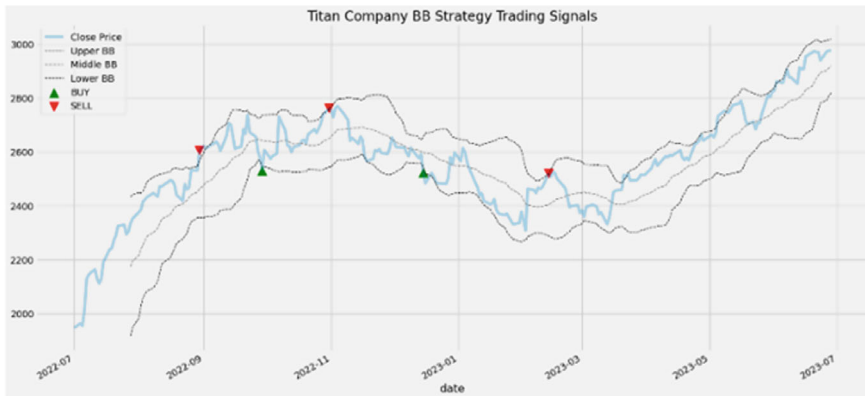


Fig. 18.10 The Bollinger Bands plot of Titan Company stock with the trading signal points identified

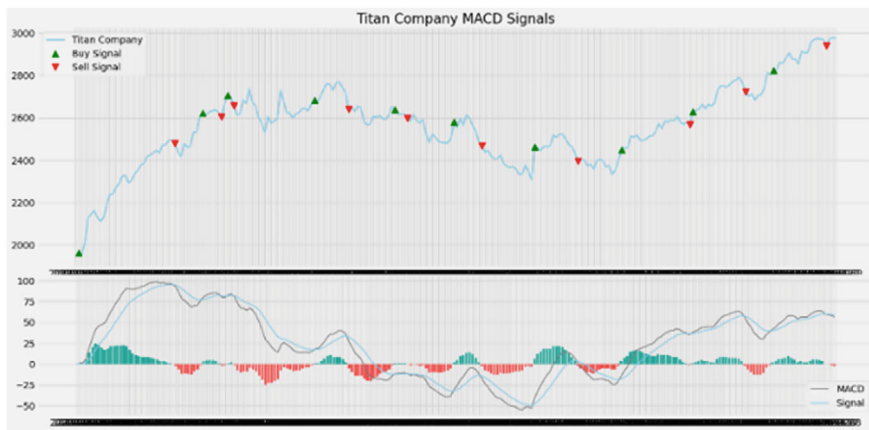


Fig. 18.11 The MACD plot of Titan Company stock with the trading signal points identified

from July 1, 2022, to June 30, 2023. Table 18.10 exhibits the annual returns of the three strategies for the 10 stocks of this sector.

Oil and Gas Sector: Figures 18.28, 18.29 and 18.30 present the Bollinger Bands, MACD, and RSI plots, respectively, of Hindustan Petroleum Corporation from July 1, 2022, to June 30, 2023. Table 18.11 exhibits the annual of the three strategies for the 10 stocks of this sector.

Pharma Sector: Figures 18.31, 18.32 and 18.33 present the Bollinger Bands, MACD, and RSI plots, respectively, of Sun Pharmaceuticals Industries from July 1, 2022, to June 30, 2023. Table 18.12 exhibits the annual returns of the three strategies for the 10 stocks of this sector. The highest return for a given stock is shown in a bold font.

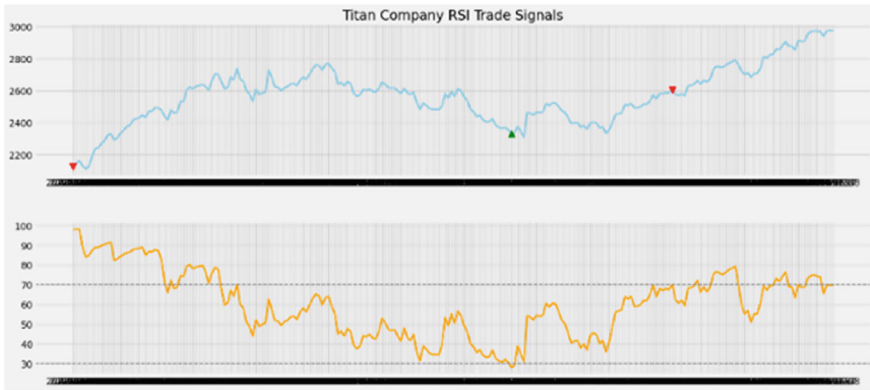


Fig. 18.12 The RSI plot of Titan Company stock with the trading signal points identified

Table 18.5 The Consumer Durables sector annual returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
TITAN	29.20	26.80	9.05
HAVELLS	16.09	4.05	9.23
CROMPTON	18.71	- 29.38	- 21.86
VOLTAS	9.72	- 10.56	3.14
DIXON	- 16.30	54.34	- 16.66
KAJARIACER	26.62	20.11	0.00
BATAINDIA	- 5.57	5.99	0.59
BLUESTARCO	20.06	50.69	5.36
RAJESHEXPO	27.70	- 10.98	- 9.73
RELAXO	- 7.95	- 8.56	- 9.51

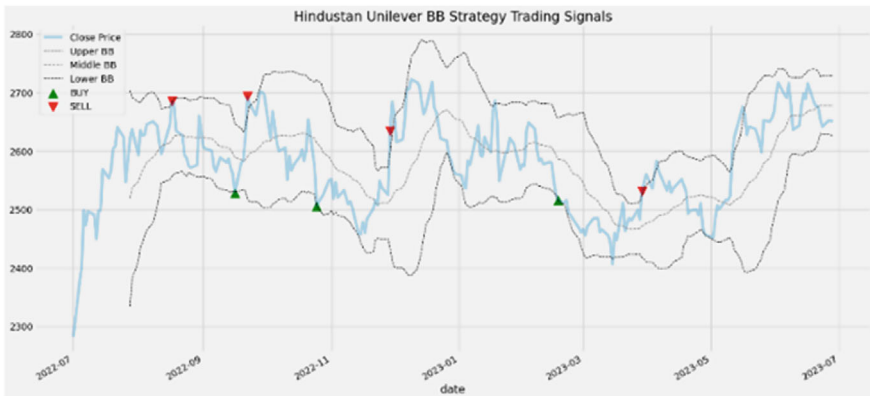


Fig. 18.13 The Bollinger Bands plot of Hindustan Unilever stock with the trading signal points identified

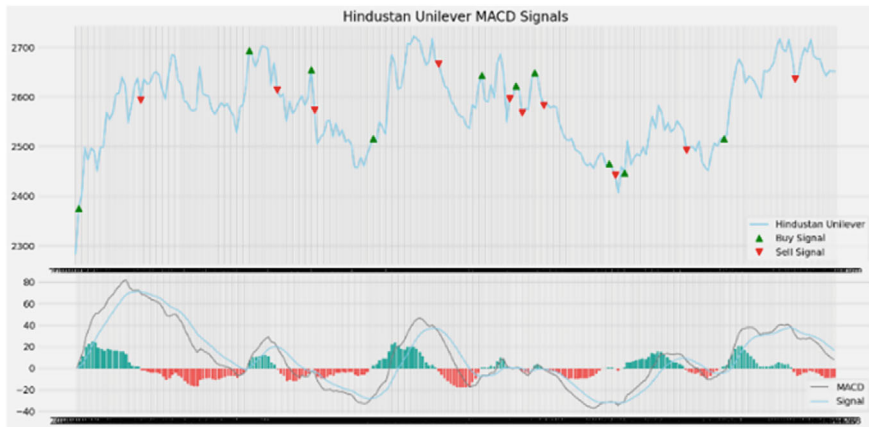


Fig. 18.14 The MACD plot of Hindustan Unilever stock with the trading signal points identified

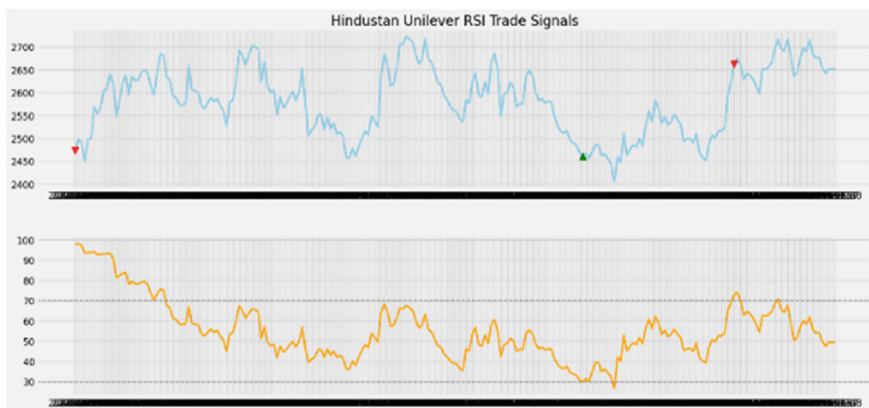


Fig. 18.15 The RSI plot of Hindustan Unilever stock with the trading signal points identified

Table 18.6 The FMCG sector annual returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
ITC	8.77	34.06	0.00
HINDUNILVR	26.31	11.89	7.42
NESTLEIND	12.79	20.42	7.31
BRITANNIA	18.61	17.40	0.00
TATACONSUM	12.95	7.62	2.97
GODREJCP	23.50	17.71	0.00
VBL	17.36	38.82	19.89
DABUR	11.76	7.44	0.00
MCDOWELL-N	2.12	13.44	-3.61
MARICO	12.04	-4.68	0.00

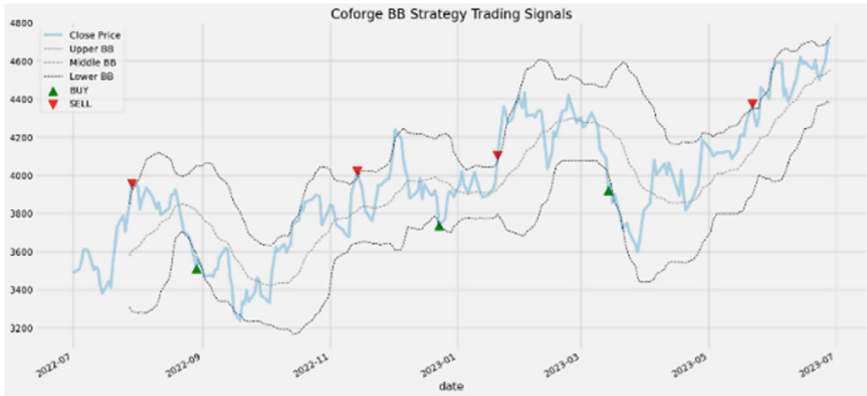


Fig. 18.16 The Bollinger Bands plot of Coforge stock with the trading signal points identified

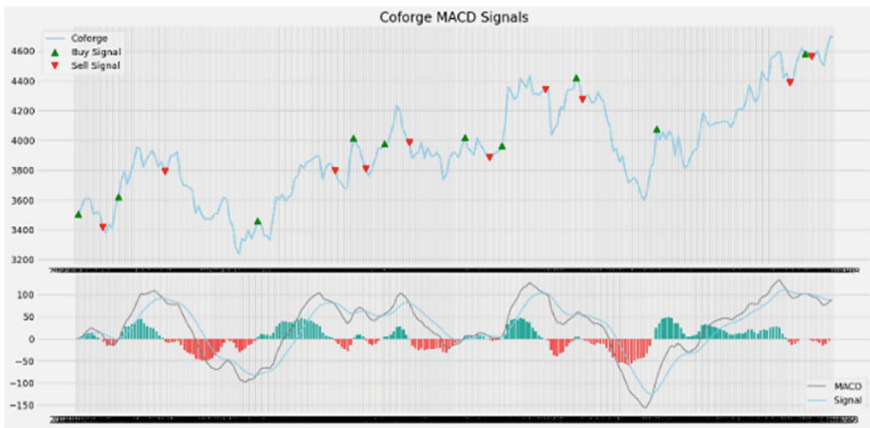


Fig. 18.17 The MACD plot of Coforge stock with the trading signal points identified

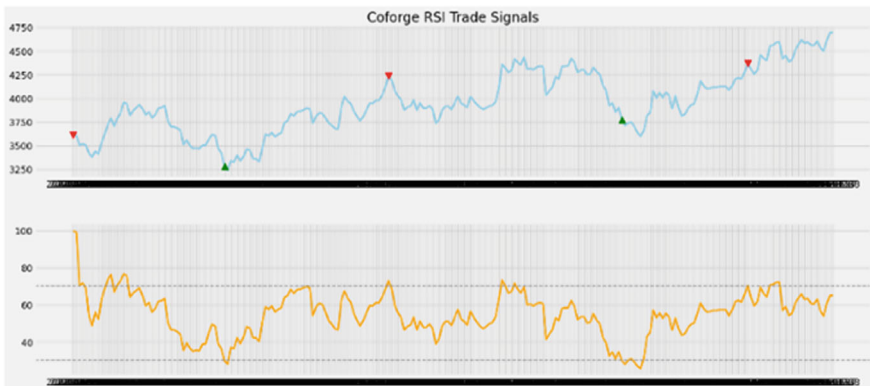


Fig. 18.18 The RSI plot of Coforge stock with the trading signal points identified

Table 18.7 The IT sector annual returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
INFY	0.16	- 5.40	- 10.99
TCS	- 0.54	4.28	3.77
WIPRO	1.03	- 1.07	6.30
TECHM	14.98	13.91	11.97
HCLTECH	4.22	14.48	0.00
LTIM	34.93	13.83	26.79
PERSISTENT	7.08	39.30	0.00
COFORGE	37.57	17.03	32.60
MPHASIS	- 11.16	1.35	3.70
LTTS	41.61	19.57	10.01

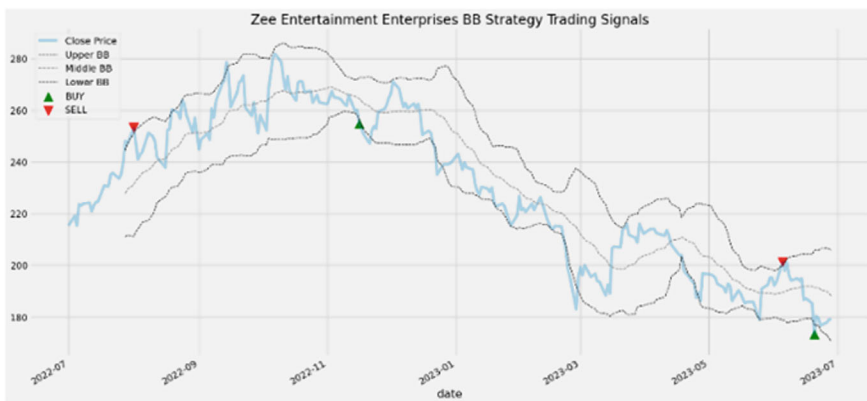


Fig. 18.19 The Bollinger Bands plot of Zee Entertainment Enterprises stock with the trading signal points identified

Private Banks Sector: Figures 18.34, 18.35 and 18.36 present the Bollinger Bands, MACD, and RSI plots, respectively, of Axis Bank, one of the ten stocks of the *private banks* sector. Table 18.13 exhibits the annual returns of the three strategies for the 10 stocks of the *private banks* sector.

PSU Banks Sector: Figures 18.37, 18.38 and 18.39 present the Bollinger Bands, MACD, and RSI plots, respectively, of the State Bank of India from July 1, 2022, to June 30, 2023. Table 18.14 exhibits the annual returns of the three strategies for the 10 stocks of this sector.

Realty Sector: Figures 18.40, 18.41 and 18.42 present the Bollinger Bands, MACD, and RSI plots, respectively, of DLF from July 1, 2022, to June 30, 2023. Table 18.15 exhibits the annual returns of the three strategies for the 10 stocks of this sector.



Fig. 18.20 The MACD plot of Zee Entertainment Enterprises stock with the trading signal points identified

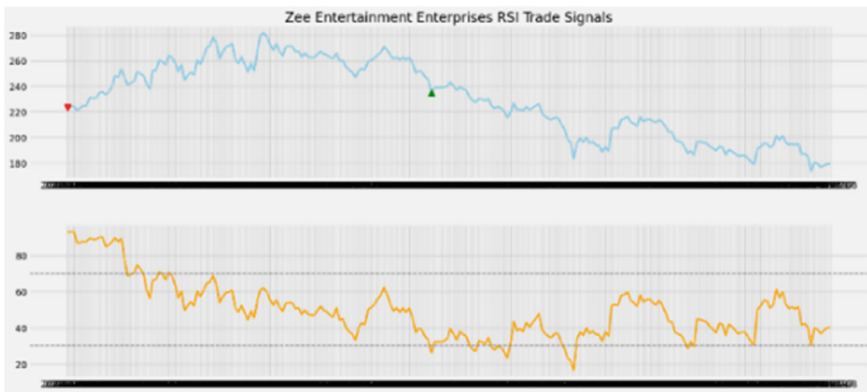


Fig. 18.21 The RSI plot of Zee Entertainment Enterprises stock with the trading signal points identified

Table 18.16 exhibits the summary of the results, in which, for each sector, the number of stocks that yielded the highest returns corresponding to each of the three strategies are listed. For example, for the *auto* sector, for the period from July 1, 2022, to June 30, 2023, two stocks yielded the highest returns using Bollinger bands, the MACD strategy yielded the highest return for six stocks, while for two stocks, the RSI strategy yielded the highest return. The column-wise totals represent the total number of stocks that yielded the highest returns corresponding to the three strategies for the said period. Bollinger Bands, MACD, and RSI strategies yielded the highest returns for 61, 62, and 17 stocks, respectively. It is evident that while the Bollinger Bands and MACD strategies performed equally well, RSI’s performance was poor. The Bollinger Bands strategy exhibited the best performance for the *private banks*

Table 18.8 The Media sector returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
ZEEL	- 5.55	- 13.25	- 31.28
PVRINOX	- 21.68	- 28.74	10.76
SUNTV	28.06	- 20.60	- 1.56
TV18BRDCST	- 2.81	- 8.29	12.12
NAZARA	- 1.18	- 7.65	21.00
DISHTV	- 18.15	70.31	- 4.29
NETWORK18	3.08	- 16.55	31.11
NAVNETEDUL	- 0.68	27.72	- 6.48
HATHWAY	4.04	- 0.31	11.11
NDTV	73.74	185.85	- 23.09

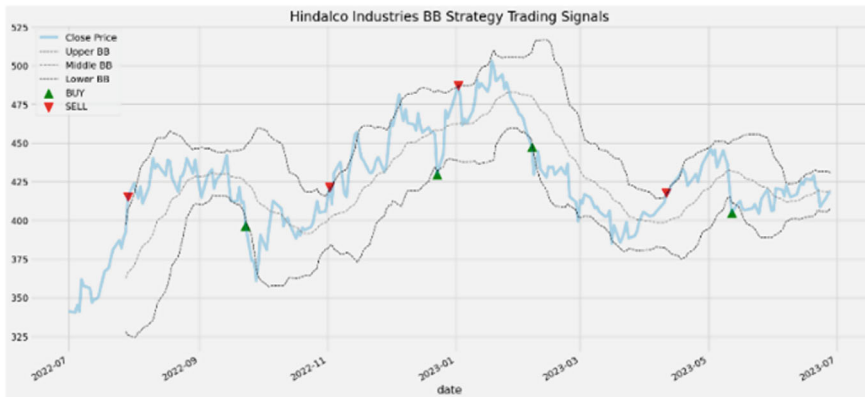


Fig. 18.22 The Bollinger Bands plot of Hindalco Industries stock with the trading signal points identified

and consumer durables sectors. For the PSU banks sector, the performance of the MACD strategy has been excellent. While the performance of the RSI strategy is found to be quite poor in general for the period of this study, for the media sector, this strategy worked reasonably well.

18.4.3 Critical Analysis of the Results

Bollinger Bands, MACD, and RSI are the three most popular technical indicators, each yielding the best returns under specific time series patterns of stock prices.

Bollinger Bands: As mentioned in Sect. 18.3, Bollinger Bands comprise a central band (a moving average) and two additional bands (each a standard deviation away

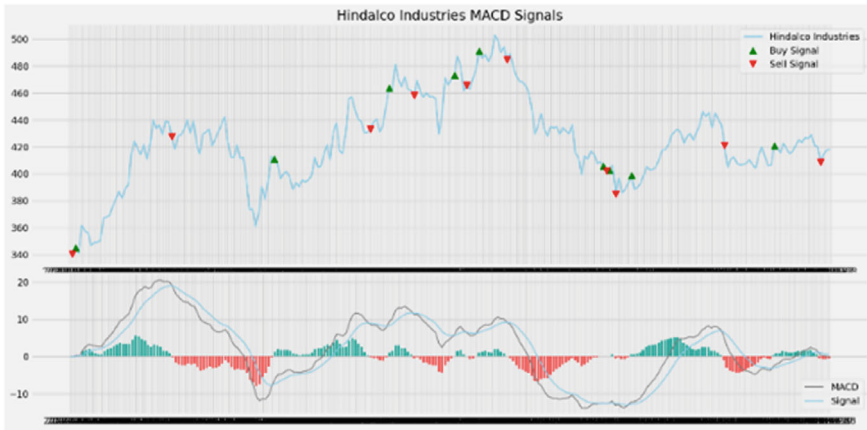


Fig. 18.23 The MACD plot of Hindalco Industries stock with the trading signal points identified

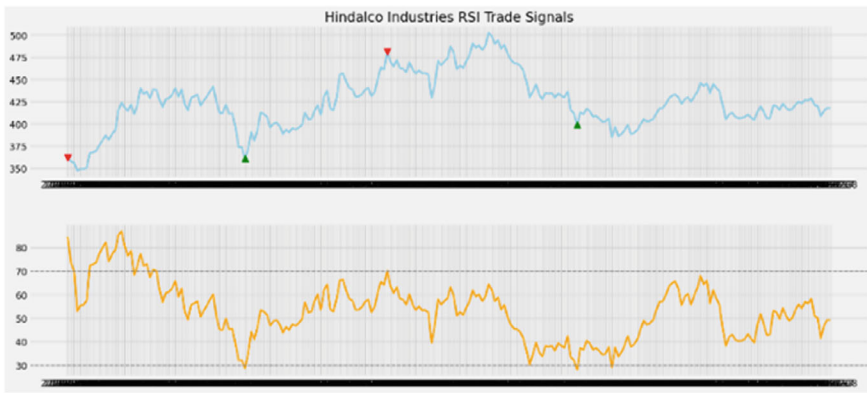


Fig. 18.24 The RSI plot of Hindalco Industries stock with the trading signal points identified

Table 18.9 The Metal sector annual returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
TATASTEEL	30.05	15.65	6.13
ADANIENIT	- 19.12	117.35	- 47.74
JSWSTEEL	15.64	8.24	34.56
HINDALCO	33.35	21.42	33.22
VEDL	22.96	2.66	- 2.78
APLAPOLLO	28.59	34.63	11.18
JINDALSTEL	19.40	8.24	12.02
SAIL	40.75	- 10.37	12.11
NMDC	1.40	- 13.76	6.21
JSL	28.09	72.75	0.00

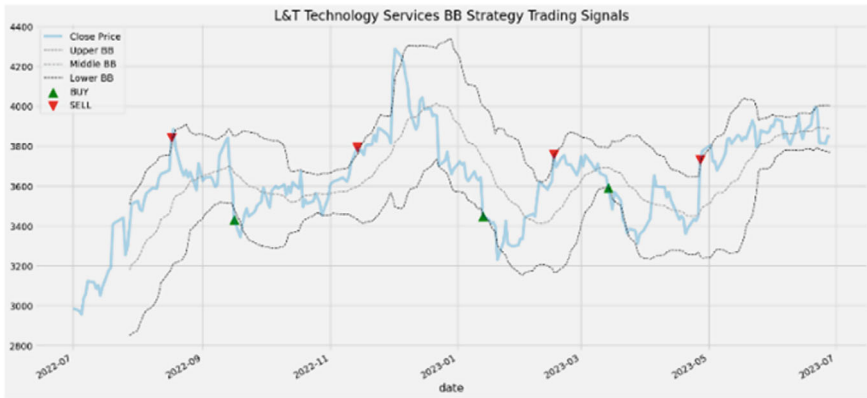


Fig. 18.25 The Bollinger Bands plot of L&T Technology Services stock with the trading signal points identified

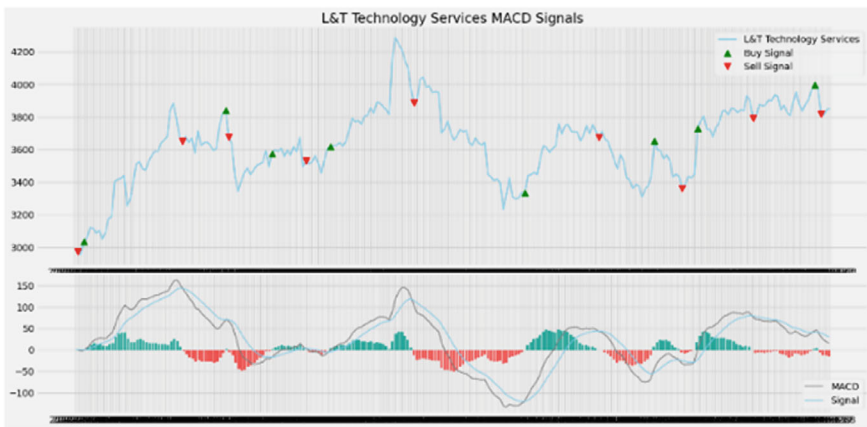


Fig. 18.26 The MACD plot of L&T Technology Services stock with the trading signal points identified

from the moving average). These bands help detect volatility and potential buy or sell signals by comparing the stock’s price to these bands. The following time series patterns are most appropriate for computations using Bollinger Bands.

- (i) *Mean reversion*: Stocks that show mean-reverting behavior, where prices oscillate around a central value (the moving average), tend to yield the highest returns with Bollinger Bands. In this pattern, prices frequently touch or cross the outer bands and revert to the middle band.

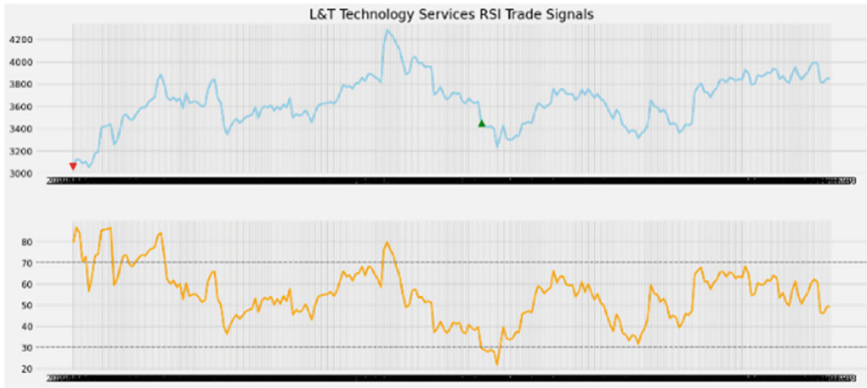


Fig. 18.27 The RSI plot of L&T Technology Services stock with the trading signal points identified

Table 18.10 The Mid-Small IT & Telecom sector returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
TATAELXSI	- 19.21	22.14	- 1.49
PERSISTENT	7.08	39.30	0.00
TATACOMM	39.34	34.96	0.00
COFORGE	37.57	17.03	32.60
MPHASIS	- 11.16	1.36	3.70
KPITTECH	23.87	9.44	0.00
CYIENT	10.71	26.38	7.12
LTTTS	41.61	19.57	10.01
SONATASOFTW	21.17	61.19	0.00
OFSS	4.41	6.97	0.00

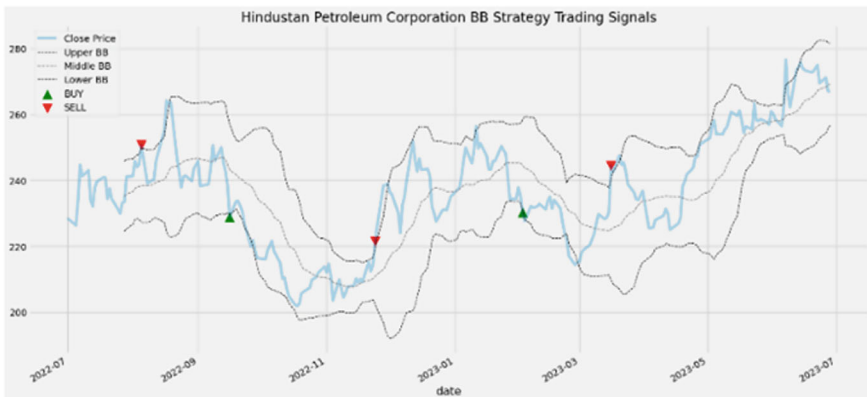


Fig. 18.28 The Bollinger Bands plot of Hindustan Petroleum Corporation stock with the trading signal points identified

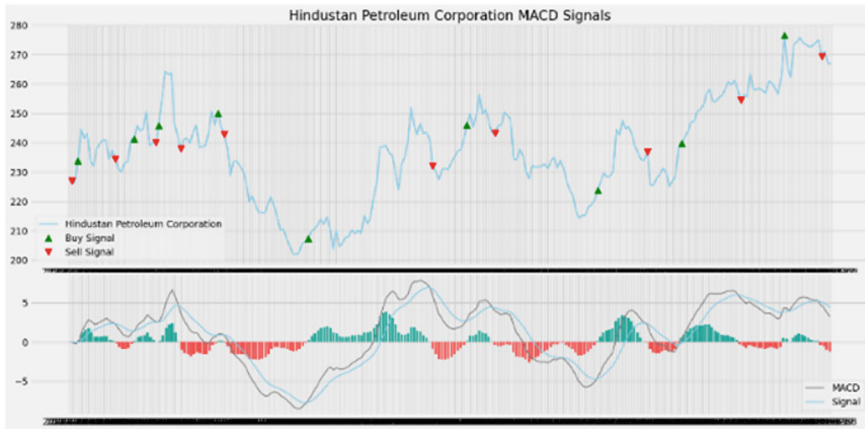


Fig. 18.29 The MACD plot of Hindustan Petroleum Corporation stock with the trading signal points identified

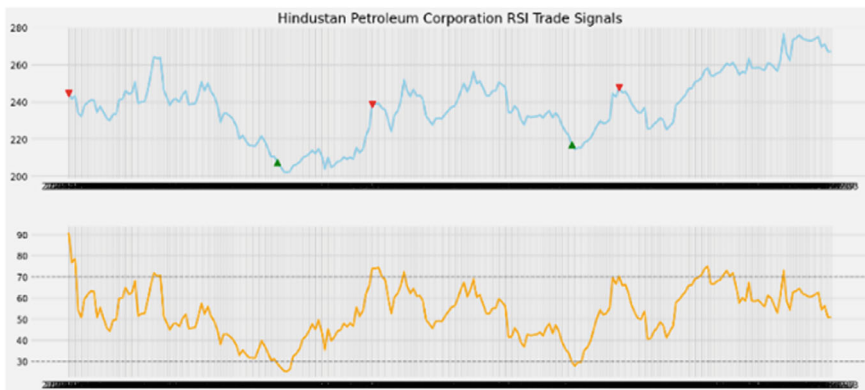


Fig. 18.30 The RSI plot of Hindustan Petroleum Corporation stock with the trading signal points identified

- (ii) *High volatility periods*: Stocks experiencing high volatility, where prices frequently move from one band to another, can provide good trading opportunities. Traders can purchase when the price reaches the lower band and sell when it hits the upper band.

Bollinger Bands yielded the highest returns for 61 stocks, which were typically highly volatile and mean-reverting in nature. More specifically, most of the consumer durables and private banks' stocks were volatile and mean-reverting. Therefore, Bollinger Bands generated the highest returns for most stocks in these two sectors.

MACD: The MACD is a momentum indicator that tracks trends by comparing two moving stock price averages. According to Sect. 18.3, the MACD line is created

Table 18.11 The oil and gas sector annual returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
RELIANCE	6.22	- 1.07	21.20
ONGC	23.42	- 2.98	10.73
BPCL	16.12	6.45	13.46
IOC	8.95	21.39	- 0.73
GAIL	7.97	- 13.09	0.00
ATGL	- 297	18.44	- 401.63
HINDPETRO	10.92	10.97	23.09
PETRONET	10.85	- 21.92	2.70
IGL	27.31	16.43	0.00
OIL	4.00	- 20.83	12.73

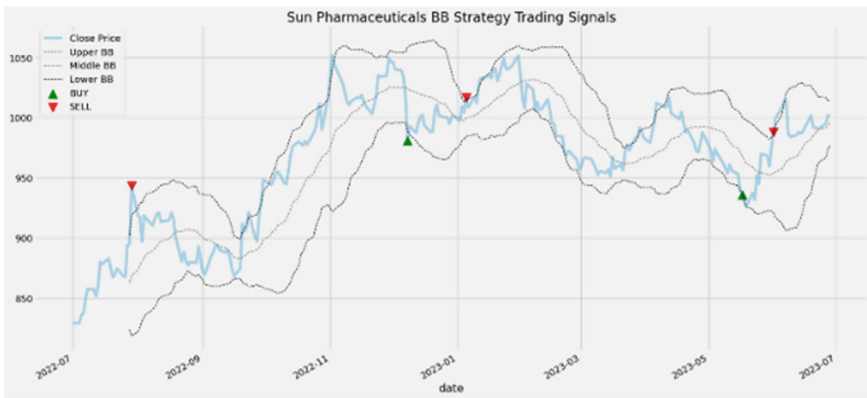


Fig. 18.31 The Bollinger Bands plot of Sun Pharmaceutical Industries stock with the trading signal points identified

by subtracting the 26-period EMA from the 12-period EMA. The signal line is the 9-period EMA of the MACD line. The time series of stock prices for which MACD usually yields the highest returns exhibit the following patterns.

- (i) *Trending markets*: MACD performs best in trending markets, where a clear upward or downward trend is characterized by clear upward or downward movements. In a bullish trend, marked by a strong upward trajectory, a buy signal occurs when the MACD line crosses above the Signal line. Conversely, a sell signal occurs when the MACD line crosses below the Signal line in a bearish trend, defined by a downward trajectory.
- (ii) *Long trends*: Long, sustained trends provide the best signals because MACD is designed to capture the trend’s momentum.

MACD yielded the highest returns for 62 stocks that exhibited long and strong trends in their time series. Financial services except banks, PSU banks, auto, and

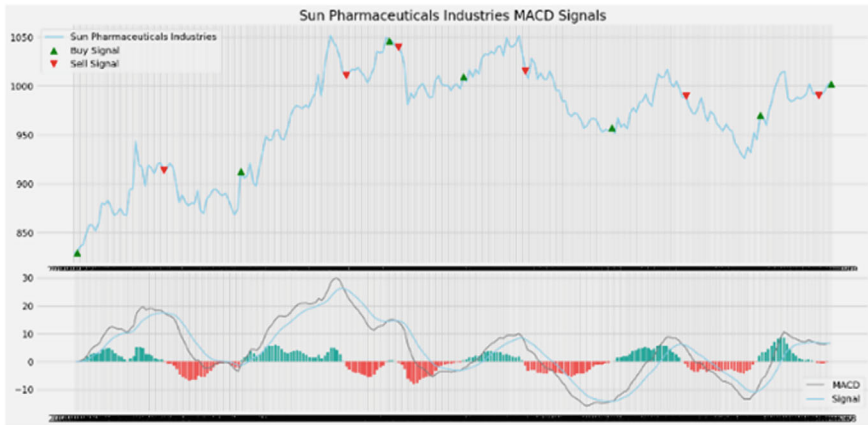


Fig. 18.32 The MACD plot of Sun Pharmaceutical Industries stock with the trading signal points identified

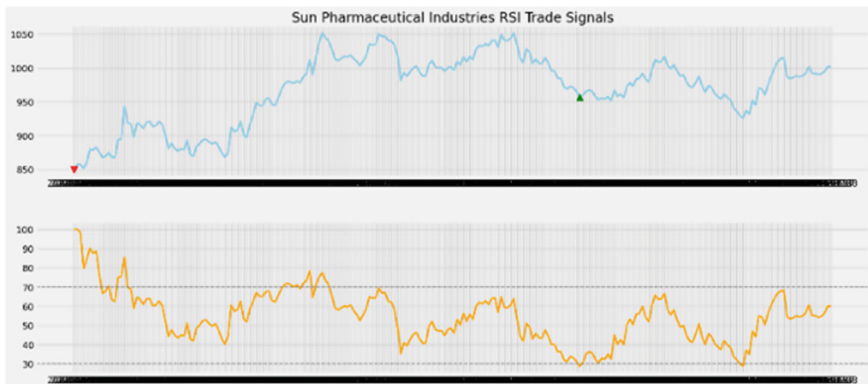


Fig. 18.33 The RSI plot of Sun Pharmaceutical Industries stock with the trading signal points identified

mid-small IT and telecom sectors exhibited stronger trends. Hence, the performance of MACD was the best for these sectors.

RSI: RSI functions as a momentum oscillator, assessing the pace and variation of price changes. As explained in Sect. 18.3, it ranges between 0 and 100 and is commonly utilized to detect overbought or oversold conditions.

Overbought or oversold patterns: RSI is most effective when stocks exhibit clear overbought conditions (above 70) or oversold conditions (below 30). Stocks frequently entering these extreme zones and returning to normal levels often present favorable trading opportunities. RSI divergence, where the price establishes a new high or low while RSI does not, can indicate potential reversals. Bullish divergence occurs when the stock price establishes a new low while the RSI forms a higher

Table 18.12 The Pharma sector annual returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
SUNPHARMA	19.91	28.16	4.47
DRREDDY	5.86	1.64	20.10
CIPLA	- 9.09	- 2.21	11.88
DIVISLAB	38.08	- 11.48	- 7.68
LUPIN	14.24	2.15	9.31
AUROPHARMA	16.55	2.36	6.64
ALKEM	15.68	12.26	0.00
TORNTPHARM	8.99	22.71	10.19
ZYDUSLIFE	12.30	36.46	5.43
LAURUSLABS	- 39.40	- 26.80	- 38.27



Fig. 18.34 The Bollinger Bands plot of Axis Bank stock with the trading signal points identified

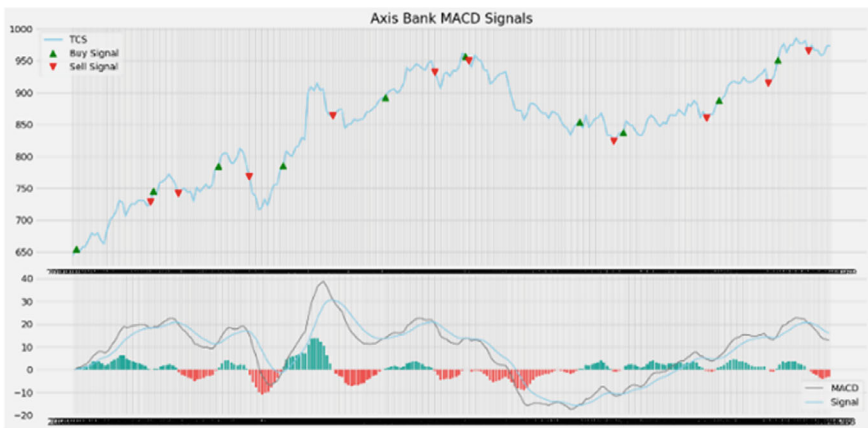


Fig. 18.35 The MACD plot of Axis Bank stock with the trading signal points identified

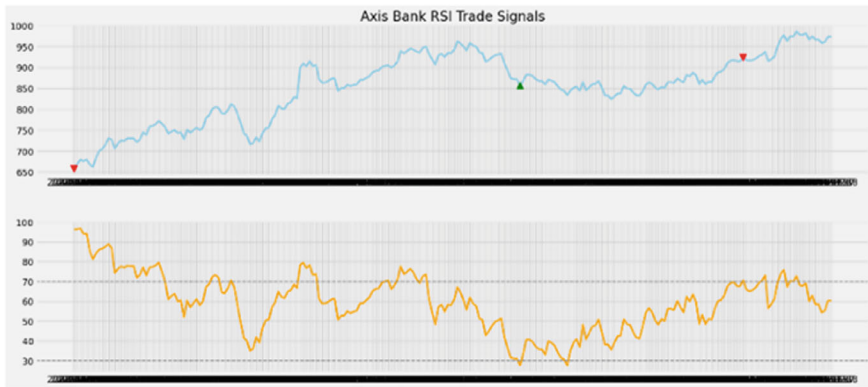


Fig. 18.36 The RSI plot of Axis Bank stock with the trading signal points identified

Table 18.13 The Private Banks sector annual returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
ICICIBANK	28.78	19.95	11.63
HDFCBANK	23.11	12.82	0.00
INDUSINDBK	24.29	11.73	11.24
KOTAKBANK	11.22	3.16	3.22
AXISBANK	30.81	32.29	6.81
FEDERALBNK	42.35	2.29	0.00
IDFCFIRSTB	22.48	38.13	0.00
BANDHANBNK	8.21	0.32	0.25
RBLBANK	39.58	56.09	18.43
CUB	- 3.00	- 9.45	- 28.64

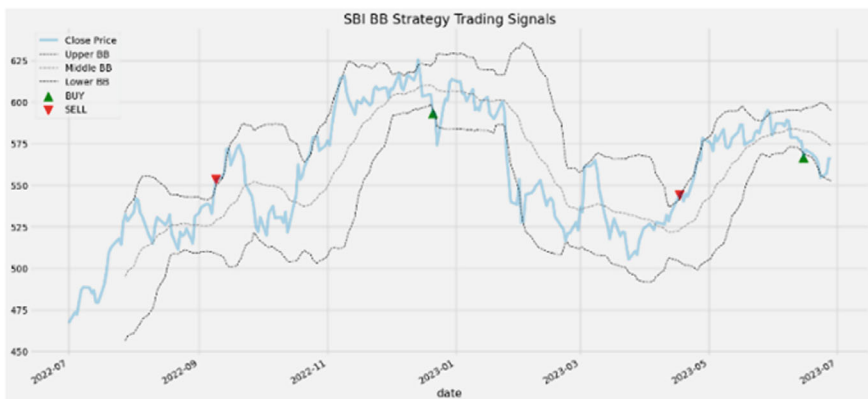


Fig. 18.37 The Bollinger Bands plot of State Bank of India stock with the trading signal points identified

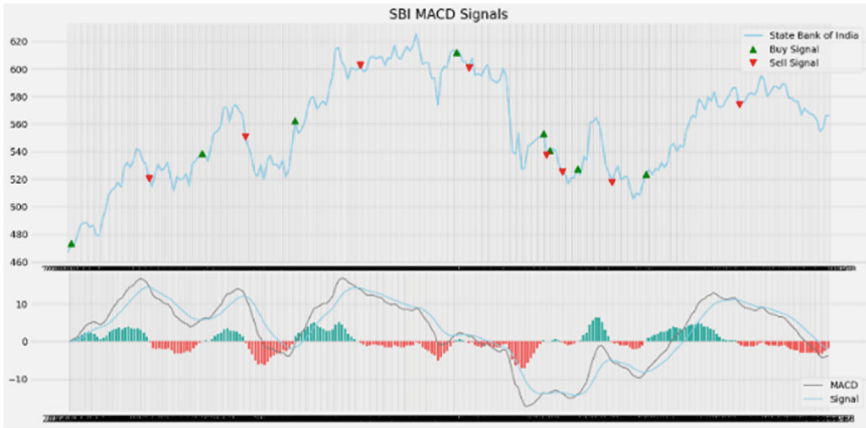


Fig. 18.38 The MACD plot of State Bank of India stock with the trading signal points identified

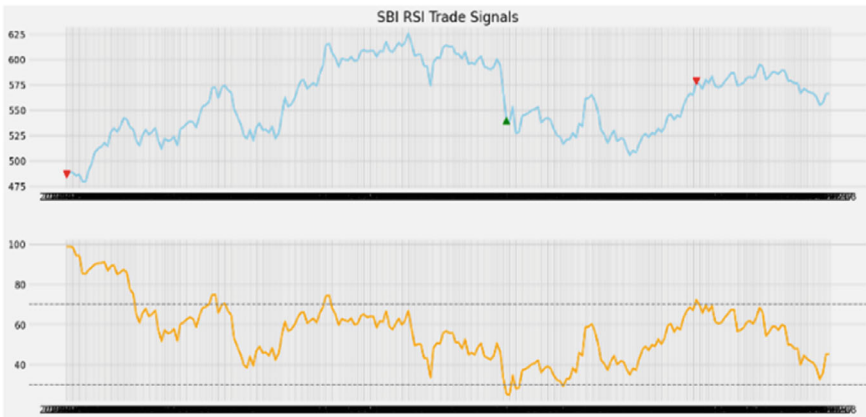


Fig. 18.39 The RSI plot of State Bank of India stock with the trading signal points identified

Table 18.14 The PSU Banks sector annual returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
SBIN	6.46	22.04	6.75
BANKBARODA	18.96	29.13	14.04
PNB	13.86	71.13	0.00
CANBK	30.23	23.60	0.00
UNIONBANK	17.17	101.29	4.04
INDIANB	24.74	101.29	23.14
BANKINDIA	11.68	41.33	2.00
MAHABANK	10.20	95.83	18.03
IOB	- 7.48	27.79	- 0.83
CENTRALBK	1.05	72.91	12.74

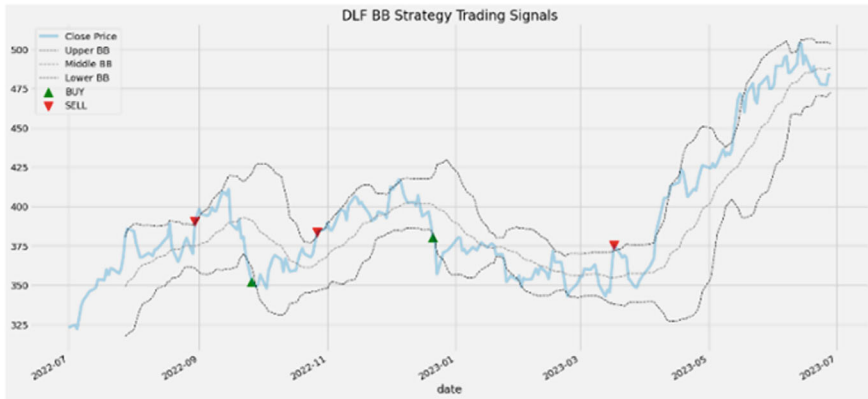


Fig. 18.40 The Bollinger Bands plot of DLF stock with the trading signal points identified

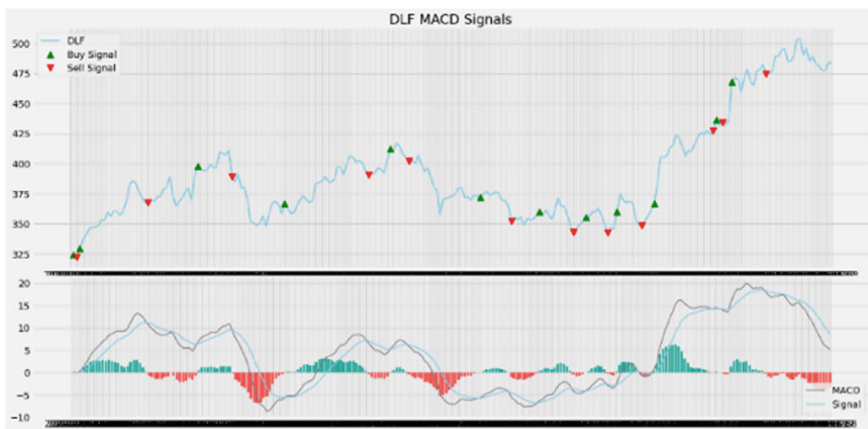


Fig. 18.41 The MACD plot of DLF stock with the trading signal points identified

low, indicating a weakening downward trend. Conversely, bearish divergence occurs when the stock price establishes a new high, but RSI forms a lower high, suggesting weakening upward momentum.

The results clearly indicate that stocks from most sectors did not exhibit overbought or oversold patterns except for the media sector stocks. Hence, RSI performed well only for 17 out of 140 stocks from 14 sectors.

It's important to recognize that each of these indicators has advantages, which are influenced by market conditions and the behavior of stock prices. Consequently, traders must choose a suitable tool based on the traits of the time series they are analyzing.

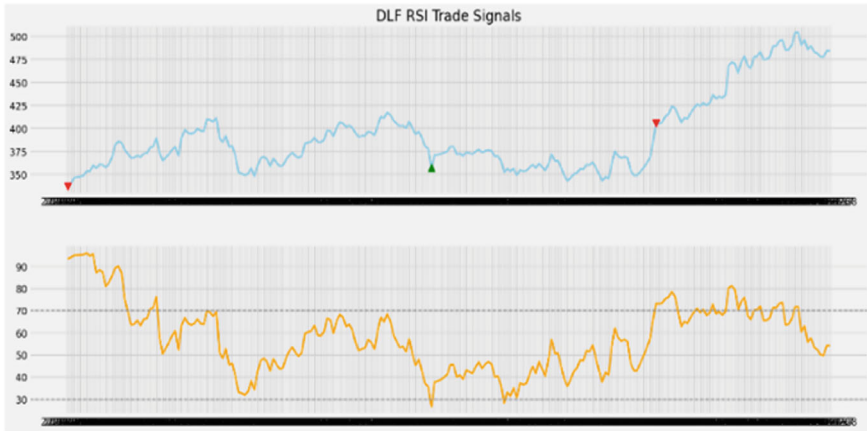


Fig. 18.42 The RSI plot of DLF stock with the trading signal points identified

Table 18.15 The Realty sector annual returns (in percentage) of the BB, MACD, and RSI methods

Stocks	BB	MACD	RSI
DLF	19.18	13.61	9.99
GODREJPROP	2.27	15.84	- 4.68
LODHA	7.16	- 19.49	11.65
PHOENIXLTD	18.27	2.00	11.45
OBEROIRLTY	41.72	7.84	17.17
PRESTIGE	22.95	26.81	15.56
BRIGADE	18.28	5.87	0.00
MAHLIFE	2.24	10.33	6.45
IBREALEST	2.57	14.38	13.04
SOBHA	22.42	1.78	- 11.53

18.5 Applications of AI and Machine Learning

The study of the effectiveness of Bollinger Bands, MACD, and RSI in the stock market can be significantly enhanced by incorporating AI and machine learning techniques. Some of these methods and approaches are discussed in this section.

- (i) *Data collection and preprocessing:* In this study, the data gathering was automated using web scraping and Yahoo Finance APIs to obtain stock price data from the NSE website.
- (ii) *Data cleaning and normalization:* In the current work, missing data detection and imputation have been done using machine learning techniques. Data normalization has also ensured consistency across different stocks and sectors.

Table 18.16 The summary results of the three technical indicators on the fourteen sectors (Period: July 1, 2022–June 30, 2023)

Sector	BB	MACD	RSI
Auto	2	6	2
Banking	5	5	0
Financial Services Ex Banks	3	7	0
Consumer Durables	7	3	0
FMCG	6	4	0
Information Technology	5	3	2
Media	2	3	5
Metal	5	3	2
Mid-Small IT & Telecom	4	6	0
Oil and Gas	5	2	3
Pharma	4	4	2
Private Banks	7	3	0
PSU Banks	1	9	0
Realty	5	4	1
Total	61	62	17

- (iii) *Enhanced feature extraction*: The technical analysis of stocks can be further improved using AI and deep learning to derive additional features from raw stock price data, such as volatility measures, sector-specific indicators, and macroeconomic factors.
- (iv) *Dimensionality reduction*: Methods like principal component analysis (PCA) from machine learning can decrease the number of features in extensive datasets while preserving the most relevant ones, enhancing model efficiency and performance.
- (v) *Predictive model development*: Machine learning models can be further developed to predict stock price movement based on technical indicators and other derived features. Models such as neural networks, gradient boosting, random forests, and the like can be designed for this purpose.
- (vi) *Signal optimization*: Use of AI to optimize buy and sell signals from Bollinger Bands, MACD, and RSI. As an illustration, reinforcement learning can be utilized to acquire optimal trading tactics by maximizing returns within a simulated setting.
- (vii) *Algorithmic trading*: Machine learning algorithms can be implemented to automate trading decisions based on real-time data, allowing for faster and potentially more profitable trades.
- (viii) *Performance analysis*: Machine learning techniques are used to statistically compare the performance of Bollinger Bands, MACD, and RSI across different stocks and sectors. Techniques such as cross-validation and bootstrapping can provide robust performance estimates.

- (ix) *Ensemble methods*: Predictions from multiple technical indicators can be combined using ensemble methods like stacking, boosting, or bagging to improve overall trading performance.
- (x) *Predictive risk analysis*: Utilizing AI and machine learning models enables the anticipation of potential risks and downturns tied to various trading tactics. Integrating machine learning methodologies can elevate methods such as conditional value at risk.
- (xi) *Dynamic portfolio optimization*: Machine learning and AI algorithms can dynamically adjust the portfolio composition based on predicted market conditions and technical indicator performance.
- (ix) *Visualization and interpretation*: While several visualization techniques have been used to exhibit the performance of the technical indicators, AI-powered tools can be further used to create dynamic and interactive visualizations of trading performance, technical indicator signals, and comparative analysis results.
- (x) *Natural language processing*: NLP techniques can be utilized to analyze news sentiment and the trends in social, integrating these insights with technical indicators to enhance trading decisions.
- (xi) *Model retraining*: The predictive models can be continuously retrained with new data to adapt to changing market conditions using AI and machine learning. Automated pipelines can be set up for ongoing data collection, model training, and evaluation.
- (xii) *Feedback loops*: Feedback loops can be established where the performance of trading strategies is monitored, and AI models are adjusted based on the outcomes to improve future performance.

18.6 Conclusion

This chapter explored three prominent and extensively utilized technical indicators—Bollinger Bands, MACD, and RSI—and conducted a comparative investigation into their efficacy within the context of the Indian stock market. The study focused on stocks selected from 14 sectors listed on the NSE of India. The top 10 stocks in each sector are determined based on their free-float market capitalization, as reported by the NSE on July 1, 2022 (NSE Website). Trading activities were conducted for one year, from July 1, 2022, to June 30, 2023, with an initial capital of Indian Rupees (INR) 100,000, employing the three technical indicators. The technical indicator that produces the highest return for each stock is identified, and a comparative analysis is performed based on the overall performance of these indicators across all 14 sectors.

The analysis reveals that Bollinger Bands, MACD, and RSI approaches yielded the highest returns for 61, 62, and 17 stocks, respectively. Notably, Bollinger Bands and MACD strategies demonstrated comparable effectiveness, whereas RSI exhibited a distinctly subpar performance. Specifically, the Bollinger Bands strategy excelled in the *consumer durables* and *private banks* sectors, while the MACD strategy stood

out in the *PSU banks* sector. Despite its generally lackluster performance, the RSI strategy demonstrated reasonable effectiveness in the *media* sector.

However, generalizing findings from a study of 140 stocks from 14 sectors listed on the NSE of India to all stock markets in the world requires careful consideration of several factors. There are several factors in support and in opposition to an attempt of generalization of the results of the study,

The factors that can be used in support of generalizations are the following. First, the study covers various sectors, suggesting that the findings are not limited to a specific industry. This diversity can enhance the robustness of the conclusions, as different sectors often exhibit different trading patterns and behaviors. Second, with 140 stocks analyzed, the sample size is relatively large. Larger sample sizes tend to provide more reliable and statistically significant results, reducing the impact of anomalies or outliers.

However, there are stronger factors against generalizations, too. First, each stock market has its own characteristics, regulations, and behaviors influenced by local economic conditions, investor behavior, and regulatory environments. The NSE operates in the context of India's economic landscape, which may differ significantly from other markets, such as those in the USA, Europe, and East Asia. Second, investor behavior can vary widely across different regions. Cultural attitudes toward risk, investment strategies, and market participation can affect how stocks are traded and how indicators perform. Third, different stock exchanges have varying regulations that can influence market behavior. Factors like trading hours, transaction costs, and the presence of certain market participants (e.g., institutional investors) can differ, impacting the effectiveness of trading strategies. Fourth, the economic context during the study period can significantly impact the results. Macroeconomic factors such as interest rates, inflation, and economic growth rates vary across regions and can affect stock performance and the effectiveness of technical indicators.

While the study's sample size and sector diversity are strengths of the current study, the external validity (i.e., the extent to which the results can be generalized to other contexts) remains in question without further evidence. Similar studies need to be conducted across different markets to generalize with higher confidence to validate the findings. Conducting similar studies in major markets such as NYSE, LSE, TSE, etc., and comparing the results would provide a stronger basis for generalization. If similar patterns are observed across these studies, the argument for generalization would be stronger. Analyzing the correlation between the NSE and other global markets can provide insights into the potential for generalization. Highly correlated markets might exhibit similar behaviors, making generalization more plausible.

Conversely, markets that operate independently may show different results. Finally, the market conditions can change over time. The period during which the study was conducted might have unique characteristics (e.g., post-pandemic recovery, economic downturn) that influenced the results. Further studies should consider different periods to ensure the robustness of the findings over time.

Another interesting research direction will encompass an investigation into constructing resilient portfolios by incorporating the three technical indicators. The

objective is to assess the performance of these portfolios, aiming to gain insights into the efficacy of the indicators in a portfolio context.

References

1. Person, J. L: A Complete Guide to Technical Trading Tactics: How to Profit Using Pivot Points, Candlesticks & Other Indicators. Wiley. ISBN: 978-471-58455-1 (2012)
2. Carta, S.M., Consoll, S., Podda, A.S., Recupero, D.R., Stanciu, M.M.: Ensembling and dynamic asset selection for risk-controlled statistical arbitrage. *IEEE Access* **9**, 29942–29959 (2021). <https://doi.org/10.1109/ACCESS.2021.3059187>
3. Chatterjee, A., Bhowmick, H., Sen, J.: Stock price prediction using time series, econometric, machine learning, and deep learning models. In: Proceedings of the IEEE Mysore Sub Section International Conference (MysuruCon'21), pp. 289–296, October 24–25, Hassan, Karnataka, India (2021). <https://doi.org/10.1109/MysuruCon52639.2021.9641610>
4. Mehtab, S., Sen, J.: A time series analysis-based stock price prediction using machine learning and deep learning models. *Int. J. Bus. Forecast. Mark. Intell.* **6**(4), 272–335 (2021). <https://doi.org/10.1504/IJBFMI.2020.115691>
5. Mehtab, S., Sen, J.: Stock price prediction using convolutional neural networks on a multivariate time series. In: Proceedings of the 3rd National Conference on Machine Learning and Artificial Intelligence (NCMLAI'20), February 1, New Delhi, India (2020). <https://doi.org/10.36227/techrxiv.15088734.v1>
6. Mehtab, S., Sen, J.: A robust predictive model for stock price prediction using deep learning and natural language processing. In: Proceedings of the 7th International Conference on Business Analytics and Intelligence (BAICONF'19), December 5–7, Bangalore, India (2019). <https://doi.org/10.36227/techrxiv.15023361.v1>
7. Mehtab, S., Sen, J., Dutta, A.: Stock price prediction using machine learning and LSTM-based deep learning models. In: Thampi, S. M., Piramuthu, S., Li, K.C., Beretti, S., Wozniak, M., Singh, D. (eds), *Machine Learning and Metaheuristics Algorithms, and Applications (SoMMA'20)*, pp 86–106, Communications in Computer and Information Science, vol. 1366. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-0419-5_8
8. Sarmiento, S.M., Horta, N.: Enhancing a pairs trading strategy with the application of machine learning. *Expert. Syst. Appl.* **158**, Art ID 113490 (2020). <https://doi.org/10.1016/j.eswa.2020.113490>
9. Sen, J.: Stock price prediction using machine learning and deep learning frameworks. In: Proceedings of the 6th International Conference on Business Analytics and Intelligence (ICBAI'18), December 20–22, Bangalore, India (2018)
10. Sen, J., Datta Chaudhuri, T.: A robust predictive model for stock price forecasting. In: Proceedings of the 5th International Conference on Business Analytics and Intelligence (BAICONF'17), December 11–13, Bangalore, India (2017). <https://doi.org/10.36227/techrxiv.16778611.v1>
11. Chen, Y-Y., Chen, W-L., Huang, S-H.: Developing arbitrage strategy in high-frequency pairs trading with filterbank CNN algorithm. In: Proceedings of the 2018 IEEE International Conference on Agents (ICA'18), pp. 113–116, July 28–31, Singapore (2018). <https://doi.org/10.1109/AGENTS.2018.8459920>
12. Chong, E., Han, C., Park, F.C.: Deep learning networks for stock market analysis and prediction: methodology, data, representations, and case studies. *Expert. Syst. Appl.* **85**, 187–205 (2017). [https://doi.org/10.1016/j.eswa.2017.04.030\(2017\)](https://doi.org/10.1016/j.eswa.2017.04.030(2017))
13. Sen J., Mehtab, S.: Accurate stock price forecasting using robust and optimized deep learning models. In: Proceedings of the IEEE International Conference on Intelligent Technologies (CONIT), pp. 1–9, June 25–27, Hubballi, India (2021). <https://doi.org/10.1109/CONIT51480.2021.9498565>

14. Mehtab, S., Sen, J.: Stock price prediction using CNN and LSTM-based deep learning models. In: Proceedings of the IEEE International Conference on Decision Aid Sciences and Applications (DASA'20), pp. 447–453, November 8–9, Sakheer, Bahrain (2020). <https://doi.org/10.1109/DASA51403.2020.9317207>
15. Mehtab, S., Sen, J., Dasgupta, S.: Robust analysis of stock price time series using CNN and LSTM-based deep learning models. In: Proceedings of the IEEE 4th International Conference on Electronics, Communication and Aerospace Technology (ICCEA'20), pp. 1481–1486, November 5–7, Coimbatore, India (2020). <https://doi.org/10.1109/ICECA49313.2020.9297652>
16. Sen, J., Mehtab, M.: Design and analysis of robust deep learning models for stock price prediction. In: Sen, J. (ed.) Machine Learning—Algorithms, Models and Applications, pp. 15–46, IntechOpen, London, UK (2021). <https://doi.org/10.5772/intechopen.99982>
17. Sen, J., Mondal, S., Mehtab, S.: Analysis of sectoral profitability of the Indian stock market using an LSTM regression model. In: Proceedings of the Deep Learning Developers' Conference (DLDC'21), September 24, Bangalore, India (2021). <https://doi.org/10.36227/techrxiv.17048579.v1>
18. Sen, J., Mehtab, S., Nath, G.: Stock price prediction using deep learning models. *Lattice: Mach. Learn. J.* **1**(3), 34–40 (2020). <https://doi.org/10.36227/techrxiv.16640197.v1>
19. Sen, J., Mehtab, S.: Long-and-short-term memory (LSTM) price prediction-architectures and applications in stock price prediction. In: Singh, U., Murugesan, S., Seth, A. (eds.) Emerging Computing Paradigms—Principles, Advances, and Applications, Wiley, USA (2022). <https://doi.org/10.1002/9781119813439.ch8>
20. Thormann, M-L., Farchmin, J., Weissner, C., Kruse, R-M., Safken, B., Silbersdorff, A.: Stock price predictions with LSTM neural networks and twitter sentiments. *Stat. Optim. Inf. Comput.* **9**(2), 268–287 (2021). <https://doi.org/10.19139/soic-2310-5070-1202>
21. Tran, D.T., Iosifidis, A., Kannianen, J., Gabbouj, M.: Temporal attention-augmented bilinear network for financial time-series data analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(5), 1407–1418 (2019). <https://doi.org/10.1109/TNNLS.2018.2869225>
22. Mehtab S., Sen, J.: Analysis and forecasting of financial time series using CNN and LSTM-based deep learning models. In: Sahoo, J.P., Tripathy, A.K., Mohanty, M., Li, K.C., Nayak, A.K. (eds.) Advances in Distributed Computing and Machine Learning, Lecture Notes in Networks and Systems, vol. 302, pp. 405–423, Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-4807-6_39
23. Li, Y., Pan, Y. A.: A novel ensemble deep learning model for stock prediction based on stock prices and news. *Int. J. Data Sci. Anal.* **13**, 139–149 (2022). <https://doi.org/10.1007/s41060-021-00279-9>
24. Zhang, Y., Li, J., Wang, H., Choi, S-C.T.: Sentiment-guided adversarial learning for stock price prediction. *Front. Appl. Math. Stat.* **7**, Art ID: 601105 (2021). <https://doi.org/10.3389/fams.2021.601105>
25. Sen, J., Dutta, A., Mehtab, S.: Stock portfolio optimization using a deep learning LSTM model. In: Proceedings of the IEEE Mysore Sub Section International Conference (MysuruCon'21), pp. 263–271, October 24–25, Hassan, Karnataka, India (2021). <https://doi.org/10.1109/MysuruCon52639.2021.9641662>
26. Sen, J., Dutta, A., Mehtab, S.: Profitability analysis in stock investment using an LSTM-based deep learning model. In: Proceedings of the IEEE 2nd International Conference for Emerging Technology (INCET'21), pp. 1–9, May 21–23, Belagavi, India (2021). <https://doi.org/10.1109/INCET51464.2021.9456385>
27. Cheng, D., Liu, Y., Niu, Z., Zhang, L.: Modeling similarities among multi-dimensional financial time series. *IEEE Access* **6**, 43404–43414 (2018). <https://doi.org/10.1109/ACCESS.2018.2862908>
28. Sen, J.: A forecasting framework for the Indian healthcare sector index. *Int. J. Bus. Forecast. Mar-Keting Intell. (IJBFMI)* **7**(4), 311–350 (2021). [https://doi.org/10.1504/IJBFMI.2022.10047095\(2022\)](https://doi.org/10.1504/IJBFMI.2022.10047095(2022))

29. Sen, J., Datta Chaudhuri, T.: A time series analysis-based forecasting framework for the Indian healthcare sector. *J. Insur. Financ. Manag.* **3**(1), 66–94 (2017). <https://doi.org/10.36227/tehrxiv.16640221.v1>
30. Sen, J.: Stock composition of mutual funds and fund style: a time series decomposition approach towards testing for consistency. *Int. J. Bus. Forecast. Mark. Intell.* **4**(3), 235–292 (2018). <https://doi.org/10.1504/IJBFMI.2018.092781>
31. Sen, J.: A time series analysis-based forecasting approach for the Indian realty sector. *Int. J. Appl. Econ. Stud.* **5**(4), 8–27 (2017). <https://doi.org/10.36227/tehrxiv.16640212.v1>
32. Sen, J.: A robust analysis and forecasting framework for the Indian mid cap sector using time series decomposition approach. *J. Insur. Financ. Manag.* **3**(4), 1–32 (2017). <https://doi.org/10.36227/tehrxiv.15128901.v1>
33. Sen, J., Datta Chaudhuri, T.: Understanding the sectors of indian economy for portfolio choice. *Int. J. Bus. Forecast. Mark. Intell.* **4**(2), 178–222 (2018). <https://doi.org/10.1504/IJBFMI.2018.090914>
34. Sen, J., Datta Chaudhuri, T.: A predictive analysis of the Indian FMCG sector using time series decomposition-based approach. *J. Econ. Libr.* **4**(2), 206–226 (2017). <https://doi.org/10.1453/jel.v4i2.1282>
35. Sen, J., Datta Chaudhuri, T.: Decomposition of time series data to check consistency between fund style and actual fund composition of mutual funds. In: Proceedings of the 4th International Conference on Business Analytics and Intelligence (ICBAI'16), December 19–21 (2016). <https://doi.org/10.13140/RG.2.2.33048.19206>
36. Sen, J., Datta Chaudhuri, T.: An investigation of the structural characteristics of the Indian IT sector and the capital goods sector—an application of the R programming language in time series decomposition and forecasting. *J. Insur. Financ. Manag.* **1**(4), 68–132 (2016). <https://doi.org/10.36227/tehrxiv.16640227.v1>
37. Sen, J., Datta Chaudhuri, T.: An alternative framework for time series decomposition and forecasting and its relevance for portfolio choice—a comparative study of the Indian consumer durable and small cap sectors. *J. Econ. Libr.* **3**(2), 303–326 (2016). <https://doi.org/10.48550/arXiv.1605.03930>
38. Sen, J., Datta Chaudhuri, T.: Decomposition of time series data of stock markets and its implications for prediction—an application for the Indian auto sector. In: Proceedings of the 2nd National Conference on Advances in Business Research and Management Practices (ABRMP'16), pp. 15–28, January 8–9 (2016). <https://doi.org/10.13140/RG.2.1.3232.0241>
39. Sen, J., Datta Chaudhuri, T.: A framework for predictive analysis of stock market indices—a study of the Indian auto sector. *J. Manag. Pract.* **2**(2), 1–20 (2015). <https://doi.org/10.13140/RG.2.1.2178.3448>
40. Sen, J., Mehtab, S., Dutta, A.: Volatility modeling of stocks from selected sectors of the Indian economy using GARCH. In: Proceedings of the IEEE Asian Conference on Innovation in Technology (ASIANCON'21), pp. 1–9, August 28–29, Pune, India (2021). <https://doi.org/10.1109/ASIANCON51346.2021.9544977>
41. Brim, A.: Deep reinforcement learning pairs trading with a double deep Q-network. In: Proceedings of the 10th Annual Computing and Communication Workshop and Conference (CCWC'20), pp. 222–227, January 6–8, 2020, Las Vegas, Nevada, USA (2020). <https://doi.org/10.1109/CCWC47524.2020.9031159>
42. Fengqian, D., Chao, L.: An adaptive financial trading system using deep reinforcement learning with candlestick decomposing features. *IEEE Access* **8**, 63666–63678 (2020). <https://doi.org/10.1109/ACCESS.2020.2982662>
43. Kim, S-H., Park, D-Y., Lee, K-H.: Hybrid deep reinforcement learning for pairs trading. *Appl. Sci.* **12**(3), Art ID 944 (2022). <https://doi.org/10.3390/app12030944>
44. Kim, T., Kim, H.Y.: Optimizing the pair-trading strategy using deep reinforcement learning with trading and stop-loss boundaries. *Complex. Financ. Mark.* **2019**, Art ID: 3582516 (2019). <https://doi.org/10.1155/2019/3582516>
45. Lei, K., Zhang, B., Li, Y., Yang, M., Shen, Y.: Time-driven feature-aware jointly deep reinforcement learning for financial signal representation and algorithmic trading. *Expert. Syst. Appl.* **140**, Art ID 112872 (2020). <https://doi.org/10.1016/j.eswa.2019.112872>

46. Li Y., Zheng, W., Zheng, Z.: Deep robust reinforcement learning for practical algorithmic trading. *IEEE Access* **7**, 108014–108022 (2019). <https://doi.org/10.1109/ACCESS.2019.2932789>
47. Lu, J.-Y., Lai, H.-C., Shih, W.-Y., Chen, Y.-F., Huang, S.-H., Chang, H.-H., Wang, J.-Z., Huang, J.-L., Dai, T.-S.: Structural break-aware pairs trading strategy using deep reinforcement learning. *J. Supercomput.* **78**, 3843–3882 (2021). <https://doi.org/10.1007/s11227-021-04013-x>
48. Park, D.-Y., Lee, K.-H.: Practical algorithmic trading using state representation learning and imitative reinforcement learning. *IEEE Access* **9**, 152310–152321 (2021). <https://doi.org/10.1109/ACCESS.2021.3127209>
49. Sen, J.: Portfolio optimization using reinforcement learning and hierarchical risk parity approach. In: Rivera, G., Cruz-Reyes, L., Dorronsoro, B., Rosete-Suarez, A, (eds.) *Data Analytics and Computational Intelligence: Novel Models, Algorithms and Applications, Studies in Big Data*, vol. 132. Springer, Switzerland (2023). https://doi.org/10.1007/978-3-031-38325-0_20
50. Sen, J.: A comparative analysis of portfolio optimization using reinforcement learning and hierarchical risk parity approaches. In: *Proceedings of the 9th International Conference on Business Analytics and Intelligence (BAICONF'22)*, December 15–17, 2022, Bangalore, India (2022). https://doi.org/10.1007/978-3-031-38325-0_20
51. Sen, J., Mehtab, S.: A comparative study of optimum risk portfolio and Eigen portfolio on the Indian stock market. *Int. J. Bus. Forecast. Mark. Intell.* **7**(2), 143–195 (2022). <https://doi.org/10.1504/IJBFMI.2021.120155>
52. Sen, J., Mehtab, S., Dutta, A., Mondal, S.: Precise stock price prediction for optimized portfolio design using an LSTM model. In: *Proceedings of the IEEE 19th International Conference on Information Technology (OCIT'12)*, pp. 210–215, December 16–18. Bhubaneswar, India (2021). <https://doi.org/10.1109/OCIT53463.2021.00050>
53. Sen, J., Mondal, S., Nath, G.: Robust portfolio design and stock price prediction using an optimized LSTM model. In: *Proceedings of the IEEE 18th India Council International Conference (INDICON'21)*, pp. 1–6, December 19–21, Guwahati, India (2021). <https://doi.org/10.1109/INDICON52576.2021.9691583>
54. Sen, J., Mondal, S., Mehtab, S.: Portfolio optimization on NIFTY thematic sector stocks using an LSTM model. In: *Proceedings of the IEEE International Conference on Data Analytics for Business and Industry (ICDABI'21)*, pp. 364–369, October 25–26, Bahrain (2021). <https://doi.org/10.1109/ICDABI53623.2021.9655886>
55. Fernandez, E., Gomez, C., Rivera, G., Cruz-Reyes, L.: Hybrid metaheuristic approach for handling many objectives and decisions on partial support in project portfolio optimisation. *Inf. Sci.* **315**, 102–122 (2015). <https://doi.org/10.1016/j.ins.2015.03.064>
56. Sen A., Sen, J.: A study of the performance evaluation of equal-weight portfolio and optimum risk portfolio on the Indian stock market. *Int. J. Bus. Forecast. Mark. Intell. (IJBFMI)* (2024) (In Press). <https://doi.org/10.48550/arXiv.2309.13696>
57. Wang, Z., Zhang, X., Zhang, Z., Sheng, D.: Credit portfolio optimization: a multi-objective genetic algorithm approach. *Borsa Istanbul Rev.* **22**(1), 69–76 (2022). <https://doi.org/10.1016/j.bir.2021.01.004>
58. Zheng, Y., Zheng, J.: A novel portfolio optimization model via combining multi-objective optimization and multi-attribute decision making. *Appl. Intell.* **52**, 5684–5695 (2022). <https://doi.org/10.1007/s10489-021-02747-y>
59. Sen, J., Dutta, A.: Design and analysis of optimized portfolios for selected sectors of the Indian stock market. In: *Proceedings of the 2022 International Conference on Decision Aid Sciences and Applications (DASA)*, pp. 567–573, March 23–25, Chiangrai, Thailand (2022). <https://doi.org/10.1109/DASA54658.2022.9765289>
60. Sen, J., Dutta, A.: A comparative study of hierarchical risk parity portfolio and Eigen portfolio on the NIFTY 50 stocks. In: Buyya, R., Hernandez, S.M., Kovvur, R.M.R., Sarma, T.H. (eds.) *Computational Intelligence and Data Analytics, Lecture Notes on Data Engineering and Communications Technologies*, vol. 142, pp. 443–460. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-3391-2_34

61. Sen, J., Dutta, A.: Portfolio optimization for the Indian stock market. In: Wang, J. (ed.) *Encyclopedia of Data Science and Machine Learning*, pp. 1904–1951, IGI Global, USA, August (2022). <https://doi.org/10.4018/978-1-7998-9220-5.ch115>
62. Sen, J., Dutta, A.: Risk-based portfolio optimization on some selected sectors of the Indian stock market. In: Borah, M.D., Laiphrakpam, D.S., Auluck, N., Balas, V.E. (eds) *Big Data, Machine Learning, and Applications, BigDML, 2021. Lecture Notes in Electrical Engineering*, vol. 1053, pp. 765–778 (2021). https://doi.org/10.1007/978-981-99-3481-2_58
63. Sen, J., Dutta, A., Mondal, S., Mehtab, S.: A comparative study of portfolio optimization using optimum risk and hierarchical risk parity approaches. In: *Proceedings of the 8th International Conference on Business Analytics and Intelligence (ICBAI'21)*, December 20–22, Bangalore, India (2021). <https://doi.org/10.13140/RG.2.2.35308.28809>
64. Sen, J., Mehtab, S., Dutta, A., Mondal, S.: Hierarchical risk parity and minimum variance portfolio design on NIFTY 50 stocks. In: *Proceedings of the IEEE International Conference on Decision Aid Sciences and Applications (DASA'21)*, December 7–8, Sakheer, Bahrain (2021). <https://doi.org/10.1109/DASA53625.2021.9681925>
65. Corazza, M., di Tollo, G., Fasano, G., Pesenti, R.: A novel hybrid PSO-based metaheuristic for costly portfolio selection problem. *Ann. Oper. Res.* **304**, 109–137 (2021). <https://doi.org/10.1007/s10479-021-04075-3>
66. Thakkar, A., Chaudhuri, K.: A comprehensive survey on portfolio optimization, stock price and trend prediction using particle swarm optimization. *Arch. Comput. Methods Eng.* **28**, 2133–2164 (2021). <https://doi.org/10.1007/s11831-020-09448-8>
67. Kaucic, M., Moradi M., Mirzazadeh, M.: Portfolio optimization by improved NSGA-II and SPEA 2 based on different risk measures. *Financ. Innov.* **5**(26), Art ID: 26 (2019). <https://doi.org/10.1186/s40854-019-0140-6>
68. Karimi, M., Tahayori, H., Tirdad, K., Sadeghian, A.: A perceptual computer for hierarchical portfolio selection based on interval type-2 fuzzy sets. *Granul. Comput.* (2022). <https://doi.org/10.1007/s41066-021-00311-0>
69. Li, Y., Zhou, B., Tan, Y.: Portfolio optimization model with uncertain returns based on prospect theory. *Complex Intell. Syst.* (2021). <https://doi.org/10.1007/s40747-021-00493-9>
70. Chou, Y-H., Jiang, Y-C., Kuo, S-Y.: Portfolio optimization in both long and short selling trading using trend ratios and quantum-inspired evolutionary algorithms. *IEEE Access* **9**, 152115–152130 (2021). <https://doi.org/10.1109/ACCESS.2021.3126652>
71. Flori, A., Regoli, D.: Revealing pairs-trading opportunities with long short-term memory networks. *Eur. J. Oper. Res.* **295**(2), 772–791 (2021). <https://doi.org/10.1016/j.ejor.2021.03.009>
72. Gupta, K., Chatterjee, N.: Selecting stock pairs for pairs trading while incorporating lead-lag relationship. *Phys. A: Stat. Mech. Its Appl.* **551**, Art ID 124103 (2020). <https://doi.org/10.1016/j.physa.2019.124103>
73. Ramos-Requena, J.P., Lopez-Garcia, M.N., Sanchez-Granero, M.A., Trinidad-Segovia, J.E.: A cooperative dynamic approach to pairs trading. *Complex. Financ. Mark.* **2021**, Art ID 7152846 (2021). <https://doi.org/10.1155/2021/7152846>
74. Sen, J.: Optimum pair-trading strategies for stocks using cointegration-based approach. In: *Proceedings of the IEEE 29th OITS International Conference on Information Technology (OCIT'22)*, December 14–16, Bhubaneswar, India (2022). <https://doi.org/10.1109/OCIT56763.2022.00076>
75. Sen, J.: Designing efficient pair-trading strategies using cointegration for the Indian stock market. In: *Proceedings of the IEEE 2nd Asian Conference on Innovation in Technology (ASIANCON'22)*, pp. 1–9, Pune, India, August (2022). <https://doi.org/10.1109/ASIANCON55314.2022.9909455>
76. Seshu, V., Shanbhag, H., Rao, S.R., Venkatesh, Agarwal, P., Arya, A.: Performance analysis of Bollinger bands and long short-term memory (LSTM) models based strategies on NIFTY50 companies. In: *Proceedings of 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, pp. 184–190 (2022). <https://doi.org/10.1109/Confluence52989.2022.9734127>
77. NIFTY 50 Wiki Page: https://en.wikipedia.org/wiki/NIFTY_50. Accessed on 15 May 2024

78. Zheng, Y., Li, X., Feng, Y.: Research on the quantitative trading strategy based on Bollinger band strategy and polynomial regression model. In: Proceedings of the 2nd International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 2022, pp. 1255–1260 (2022). <https://doi.org/10.1109/ICDSCA56264.2022.9988398>
79. Lauguico, S., Concepcion, RII., Alejandro, J., Macasaet, D., Tobias, R.R., Bandala, A., Dadios, E.: A fuzzy logic-based stock market trading algorithm using Bollinger bands. In: Proceedings of IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), Laoag, Philippines, pp 1–6 (2019). <https://doi.org/10.1109/HNICEM48295.2019.9072734>
80. Au, S.C., Keung, J.W.: New technique for stock trend analysis-volume-weighted squared moving average convergence & divergence. In: Proceedings of the 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy, pp. 987–988 (2023). <https://doi.org/10.1109/COMPSAC57700.2023.00140>
81. Deac, G.-A., Iancu, D.-T.: Trading strategy hyper-parameter optimization using genetic algorithm. In: Proceedings of the 24th International Conference on Control Systems and Computer Science (CSCS), Bucharest, Romania, pp. 121–127 (2023). <https://doi.org/10.1109/CSCS59211.2023.00028>
82. Chen, Y., Huang, L-E., Wang, P-H., Tang, J-H., Hsu, K-N., Chou, Y-H., Kuo, S-Y.: A dynamic stock trading system using GNQTS and RSI in the U.S. stock market. In: Proceedings of 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, pp. 456–461 (2022). <https://doi.org/10.1109/SMC52423.2021.9659251>
83. Zatwarnicki, M., Zatwarnicki, K., Stolarski, P.: Effectiveness of the relative strength index signals in timing the cryptocurrency market. *Sensors (Basel)* **23**(3), Art Id: 1664 (2023). <https://doi.org/10.3390/s23031664>
84. NSE Website: <http://www1.nseindia.com>. Accessed on 15 May 2024
85. Google Colab: <https://colab.research.google.com>. Accessed on 15 May 2023

Chapter 19

PD-Monitor: A Self-management App for Monitoring Patients with Parkinson's Disease



Giner Alor-Hernández , Laura-Nely Sánchez-Morales ,
Francisco-Javier García-Dimas, Nancy-Aracely Cruz-Ramos ,
and José-Luis Sánchez-Cervantes 

Abstract Parkinson's disease is a neurodegenerative disease that affects the nervous system and usually occurs in adulthood. Although different treatments exist, artificial intelligence (AI) represents an important contribution to the treatment of Parkinson's. In this sense, large volumes of data can be processed, and patterns can be identified to facilitate the early diagnosis of Parkinson's disease (PD). In addition, AI can provide additional information for decision-making and personalized PD monitoring. Therefore, we propose a set of algorithms based on AI techniques for symptom monitoring of PD patients. Furthermore, we present a software module as a proof-of-concept of the proposed algorithms. This module makes recommendations to control the symptoms of the disease through notifications and alerts for PD patients, using artificial intelligence techniques. In addition, we present the architecture design, which consists of two main functionalities: (1) Voice detection to extract information from the patient's voice, and (2) Freehand drawing detection to identify tremors in patients' hands. Finally, two case studies are presented as a proof of concept of the proposed module.

Keywords Artificial intelligence techniques · Neurodegenerative diseases · Parkinson's disease

G. Alor-Hernández (✉) · F.-J. García-Dimas · N.-A. Cruz-Ramos · J.-L. Sánchez-Cervantes
Tecnológico Nacional de México/I.T., Orizaba, México
e-mail: giner.ah@orizaba.tecnm.mx

F.-J. García-Dimas
e-mail: M21011173@orizaba.tecnm.mx

N.-A. Cruz-Ramos
e-mail: d08011250@orizaba.tecnm.mx

J.-L. Sánchez-Cervantes
e-mail: jose.sc@orizaba.tecnm.mx

L.-N. Sánchez-Morales
CONAHCYT-Tecnológico Nacional de México/I. T., Orizaba, México
e-mail: laura.sanchez@conahcyt.mx

19.1 Introduction

The number of patients with Parkinson's disease (PD) worldwide has a prevalence of more than 8.5 million people [1]. According to the Ref. [1], it is a neurodegenerative disease that affects the nervous system of patients affected by this disease. Parkinson's is the second neurodegenerative disease, and its incidence is increasing due to the aging of the population [2]. The Ref. [3] in the USA has reported that 4% of the Parkinson's population was diagnosed before the age of 50. Men are 1.5 times more likely to develop Parkinson's disease compared to women. In addition, the Ref. [3] estimated in their study that, by 2030, 1.2 million people in the USA will be living with Parkinson's disease.

Different techniques have been reported in the literature to detect Parkinson's disease. Some of these techniques are based on voice analysis, freehand drawing analysis, body movement, hand movement, or the use of magnetic resonance imaging which is one of the most reliable techniques. In the same vein, PD patients need technological solutions that allow them to improve their quality of life, regardless of the diagnostic technique.

Therefore, our contribution is a set of algorithms based on Artificial Intelligence (AI) techniques implemented in the PD-Monitor module as a proof of concept. PD-Monitor is a module of recommendations, notifications, and alerts to monitor the main symptoms of patients with PD. The main characteristic of this module is that it is the least invasive, for this purpose it has been adapted with two functions: (1) Voice detection to extract information from the patient's voice, and (2) Freehand drawing detection to identify tremors in patients' hands.

The motivation of this research is to propose a comprehensive approach to the care of PD patients. This comprehensive approach is based on voice analysis and freehand drawing techniques. It can be scalable to other biometric variables such as stuttering, and hand tremors, which are the most visible in patients with Parkinson's disease.

The chapter is composed of 6 sections, Sect. 19.1 is an introduction, and Sect. 19.2 introduces a set of works identified in the literature. Section 19.3 describes the design of the PD-Monitor architecture. Section 19.4 describes the PD-Monitor voice and freehand drawing modules. Section 19.5 presents two cases of study as a proof of concept of PD-Monitor. Finally, Sect. 19.6 describes the conclusions and future work.

19.2 Related Work

This section presents a comprehensive literature review of the solutions developed for monitoring patients with Parkinson's disease.

Reference [4] proposed an application for mobile devices. The main functionality consisted of analyzing the information obtained by a smartphone positioned on the patient's arm to transmit the data to the MATLAB software. The results indicated

an accuracy of 95% and a Kappa coefficient of 90%. Reference [5] developed an application for tremor quantification called TREMOR12. The application obtained acceleration and rotation samples of hand tremors. The data from these samples were exported as a comma-separated 3-value file for further analysis. The results showed that the application was able to detect and record tremor characteristics in PD patients. While, in Ref. [6] realized an IoT-based biomedical telemonitoring system. The system provided remote care to patients with Parkinson's, heart disease, and diabetes, using sensors embedded in shoes and a belt to monitor vital points. Data was transmitted via IoT and stored in the cloud for future medical analysis, including activity tracking, location, and fall detection through an Android-based application. Reference [7] described Parkinson's Home Exercises, an app that contains videos of more than 50 home exercises, as well as movement tips and instructions for daily exercises and mobility. This app was designed for use by patients and therapists. The tips and high-quality video exercises have been compiled by expert researchers and therapists in the field.

Reference [8] developed an application for mobile devices called STOP, which integrated a game to monitor PD symptoms and a medication intake diary. An evaluation was conducted with PD patients who reported "a sense of control". In addition, the evaluation generated meaningful information about the future work of STOP. Meanwhile, Ref. [9] presented a user-centered design (UCD) process of an interface for Parkinson's disease (PD) patients to help them better manage their symptoms. The interface enabled the visualization of symptom and medication information collected through an Internet of Things (IoT)-based system. The IoT system consisted of a smartphone, a wrist sensor, a bed sensor, and an electronic dosing device. Reference [10], proposed a new machine-learning approach to assess the severity of PD patients. The approach generated a mobile Parkinson's disease (mPD) score from a set of activities (finger tapping, voice, gait, reaction time, and balance) monitored by HopkinsPD application in PD patients before and after their medication. Reference [11] developed a mobile device application to identify PD by measuring a short period of monosyllabic speech. In this case, an algorithm that measures the frequency over a period of time was used to test the characteristics of a PD patient's voiceprint in the spectrogram domain. In addition, a convolutional neural network called DeepVoice was developed to complete the identification. The results show that DeepVoice achieved an accuracy of $90.45 \pm 1.71\%$ with an audio segment as short as 10 s. Reference [12] conducted a systematic review to analyze the capabilities, challenges, and impact of mobile self-directed health research (mHealth) mobile apps such as ResearchKit. As a result, 36 ResearchKit apps were identified. Most of the apps were used to generate datasets for secondary research.

In another perspective, Ref. [13] presented APParkinson, an application that aims to provide support to the Parkinson's community. This application facilitated the control of users' medical information through different functions such as setting medication alerts, taking notes on symptoms, recording notes of interest, and consultation dates, and accessing a series of recommended exercises. Reference [14] conducted a review of PD-related apps available for iOS and Android. Findings included apps for symptom control, exercise, informational apps, and apps for

evaluating PD scales or collecting data for clinical trials. In addition, apps were found for treating PD symptoms, but no app-controlled all symptoms. In Ref. [15] evaluated the usefulness of mobile devices and smartwatches in the treatment of Parkinson's disease (PD). Over a 6-month period, 51 patients used these devices to record movement data, symptom severity, and medication. The physicians reviewed the data at periodic visits, providing feedback to improve the system. Although only 39% completed the study, 83,432 h of data were shared. Symptom and medication reporting was low (40–60%), but other data were adequately transmitted.

While, Ref. [16] conducted a study to measure the acceptability and feasibility of self-management of Parkinson's patients using smartphone applications. The study was applied to 204 participants with PD, of whom 82.84% indicated that they preferred the use of these apps. Reference [17] explored the feasibility of using a mHealth platform for remote PD monitoring, based on a smartphone, a watch, and a pair of smart templates; the system is called PD_manager. They also explored PD-related determinants and validated a tremor assessment method with daily activity data. Of the 75 participants, 65 (87%) completed the study, using the system for 11.57 days on average. The algorithm proved to be effective in detecting and assessing tremors. From another perspective, Ref. [18] presented Apkinson, an application for mobile devices for motor assessment and monitoring of PD patients. Apkinson was based on posture, articulation and pronunciation in speech, regularity and freezing of gait in walking, and hand attitude. Most of the metrics were adequate to discriminate or identify PD patients. In addition, the application helped to track progression accurately. Reference [19] developed an electronic diary (eDiary) specifically for treating and exploring PD, based on ecological momentary assessments (EMA). The eDiary was applied in a group of 20 PD patients, for 14 consecutive days, without adjusting their free-living routines. Correlations were found between the answers given that supported the internal validity of the eDiary.

On the other hand, Ref. [20] developed the mPower 2 application for tracking the status of PD patients using questionnaires and phone sensor data. The main functions of mPower included (1) Innovative activity-based measurements of Parkinson's symptoms, such as finger tapping, tremor measurement, and walking; (2) Knowledge sharing and linkage with researchers; and (3) Tracking triggers and symptom variations. Another application that was presented by Ref. [21] is Rhythm—Parkinson's Gait App. This app encourages correct gait coordination in Parkinson's patients based on listening to rhythmic audio. It is not a medical app, it does not provide medical information, treatment, or diagnosis. In Ref. [22] a review of the use of Digital Biomarkers (DMs) for PD monitoring was presented. As part of the findings, they concluded that MDs derived from speech, gait, voice, handwriting, and face movement have great potential in the field of PD.

Meanwhile, Ref. [23] is a digital tool focused on Parkinson's patients to treat symptoms at home through personalized daily therapies and a dedicated support service. This application supports three main symptoms: mobility (walking), dexterity (fine hand movements), and speech (speech and language).

In Ref. [24] presented a study on a pilot test of the vCare system. The trial consisted of evaluating 10 PD patients using vCare. The study evaluated aspects of quality-of-life improvement in PD patients, adherence to the home care and rehabilitation plan, risk factor reductions, and personalization and health promotion, as well as evaluating the usability and level of satisfaction with the use of vCare. In Ref. [25] presented the design of a physical therapy support system for remote monitoring of patients with PD. The system recorded the physical activity and falls of the patients during 6 weeks, and the usability and usefulness of the system were evaluated. Otherwise, Ref. [26] conducted a study to identify motivations and barriers to monitoring PD symptoms among patients and professionals and understand the benefits and limitations of wearable sensors. The study involved 434 PD patients and 166 healthcare professionals specialized in PD care (Table 19.1).

We found that most of the papers reviewed in this section address proposals related to symptom monitoring of patients with PD. In this regard, we can conclude the following: (1) Most of these papers used wearable devices to collect information from patients with PD. (2) Approaches involving the analysis of two or more techniques for symptom monitoring in patients with PD were not found.

The following section presents the design of the proposed PD-Monitor architecture.

19.3 PD-Monitor's Architecture

The PD-Monitor's architecture is based on a layered approach to achieve an efficient and modular organization. In this architecture, each layer is designed to have a clear and defined responsibility, which allows a better organization and maintenance of the code. In addition, by following a layered architecture, it is easy to remove each layer independently, if necessary.

In general, the process starts with the query of the patient data, followed by the patient's assessment information, as well as the retrieval of the variables that measure voice characteristics and freehand drawing. This information is analyzed by the recommendation system with the machine learning algorithm, and then a message type (recommendation, notification, or alert) is assigned based on the results. The message is sent through the Twilio API to the caregiver's WhatsApp® application and is also displayed through the mobile application.

The PD-Monitor's architecture is shown in Fig. 19.1, which is divided into four layers, each layer is described in detail below.

Integration layer: This is responsible for integrating PD-Monitor with the necessary internal and external services, such as the Twilio API for sending WhatsApp messages and displaying the. Here the controllers or adapters that enable communication between the service layer and the presentation layer are implemented.

Table 19.1 Comparative analysis of the state of the art

Author	Contribution	Technologies	Results
[8]	Development and evaluation of an application to monitor medication intake and symptoms in patients with PD	Smartphone compatible with Android and iOS	STOP application evaluation
[9]	A system using an UCD approach for collecting sleep and motor function measurements in PD patients	Bed sensor Wrist sensor Electronic dosing device Smartphone	The development process of a mock-up of the interface for the management of PD patients' symptoms
[15]	An evaluation of the usefulness of mobile devices and smartwatches for PD treatment	Mobile devices Smartwatches	Analysis of PD patient data collected through the Intel® pharma analytics platform
[18]	Apkinson an application for monitoring PD patients	Sensors embedded on the smartphone (microphone, accelerometer and gyroscope) Recurrent neural network (RNN) Automatic speech recognizer (ASR)	Apkinson's experiments indicated that most of the characteristics evaluated were adequate to discriminate between patients and controls
[19]	Development of a Parkinson's disease-specific eDiary using EMA, and validation of the EMA method using a large group of PD patients	EMA Smartphone Accelerometers and gyroscopes	Design of a questionnaire based on the EMA method for monitoring Parkinson's disease in free living
[20]	mPower 2 an application for symptom monitoring in PD patients	Phone sensors	Follow-up of PD symptoms and collaboration with researchers
[24]	A study to evaluate the vCare system as a telerehabilitation tool, a daily life monitoring, and a virtual trainer in Parkinson's patients	Smart band Sensors: movement and presence STAT-ON device	The vCare system showed high adherence to all measured outcomes
[25]	A physical therapy support system for remote monitoring of patients with PD	A wearable sensor in the form of a necklace (the "GoSafe") And Vital@Home app for android	The usability of the system was assessed as positive, and the perceived usefulness varied among patients

Service layer: This is responsible for providing the necessary services for PD-Monitor, such as reading voice data, freehand drawings, the library of recommendations, and patient information, stored in the database through an API to access the other layers. In addition, there is also the processing of this data using the artificial

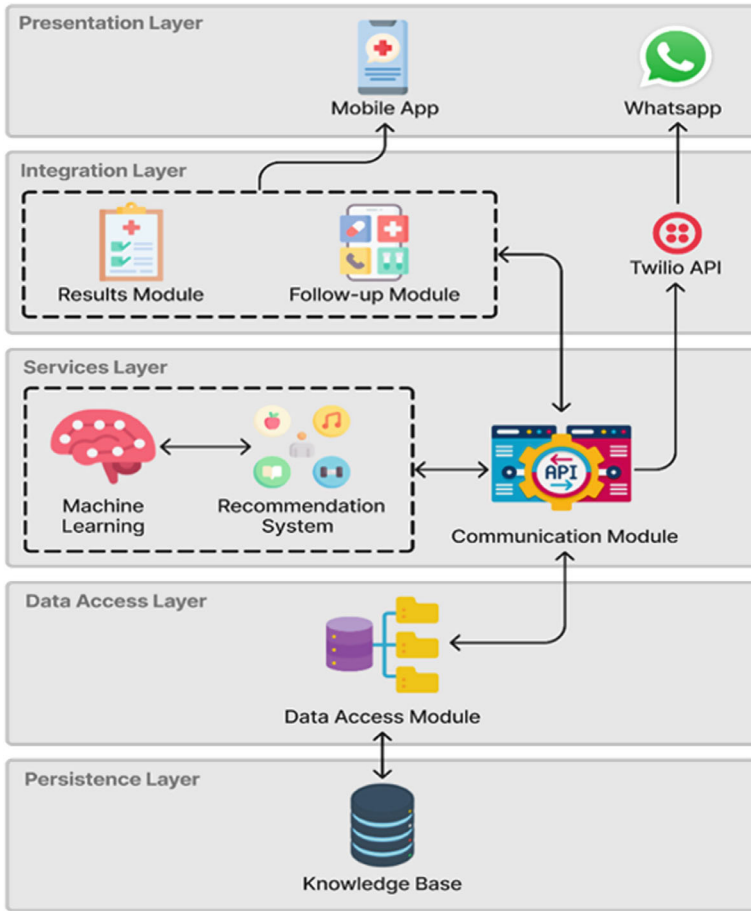


Fig. 19.1 PD-Monitor’s architecture

intelligence algorithm. Here the services that perform the necessary operations for PD-Monitor are implemented.

Data access layer: This layer is responsible for interacting with the database where the patients’ voice and freehand drawing test data are stored, as well as the recommendations that were sent and the library of messages (recommendations, notifications, and alerts). The repositories that allow reading and writing to the database are implemented here.

Persistence layer: It is responsible for storing all the data used by the PD-Monitor’s upper layers.

Regarding the design and implementation of the persistence model for the integration of the different data components, a modular strategy was followed. A distributed design was chosen to allow the scalability of the data model in the future. In this

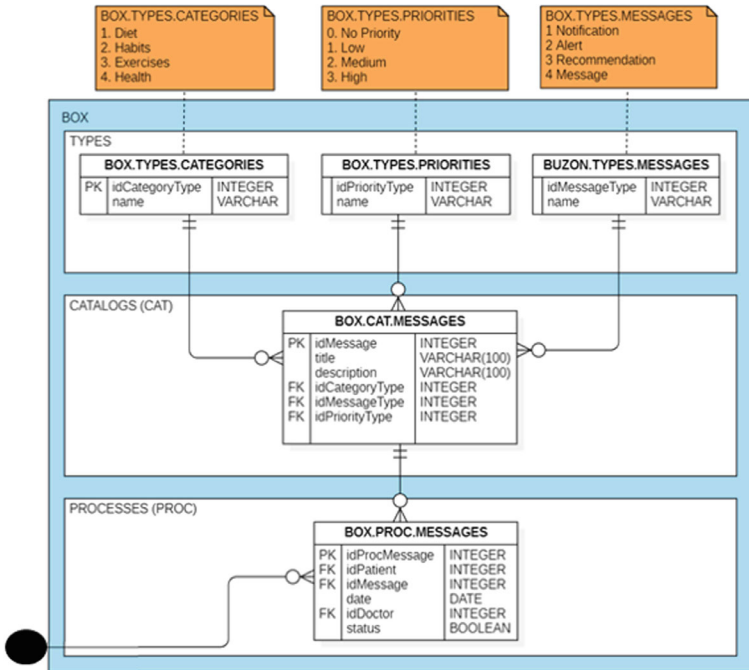


Fig. 19.2 Relational database schema: part 1

sense, the database model was divided into independent logical modules, each with its limited context schema. This was done to reuse the schema if new techniques for Parkinson’s disease monitoring needed to be included. In addition, a relational database architecture was used to facilitate data query and manipulation.

Figures 19.2 and 19.3 shows the relational schema for the database model. Figures 19.2 and 19.3 shows three different contexts, the first context in blue color (Fig. 19.2), integrates the most important entities to make the assignment of recommendations, notifications, and alerts.

The second context in green color (Fig. 19.3), was added to relate users with the role of patients, patients related to caregivers, and physicians. The third in yellow color has the entities with all the information related to the freehand drawing module.

The following section presents the voice and freehand drawing modules.

19.4 PD-Monitor Voice and Freehand Drawing Modules

PD-Monitor offers an approach to the monitoring of Parkinson’s disease symptoms. This approach aims to increase the effectiveness of medical interventions and improve the patient’s quality of life.

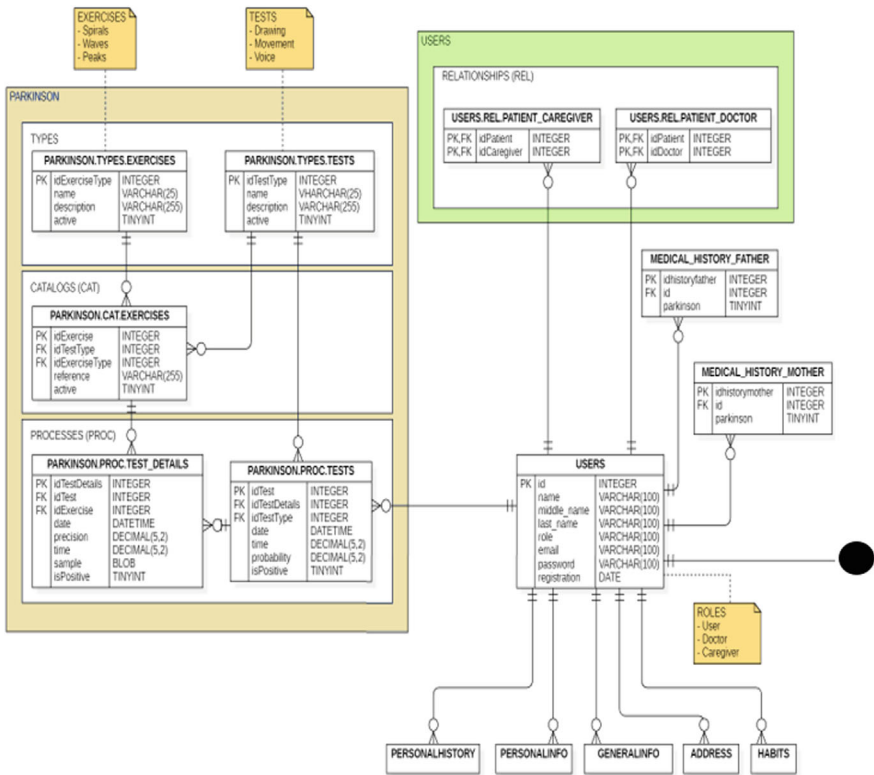


Fig. 19.3 Relational database schema: part 2

19.4.1 Voice Module

The voice module is responsible for extracting information from patients’ voice recordings. It is integrated by an external module that processes the data from the voice recordings and enters them into a database. The records are automatically evaluated each time new voice recordings are updated. Based on the consultation of this information, the recommendation module generates a recommendation, notification, or alert that is sent to the patient via the mobile application and WhatsApp.

The voice module is based on the Oxford Parkinson’s Disease Detection Dataset, created by Max Little of the University of Oxford in collaboration with the National Speech and Voice Center in Denver, Colorado. This data set contains 195 voice recordings from a total of 31 individuals, 23 of whom have PD. Also, it contains 23 attributes to determine the progression of PD. In this case, the following attributes were selected: (1) MDVP: Fo (Hz), which is the average vocal fundamental frequency; (2) MDVP: Flo (Hz), which is the minimum vocal fundamental frequency; (3) PPE, and (4) SPREAD1, that are nonlinear measures of fundamental frequency variation.

Table 19.2 Analysis of selected variables

	MDVP:Fo(Hz)	MDVP:Flo(Hz)	PPE	SPREAD1
Count	195	195	195	195
Mean	154.228641	116.324631	0.206552	-5.684397
Std	41.390065	43.521413	0.090119	1.090208
Min	88.333000	65.476000	0.044539	-7.964984
25%	117.572000	84.291000	0.137451	-6.450096
50%	148.790000	104.315000	0.194052	-5.720868
75%	182.769000	140.018500	0.252980	-5.046192
Max	260.105000	239.170000	0.527367	-2.434031

On the other hand, assignment rules were established as an essential element in the recommendations, notifications, and alerts module for patients with Parkinson's disease, to guarantee its correct functioning. For this purpose, we used the four attributes (MDVP: Fo (Hz), MDVP: Flo (Hz), PPE, and SPREAD1) and performed an analysis. This analysis is shown in Table 19.2, which includes the number of non-zero elements (count), the mean of the elements (mean), the standard deviation of the elements (std), the minimum value (min), the perceptual 25% (25%), perceptual mean (50%), perceptual 75% (75%), and the maximum value (max).

On the other hand, for the definition of the assignment rules, the dataset used to train the decision tree was cleaned. The result of the data-cleaning process reduced the dataset to only 147 records. In this case, a node in the decision tree represents a rule for a specific attribute. For classification, this rule separates values belonging to different classes. For regression, this rule separates the values to reduce the error optimally for the selected parameter criterion. The decision tree was programmed using Python and the scikit-learn library with a maximum depth of 6 and a minimum number of sheets of 1. The model was evaluated using the accuracy metric. Figure 19.4 shows the construction of the nodes and the rules generated for the decision-making of the class variable.

After generating the decision tree, an analysis was performed to determine the rules to be assigned to each type of message. Figure 19.5 shows the rules assigned to determine if a recommendation, notification, or alert is issued.

The following section describes the freehand drawing module.

19.4.2 Freehand Drawing Module

In parallel, the freehand drawing module is responsible for evaluating drawing made by the patient to detect hand tremors, a common feature of Parkinson's disease. This module also works in conjunction with an external system that calculates a

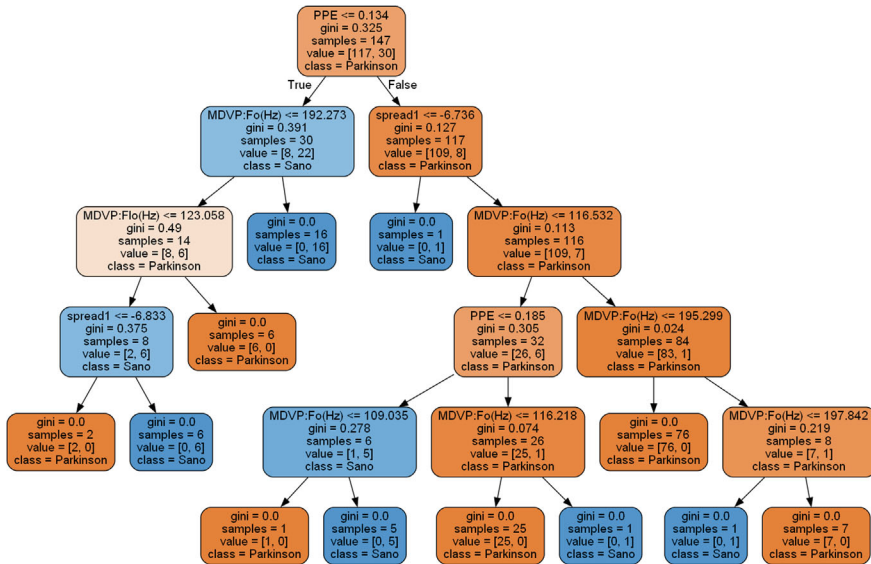


Fig. 19.4 Decision tree of the voice module

1–100 probability that the patient has Parkinson’s, based on the analysis of the free-hand drawing module. As with the voice module, once these results are obtained, recommendations, notifications, or alerts are generated and sent to the patient.

This module works with a service for the classification model and requires two datasets. One dataset for training the algorithm, and another is to perform freehand exercise classification tests for PD detection. In this case, the Parkinson’s Drawing and Parkinson Disease Spiral Drawings data sets were used. Parkinson’s Drawing contains images of spiral and wave drawings correlated in the composite drawing pen pressure analysis and represented by 2 main attributes. These attributes are determined by the X and Y axis values. This dataset contains images divided into training and test groups to compare (or reproduce) the results of the original publication presented in Refs. [27, 28]. The dataset groups are divided as follows: (a) Spirals—Training—Healthy, (b) Spirals—Test—Healthy, (c) Spirals—Training—Sick, (d) Spirals—Test—Sick, (e) Waves—Training—Healthy, (f) Waves—Test—Healthy, (g) Waves—Training—Sick, and (h) Waves—Test—Sick. Also, this repository has samples from a group of twenty-eight healthy patients and twenty-seven PD patients evaluated on the Unified PD Rating Scale [27, 28].

The Parkinson Disease Spiral Drawings, contains handwriting data from 62 people with Parkinson’s and fifteen healthy individuals distributed over a dataset of seventy-seven instances and six attributes. The attributes are defined by (a) stroke accuracy in the X and Y axes, (b) pressure exerted on the canvas, (c) grip angle, (d) execution time, and (f) test identifier [29].

For the classification model, a comparative analysis of nine classification algorithms was performed: KNN, Decision Tree, SVM, Gradient Boosting (GB),



Fig. 19.5 Rules for assigning voice-related recommendations, notifications, and alerts

LightGBM, XGBoost, Random Forest (RF), CatBoost, and AdaBoost. The results of the analysis indicated that the RF classifier obtained the best values in terms of accuracy and specificity, with an average score of 90.0 and 91.3% respectively, in the PD detection tests. The second best-performing classifier was GB, with an accuracy score of 81.8% and specificity of 82.6%. 81.8% accuracy and 82.6% specificity. Table 19.3 shows the results obtained by each of the classifiers during the analysis.

According to the results in Table 19.3, RF obtained the best performance scores in determining the classifiers, which makes it the main model to implement. On the other hand, GB obtained a good performance in terms of accuracy and is positioned as the second algorithm to implement. Additionally, the SVM algorithm was also considered for implementation. These algorithms were implemented using the Python Sickit-Learn library.

Table 19.3 Comparative results of the classification algorithms

No. test	Algorithm	Accuracy	Sensitivity	Specificity	F1 score
1	KNN	0.808	0.913	0.783	0.857
2	Decision tree	0.679	0.826	0.609	0.745
3	SVM	0.792	0.826	0.783	0.809
4	GB	0.818	0.783	0.826	0.800
5	LightGBM	0.792	0.826	0.783	0.809
6	XGBoost	0.760	0.826	0.739	0.792
7	RF	0.909	0.870	0.913	0.889
8	CatBoost	0.739	0.739	0.739	0.739
9	AdaBoost	0.739	0.739	0.739	0.739

The assignment of recommendations, notifications, and alerts using the freehand drawing module required the definition of a set of rules. The results of the spiral and wave drawing evaluations were considered to define these rules. That is the percentage range from 1 to 100 to indicate whether a person has Parkinson's disease conditions in the hands and fingers. These percentages were stored in the follows variables: (a) `res_eval_waves`, which is the result of the wave drawing evaluation; (b) `res_eval_esp`, which is the result of the spiral drawing evaluation; and (c) `avg_res_eval`, which is the average of the results of the wave and spiral drawing evaluations. However, an analysis was performed to determine that only the variable “`avg_res_eval`” would be used because this variable averages the result of the wave and spiral evaluations, which helps to reduce the probability of a bias in the results.

Figure 19.6 shows the nodes and rules constructed for decision-making of the class variable.

The rules and parameters established related to freehand drawings are shown in Fig. 19.7.

19.5 Case Study: Comprehensive Monitoring of People with Parkinson's Disease Using Voice and Freehand Drawing

This case study addresses the implementation of PD-Monitor for monitoring patients with Parkinson's disease. PD-Monitor is based on two modules working in parallel: one oriented to voice recording analysis and the other to freehand drawing analysis. Both modules seek to assign recommendations, notifications, and alerts to patients to control the symptoms of the disease. The alerts and recommendations are based on medical guidelines compiled from experts, articles, and organizations specialized in the management of Parkinson's disease. The implementation is complemented by



Fig. 19.6 Decision tree of the freehand drawing module

notifications via WhatsApp. Both modules interact with a common database through web services.

Once the results of the evaluations, whether voice or freehand drawings, are obtained, they are sent to PD-Monitor. PD-Monitor assigns a recommendation, notification, or alert, which is then sent to the user through the application for mobile devices and WhatsApp, using the Twilio API for the later.

PD-Monitor is designed to be used with a mobile device. The proposed interface is shown in Fig. 19.8, the first shows the home interface with the list of active patients (See Fig. 19.8a). In Fig. 19.8b, the mailbox in the recommendations section is shown. Figure 19.8c shows the details of the recommendations. Figure 19.8d shows the recommendations report, and Fig. 19.8e shows the recommendations report with customized filters.

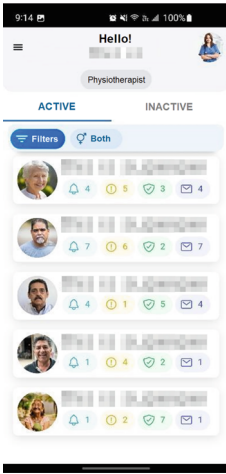
Two case studies are presented below as a proof of concept for the development of PD-Monitor.



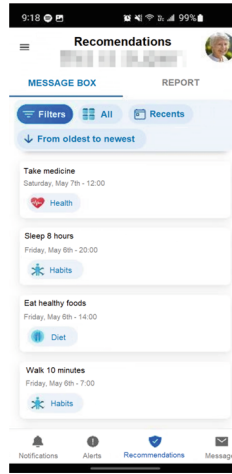
Fig. 19.7 Rules for assigning recommendations, notifications, and alerts for variables related to freehand drawing module

19.5.1 Patient Monitoring Using Voice Module

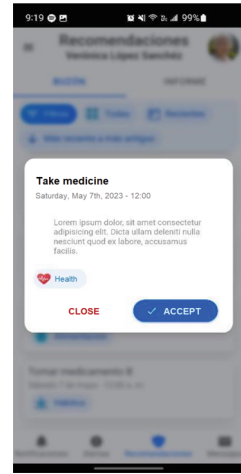
PD-Monitor was used on a 60-year-old man, who received notifications and alerts of over approximately one month. During the monitoring period, the patient interacted with the PD-Monitor application for mobile devices.



(a) Home interface - Active patients



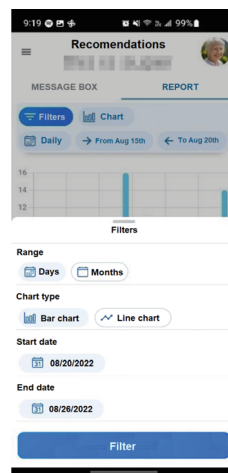
(b) Recommendations section - Mailbox subsection



(c) Recommendations section - Selected recommendation detail



(d) Recommendations section - Report subsection



(e) Recommendations section - Report subsection with custom filters

Fig. 19.8 PD-Monitor interfaces

The user was required to perform voice tests for one month (from February 7 to March 7, 2023), keeping his records through the monitoring module. During this phase, accurate data related to his biomedical parameters were collected to quantify the progress of his disease.

The tests were performed twice a week, at the end of which 9 records were obtained. The first record obtained showed normal values, subsequently four continuous records were obtained resulting in notifications that were assigned to the patient. For the remaining two weeks, the records indicated that the patient's condition started to become severe as the value variables contained critical values, therefore, the module assigned four alerts to the patient.

The records are shown in Table 19.4.

After obtaining the data through the API-REST, an analysis of the biomedical variables presented in Table 19.4 was performed. The recommendations, notifications, and alerts were generated through the decision tree presented in Fig. 19.4.

Patient records were evaluated sequentially according to dates. In this case, notifications were generated for the second, third, fourth, and fifth recordings. For the sixth, seventh, and ninth records, alerts were generated.

The second patient record concerned oculofacial praxis, then the patient was notified to continue practicing the facial muscle movement exercises to maintain good coordination when speaking (See Fig. 19.9a). The message was suggested with the aim of preserving and improving the patient's ability to articulate words accurately and fluently, which will contribute to effective communication with others. In the third record, a notification was assigned to encourage the practice of orofacial gymnastics (See Fig. 19.9a). This since a small change in the patient's way of speaking was detected. The objective of these exercises is to improve the clarity of the patient's words, which contributes to more effective communication. In fourth record a notification was assigned to encourage articulation exercises when speaking (See Fig. 19.9b) because a small change in the patient's voice was detected. By working on the mobility and control of the mouth, the patient will overcome the difficulties

Table 19.4 Records of voice tests

Id	Date	Biomedical variables of the voice		
		PPE	mdvp: fo	Spread1
1	07/02/23	116.128	109.714	-5.814103
2	10/02/23	116.329	96.871	-5.596154
3	14/02/23	197.184	174.598	-2.826923
4	17/02/23	116.128	110.379	-3.737179
5	21/02/23	204.644	89.341	-5.275641
6	24/02/23	128.024	122.115	-7.057692
7	28/02/23	186.094	177.698	-6.647436
8	03/03/23	107.256	104.621	-6.160256
9	07/03/23	116.733	111.707	-4.480769

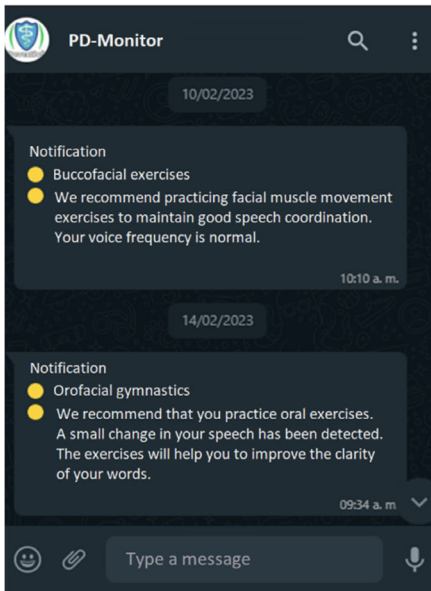
associated with the change in their voice and achieve more effective communication. In the fifth record, a notification containing reading exercises was assigned (See Fig. 19.9b) because a slight alteration in the patient's voice frequency variation was detected. Reading aloud helps to improve the clarity of pronunciation, as it allows working on word articulation and proper intonation. By practicing these exercises, patients strengthen their ability to communicate verbally, overcoming the difficulties associated with the alteration in voice frequency variation.

Regarding the alerts generated, the following results were observed. In record number 6, an alert was assigned containing mouth movement exercises including lips and tongue (See Fig. 19.9c) because a loss in speech skills was detected. The system detected a slight difference in the way the patient modulated his voice. The goal is to improve coordination and accuracy in articulating speech sounds. In record seventh, an alert was assigned and suggested the need to increase the exercises of tongue movements since it was detected an alteration in the variation of the patient's voice frequency (See Fig. 19.9c). The goal is to improve the coordination and agility of the tongue in the production of speech sounds. In eighth record, an alert was assigned and it was identified that the patient is possibly experiencing difficulties when drinking water (See Fig. 19.9d). The use of a flexible straw is beneficial for the patient, as it facilitates the process of bringing the liquid to the mouth, avoiding spills or additional difficulties due to tremors and/or movements due to possible mouth tremors. This ensures that the patient hydrates more comfortably and safely. For ninth record, an alert was assigned and it was identified that the patient is likely to experience breathing difficulties (See Fig. 19.9d). The alert to perform more controlled breathing exercises is based on the need to improve the patient's speaking. By working on breathing control, patients achieve more effective communication and increased confidence.

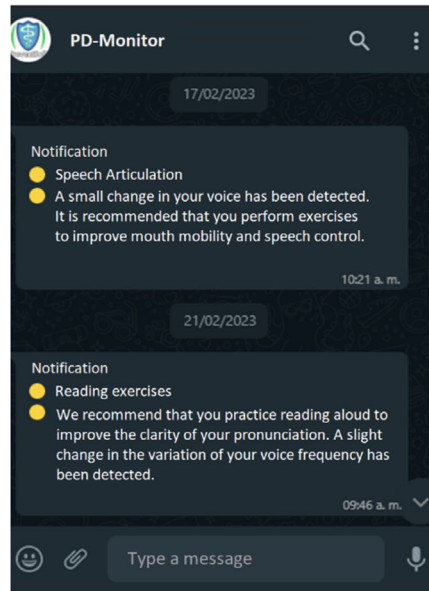
19.5.2 Patient Monitoring Using Freehand Drawing Module

The patient was asked to perform freehand drawings tests for one month (from February 14 to March 14, 2023), keeping his records through the monitoring module. This period was used to obtain the results of the evaluations performed on his drawing skills to detect anomalies in his writing skills, and general hand and finger movement disorders. These results were used to quantify the progression of the disease symptoms. The module for assigning recommendations, notifications, and alerts consulted the results of the evaluations and analyzed the values to assign a message from the collected library.

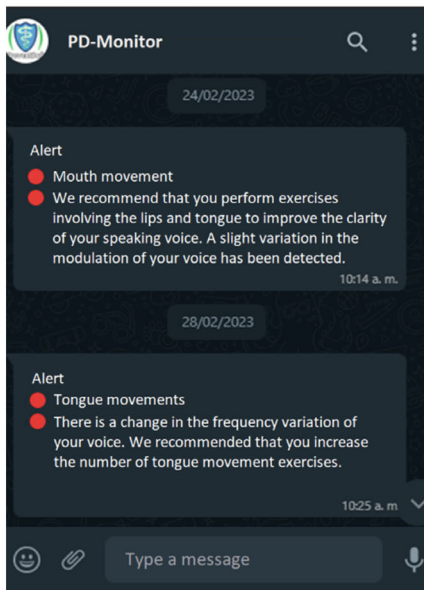
The tests were performed twice a week, obtaining 9 records at the end. During the first two weeks, the records showed non-severe results, and the system generated an alert for the patient. On the fifth record, slightly elevated results were obtained, resulting in an alert. The sixth record again lowered the percentage and an alert was assigned. The remaining three records indicated that the patient's condition



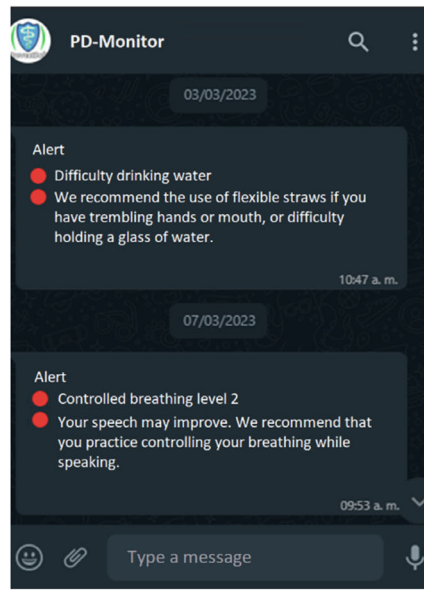
(a) Notifications assigned to registers 2 and 3



(b) Notifications assigned to registers 4 and 5



(c) Alerts assigned to registers 6 and 7



(d) Alerts assigned to registers 8 and 9

Fig. 19.9 Notifications and alerts using voice module

Table 19.5 Records of freehand drawing module

Id	Evaluation date	Percentage of drawing evaluation
1	14/02/23	46
2	17/02/23	54
3	21/02/23	53
4	24/02/23	67
5	28/02/23	75
6	03/03/23	74
7	07/03/23	89
8	10/03/23	92
9	14/03/23	96

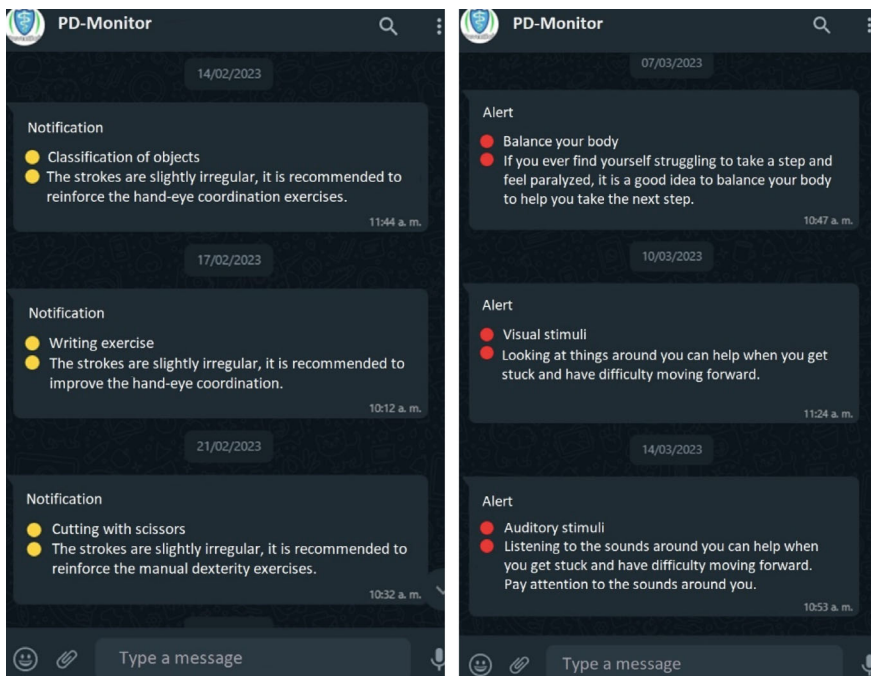
was starting to become serious since values were critical, consequently the module assigned alerts to the patient. The records are shown in Table 19.5.

According to the rules established in Sect. 4.2, recommendations, notifications, and alerts were generated. The classification of the messages was performed according to the results of the patient's freehand drawing evaluations. The rules established for Parkinson's detection by freehand drawing were applied, as shown in Fig. 19.6. Patient records were evaluated by the recommendations, notifications, and alerts module sequentially, according to the dates on which they were made. For sending notifications and alerts, messaging services were implemented through WhatsApp. This was done to guarantee an optimal user experience and maintain efficient and constant communication.

According to the results of the freehand drawing tests, recommendations, alerts, and real-time notifications were sent to the caregiver's WhatsApp account. Since the caregivers are responsible for carrying out such activities together with the patients.

In the first record, object classification was suggested and it is highlighted that the patient's drawings present slight irregularities (See Fig. 19.8a). Based on this observation, it is recommended to increase hand-eye coordination exercises. This notification is based on the need to improve the patient's motor and visual skills to effectively perform daily tasks and contribute to his independence. The patient's second record suggested performing a handwriting exercise due to slight irregularities presented in the patient's handwriting drawings (See Fig. 19.8a). Therefore, it was recommended to increase handwriting practice. The goal of this exercise is to improve the accuracy and fluency of the patient's handwriting. By increasing the frequency of handwriting, the patient has more opportunities to develop and refine his fine motor skills, which helps to improve his manual skills. In third record, it was recommended to increase the manual dexterity exercises, due to slight irregularities present in the drawing (See Fig. 19.8a). The objective of this exercise is to improve dexterity and precision in the patient's hand and finger movements. By practicing trimming with scissors, the patient strengthens the hand muscles and improves hand-eye coordination.

On the other hand, in records seven, eight and nine the following alerts were sent: in seventh record, the advice given was related to body rocking to help in situations where the patient experiences difficulty in taking a step and feels sensations of paralysis. The body rocking alert is intended to stimulate mobility and overcome feelings of paralysis (See Fig. 19.10b). In eighth record, the use of visual stimuli was suggested to help the patient in situations where he experiences difficulty in moving forward and feels that he is stuck in motion (See Fig. 19.10b). It is suggested that the patient keep his eyes open and look at things around him. The alertness to use visual stimuli is based on the idea that visual information is an important support for overcoming obstacles and maintaining the flow of movement. For the ninth recording, the use of auditory stimuli was suggested to help the patient in situations in which he/she experiences difficulties in moving forward and feels that he/she is stuck in the movement. It is suggested that the patient pay attention to surrounding sounds and use them for support. By paying attention to environmental sounds, the patient remains aware of his or her surroundings and uses them as cues to move forward (See Fig. 19.10b).



(a) Notifications assigned to registers 1, 2 and 3

(b) Alerts assigned to registers 7, 8 and 9

Fig. 19.10 Notifications and alerts using freehand drawing module

19.6 Conclusions and Future Work

Increased life expectancy and declining birth rate lead to an increase in the population of older adults, predicting an increase in cases of Parkinson's disease (PD), a progressive condition with no cure. The need to address symptoms and improve quality of life led to the development of a module of recommendations, notifications, and alerts for Parkinson's patients using Artificial Intelligence. PD-Monitor is an app which uses an algorithm to assign recommendations, notifications, and alerts, in addition to working with push notifications and messages via WhatsApp. The results demonstrate that personalized and timely messages are effective for symptom management, highlighting their difference to concerning traditional apps. As future work, we intend to implement modules for patient monitoring, which will provide more accurate and enriched information coverage. For example, using wearable devices such as smartwatches for real-time monitoring of Parkinson's patients. This will allow monitoring of the health status of patients through biometric variables. Moreover, the integration of other popular messaging platforms, such as Telegram, Messenger, or text messaging (SMS), could offer additional options to users and adapt to their communication preferences. In addition, there are plans to enrich the library of messages with more categories, such as posture, gait, and medical treatments, among others, to provide a more complete and personalized treatment.

In summary, the development of this AI-based recommendations, notifications, and alerts module showed promising results in the management of Parkinson's patients' symptoms. Its personalized and timely approach makes a significant difference compared to traditional applications. This work lays the groundwork for future improvements and extensions, to provide more comprehensive and effective support to patients in their fight against Parkinson's disease.

Acknowledgements This study segment was funded by the National Council of Humanities, Sciences, and Technologies of Mexico (CONAHCYT) and the Mexican Ministry of Public Education (SEP) through the PRODEP program. The writers also express their appreciation to the Tecnológico Nacional de México (TecNM) for supporting this research.

References

1. World Health Organization: Parkinson disease. <https://www.who.int/news-room/fact-sheets/detail/parkinson-disease>
2. Bárbara, Y., Noa, P., Jorge Lázaro, C.C., Alexander, E. del V.: La actividad física en el adulto mayor con enfermedades crónicas no transmisibles. Podium. Revista de Ciencia y Tecnología en la Cultura Física. **16** (2021)
3. Parkinson's Foundation: Understanding Parkinson's. <https://www.parkinson.org/understanding-parkinsons/statistics>
4. Fraiwan, L., Khnouf, R., Mashagbeh, A.R.: Parkinsons disease hand tremor detection system for mobile application. J. Med. Eng. Technol. **40**, 127–134 (2016). <https://doi.org/10.3109/03091902.2016.1148792>

5. Kubben, P.L., Kuijf, M.L., Ackermans, L.P.C.M., Leentjes, A.F.G., Temel, Y.: TREMOR12: an open-source mobile app for tremor quantification. *Stereotact. Funct. Neurosurg.* **94**, 182–186 (2016). <https://doi.org/10.1159/000446610>
6. De Silva, A.H.T.E., Sampath, W.H.P., Sameera, N.H.L., Amara-singhe, Y.W.R., Mitani, A.: Development of a wearable tele-monitoring system with IoT for bio-medical applications. In: 2016 IEEE 5th Global Conference on Consumer Electronics, pp. 1–2. IEEE (2016) <https://doi.org/10.1109/GCCE.2016.7800404>
7. European Foundation for Health and Exercise: Parkinson Home Exercises App. https://www.efox.nl/parkinson_app.html
8. Kuosmanen, E., Kan, V., Visuri, A., Vega, J., Nishiyama, Y., Dey, A.K., Harper, S., Ferreira, D.: Mobile-based monitoring of Parkinson’s disease. In: Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia, pp. 441–448. ACM, New York, NY, USA (2018) <https://doi.org/10.1145/3282894.3289737>
9. Memedi, M., Tshering, G., Fogelberg, M., Jusufi, I., Kolkowska, E., Klein, G.: An interface for IoT: Feeding back health-related data to Parkinson’s disease patients. *J. Sens. Actuator Netw.* **7** (2018). <https://doi.org/10.3390/jsan7010014>
10. Zhan, A., Mohan, S., Tarolli, C., Schneider, R.B., Adams, J.L., Sharma, S., Elson, M.J., Spear, K.L., Glidden, A.M., Little, M.A., Terzis, A., Ray Dorsey, E., Saria, S.: Using smartphones and machine learning to quantify Parkinson disease severity the mobile Parkinson disease score. *JAMA Neurol.* **75**, 876–880 (2018). <https://doi.org/10.1001/jamaneurol.2018.0809>
11. Zhang, H., Wang, A., Li, D., Xu, W.: DeepVoice: a voiceprint-based mobile health framework for Parkinson’s disease identification. In: 2018 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 214–217. IEEE (2018) <https://doi.org/10.1109/BHI.2018.8333407>
12. Schmitz, H., Howe, C.L., Armstrong, D.G., Subbian, V.: Leveraging mobile health applications for biomedical research and citizen science: a scoping review (2018). <https://doi.org/10.1093/jamia/ocy130>
13. Caballero Sucunza, A.: APParkinson. <https://play.google.com/store/apps/details?id=com.kiaranet.anapar&hl=es&gl=US>
14. Estévez-Martín, S., Cambronero, M.E., García-Ruiz, Y., Llana, L.: Mobile applications for people with Parkinson’s disease: a systematic search in app stores and content review. *J. Univ. Comput. Sci.* **25**, 740–761 (2019)
15. Elm, J.J., Daeschler, M., Bataille, L., Schneider, R., Amara, A., Es-pay, A.J., Afek, M., Admati, C., Teklehaimanot, A., Simuni, T.: Feasibility and utility of a clinician dashboard from wearable and mobile application Parkinson’s disease data. *NPJ Dig. Med.* **2** (2019). <https://doi.org/10.1038/s41746-019-0169-y>
16. Hu, J., Yuan, D.Z., Zhao, Q.Y., Wang, X.F., Zhang, X.T., Jiang, Q.H., Luo, H.R., Li, J., Ran, J.H., Li, J.F.: Acceptability and practicability of self-management for patients with Parkinson’s disease based on smartphone applications in China. *BMC Med. Inf. Decis. Mak.* **20** (2020). <https://doi.org/10.1186/s12911-020-01187-x>
17. Gatsios, D., Antonini, A., Gentile, G., Marcante, A., Pellicano, C., Macchiusi, L., Assogna, F., Spalletta, G., Gage, H., Touray, M., Timotijevic, L., Hodgkins, C., Chondrogiorgi, M., Rigas, G., Fotiadis, D.I., Konitsiotis, S.: Feasibility and Utility of mHealth for the remote monitoring of Parkinson disease: ancillary study of the PD_manager randomized controlled trial. *JMIR Mhealth Uhealth* **8**, e16414 (2020). <https://doi.org/10.2196/16414>
18. Orozco-Arroyave, J.R., Vásquez-Correa, J.C., Klumpp, P., Pérez-Toro, P.A., Escobar-Grisales, D., Roth, N., Ríos-Urrego, C.D., Strauss, M., Carvajal-Castaño, H.A., Bayerl, S., Castrillón-Osorio, L.R., Arias-Vergara, T., Künderle, A., López-Pabón, F.O., Parra-Gallego, L.F., Eskofier, B., Gómez-Gómez, L.F., Schuster, M., Nöth, E.: Apkinson: the smartphone application for telemonitoring Parkinson’s patients through speech, gait and hands movement. *Neuro. Dis. Manag.* **10**, 137–157 (2020). <https://doi.org/10.2217/nmt-2019-0037>
19. Habets, J., Heijmans, M., Herff, C., Simons, C., Leentjens, A.F.G., Temel, Y., Kuijf, M., Kubben, P.: Mobile health daily life monitoring for Parkinson disease: development and validation of ecological momentary assessments. *JMIR Mhealth Uhealth* **8** (2020). <https://doi.org/10.2196/15628>

20. Sage Bionetworks: App Store. <https://apps.apple.com/us/app/parkinson-mpower-2/id1375781575>
21. Ink, A.: Rhythm—Parkinson's Gait App. <https://apps.apple.com/do/app/rhythm-parkinsons-gait-app/id1593081843>
22. Fröhlich, H., Bontridder, N., Petrovska-Delacréta, D., Glaab, E., Kluge, F., Yacoubi, M. El, Marín Valero, M., Corvol, J.-C., Eskofier, B., Van Gyseghem, J.-M., Lehericy, S., Winkler, J., Klucken, J.: Leveraging the potential of digital technology for better individualized treatment of Parkinson's disease. *Front Neurol.* **13** (2022). <https://doi.org/10.3389/fneur.2022.788427>
23. Beats Medical: Beats Medical Parkinson's App. https://play.google.com/store/apps/details?id=com.beatsmedical.parkinsonsapp&hl=es_MX&gl=US
24. Del Pino, R., de Echevarría, A.O., Díez-Cirarda, M., Ustarroz-Aguirre, I., Caprino, M., Liu, J., Gand, K., Schlieter, H., Gabilondo, I., Gómez-Esteban, J.C.: Virtual coach and telerehabilitation for Parkinson's disease patients: vCare system. *J. Public Health (Germany)* (2023). <https://doi.org/10.1007/s10389-023-02082-1>
25. van den Bergh, R., Evers, L.J.W., de Vries, N.M., Silva de Lima, A.L., Bloem, B.R., Valenti, G., Meinders, M.J.: Usability and utility of a remote monitoring system to support physiotherapy for people with Parkinson's disease. *Front Neurol.* **14** (2023). <https://doi.org/10.3389/fneur.2023.1251395>
26. Evers, L.J.W., Peeters, J.M., Bloem, B.R., Meinders, M.J.: Need for personalized monitoring of Parkinson's disease: the perspectives of patients and specialized healthcare providers. *Front Neurol.* **14** (2023). <https://doi.org/10.3389/fneur.2023.1150634>
27. Scot, K.: Parkinson's Drawings. <https://www.kaggle.com/datasets/kmader/parkinsons-drawings>
28. Zham, P., Kumar, D.K., Dabnichki, P., Poosapadi Arjunan, S., Raghav, S.: Distinguishing different stages of Parkinson's disease using composite index of speed and pen-pressure of sketching a spiral. *Front Neurol.* **8** (2017). <https://doi.org/10.3389/fneur.2017.00435>
29. Team AI: Parkinson Disease Spiral Drawings. <https://www.kaggle.com/datasets/team-ai/parkinson-disease-spiral-drawings>

Chapter 20

BootstrapPLS-Fuzzy-Genetic Hybridization to Predict the Effect of the Solidarity Economy and Sustainability on Competitiveness: The Case of a Farmer's Market in Mexico



Miguel Reyna-Castillo , Alejandro Santiago , Xóchitl Barrios-del-Angel ,
Lisbeth América Brandt-García , Daniel Bucio-Gutierrez ,
Yolanda Aranda-Jiménez , and Laura Moreno-Chimely 

Abstract This study explores a new hybrid approach that combines a structural nonparametric technique and a fuzzy mathematical technique to predict the effect of social and solidarity economy and sustainable development on business competitiveness under uncertainty using the representative case of a rural farmers' market in Mexico. We used data from a survey of 100 rural entrepreneurs from a Node for Promoting the Social and Solidarity Economy (NODESS). The methodology was empirical-structural and mathematical-analytical, which is helpful for logically assessing the internal consistency of complex relationships. The measurements were

M. Reyna-Castillo

Faculty of Architecture, Design and Urbanism, The Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCyT), Autonomous University of Tamaulipas, Ciudad Victoria, Mexico
e-mail: mreyna@docentes.uat.edu.mx

A. Santiago (✉)

Faculty of Engineering Tampico, Autonomous University of Tamaulipas, Ciudad Victoria, Mexico
e-mail: aurelio.santiago@uat.edu.mx

X. Barrios-del-Angel

Tampico School of Commerce and Administration, Autonomous University of Tamaulipas, Ciudad Victoria, Mexico
e-mail: axbarrios@docentes.uat.edu.mx

L. A. Brandt-García · D. Bucio-Gutierrez · Y. Aranda-Jiménez · L. Moreno-Chimely

Faculty of Architecture, Design and Urbanism, Autonomous University of Tamaulipas, Ciudad Victoria, Mexico
e-mail: lbrandt@docentes.uat.edu.mx

D. Bucio-Gutierrez

e-mail: danielbucio@docentes.uat.edu.mx

Y. Aranda-Jiménez

e-mail: yaranda@docentes.uat.edu.mx

L. Moreno-Chimely

e-mail: lmoreno@docentes.uat.edu.mx

normalized and validated in the first phase using a non-parametric structural technique of an algorithm based on Partial Least Squares (PLS-SEM). For the predictive calculation, a Bootstrapping with $n = 10,000$ and the Inference of a Fuzzy Genetic System (FGS) were used in the second stage. The results show the functionality of our hybrid model for predicting social variables under uncertainty. The specific results show that both the attributes of the solidarity family business and those of sustainable development benefit the competitiveness of the NODESS studied, manifesting themselves in terms of advantage, growth, and competitiveness through the identification of intangible aspects. Finally, policy and social implications are provided.

Keywords PLS-Fuzzy hybridization · Prediction under uncertainty · Social solidarity economy · Sustainable development · Business competitiveness · Farmers' market in Mexico

20.1 Introduction

The SSE has emerged as a crucial paradigm in the global context, responding to the challenge of forging fairer and more equitable societies. As highlighted by Tadesse and Elsen [1], the SSE represents a necessary shift in the prevailing economic discourse, prioritizing development and social welfare over purely profit-oriented growth. This perspective is more relevant in the current global landscape, marked by crises such as the one triggered by the COVID-19 [2] pandemic. The urgency of reinventing economies, deviating from conventional models, and focusing on social welfare, solidarity, and collaboration is evident [3].

Various regions of the world have adopted SSE as a practical approach to fostering sustainable development and business competitiveness. Research in the United Kingdom, Portugal, Brazil, Senegal, Madrid, and Athens has revealed that SSE-based enterprises strengthen business performance and contribute to the sustainable development of local and international communities [2, 4]. The social relevance of the SSE is manifested in its ability to drive social change, social development, community cohesion, and individual empowerment [1]. In the Latin American context, and specifically in Mexico, initiatives such as the Nodes for the Promotion of the NODESS seek to promote local economic development in productive rural regions collectively [5, 6]. As in other regions, Latin America finds a path to sustainable environmental, economic, and social development in the SSE. The SSE represents an essential pillar for the social and economic sustainability of the region, highlighting the importance of rural family businesses in generational continuity and environmental preservation [7].

20.1.1 Social Solidarity Economy and Business Competitiveness

The SSE paradigm is evident in rural enterprises, where the close relationship between their collaborative values and their competitiveness emerges as an essential link for progress towards sustainable development [8]. Recent research has shown that SSE-based initiatives strengthen business performance in rural communities, positively impacting local and international competitiveness [4]. The connection between SSE and competitiveness lies in the ability of these initiatives to create strong networks at the community and international levels, driving sustainable development and greater competitiveness in rural environments [2].

Evidence in different regions shows that companies based on the SSE paradigm thrive by integrating into international and global networks, aligning with sustainable development [2, 4]. In this sense, the initiatives led by the SSE offer opportunities to exchange experiences and successes at the local level, strengthening the implementation of practices that boost economic competitiveness. In addition, research by Md Ajis et al. [9] highlights the importance of government information governance on successful cases of SSE in Malaysia, significantly conditioning the effectiveness of these initiatives. SSE is presented as a critical catalyst for economic development and competitiveness in rural family businesses. Active participation in collaborative networks, cooperation between companies, and a focus on social welfare are inherent aspects of SSE that strengthen the competitiveness of these companies in rural areas [10]. Thus, it is evident that adopting SSE principles contributes to social, economic, and environmental Sustainability and boosts the competitiveness of rural family businesses. In summary, the first hypothesis of this work can be deduced from the evidence shown:

H1. The attributes of the Social Solidarity Economy predict Business Competitiveness in a farmer's market in Mexico.

20.1.2 Sustainability and Business Competitiveness

The intrinsic relationship between Sustainability or Sustainable Development and rural business competitiveness is essential for long-term progress and resilience. By adopting sustainable practices, rural family businesses not only contribute to the care of the natural environment but also strengthen their competitive position in the market.

From an organizational perspective, Organizational Sustainability in rural family businesses becomes critical for their long-term success. Effective leadership and business competitiveness are intertwined, highlighting the importance of adopting sustainable practices in today's era [11]. The consideration of intangible assets, such as family and social capital, reveals that these elements influence not only knowledge and experience but also social networks and the ability to establish beneficial

relationships with various stakeholders, fundamental aspects for sustainable development [11, 12].

Other studies show the convergence of diverse theoretical currents and empirical findings that support the positive relationship between sustainable development and competitiveness in the regional agricultural sector. In line with the results of Machorro-Ramos et al. [13], it has been established that the orientation towards the pursuit of social and environmental Sustainability is positively associated with the traits of intangible capital in rural agricultural enterprises. Competitive advantage in the agricultural sector has been linked to the ability of companies to embrace sustainable and responsible practices [14, 15]. Companies that incorporate aspects of sustainable development can positively differentiate themselves in the market, contributing to their competitiveness.

Regarding specific considerations of the Agricultural Sector, rural family businesses with a strong link to the land and the community tend to prioritize community well-being and environmental Sustainability to ensure business continuity across generations [16]. Likewise, the self-perception of competitive advantage in the agricultural sector can be directly linked to practices that promote sustainable development, as has been corroborated in previous research [17]. These arguments, supported by the scientific literature cited, postulate that there is a direct and positive association between the variables of sustainable development and competitiveness in the regional agricultural sector, specifically in rural family businesses in the southern area of Tamaulipas. In summary, the second hypothesis of this work can be deduced from the evidence shown:

H2. The attributes of Sustainability predict Business Competitiveness in the case of a farmer's market in Mexico.

To support the approach to rural business competitiveness, Barney's [18] Resource-Based Vision (RBV) Theory emerges as a robust conceptual framework to elucidate the interrelationship between the sustainable SSE and the competitive performance of rural firms. The central hypothesis of RBV, according to Barney [18], posits that sustainable competitive performance is shaped primarily by firms' internal valuable resources. In this context, knowledge management aimed at identifying and exploiting intangible assets, such as solidarity cooperation in rural family businesses, is a crucial source of sustainable competitiveness performance [19–21]. Therefore, this paper aims to predict the effect of the social solidarity economy and sustainable development on business competitiveness under uncertainty using the representative case of a rural peasant market in Mexico. We used data from a survey of $n = 100$ rural entrepreneurs from a Node for Promoting the Social and Solidarity Economy, for its acronym in Spanish, NODESS. The methodology was structural and mathematical analytical, which is helpful for logically assessing the internal consistency of complex relationships. The measurements were normalized and validated in the first phase using a non-parametric structural technique of an algorithm based on Partial Least Squares (PLS-SEM). For the predictive calculation, a bootstrapping of $n = 10,000$ and the Inference of a fuzzy genetic system were used in the second stage.

This paper aims to contribute to exploring a hybrid approach between PLS-SEM and Fuzzy-Genetic algorithm, which is little used in the academic literature. This combination of methodologies offers several significant contributions to the prediction field at both theoretical and practical levels. On the one hand, applying PLS-SEM allows a robust assessment of the structural model's validity; moreover, bootstrapping validation ensures the results' reliability, especially in contexts with small samples and non-normal distributions. On the other hand, incorporating a fuzzy inference system based on genetic algorithms adds a layer of precision and adjustment to the prediction of business competitiveness. This technique allows for more accurate modeling of complex relationships between social and economic variables, especially in contexts where relationships are non-linear, and there are multiple levels of uncertainty. The methodology, results, discussion, implications, and conclusions are presented below.

20.2 Methodology

The study is based on the perspective of Wacker [22], highlighting the essential complementarity of various methodologies to address complex facets of phenomena. Two main techniques are used: empirical statistical sampling and mathematical analysis. According to Wacker [22], empirical statistical sampling verifies theoretical assumptions in large populations, while mathematical analysis explores underlying conditions.

20.2.1 *Participants*

A cross-sectional strategy was implemented for data collection at a specific time to analyze the relationships proposed in the theoretical model. The sample consisted of 100 surveys targeting rural family businesses from multiple sectors in the southern region of Tamaulipas. This choice was supported considering the relevance and significant contribution to the regional economy of entrepreneurs belonging to a NODESS. The Likert questionnaires, with a scale of 0–4, were collected in person, thus ensuring the participation of the interested parties. Table 20.1 shows the characteristics of the participants.

20.2.2 *Measurement*

The instrument was developed through the review of relevant empirical documents and the application of a Delphi methodological approach, in which representatives of rural family businesses participated. A measurement instrument consisting of

Table 20.1 Characteristics of the sample

Characteristics	Frequency	%
Gender		
Female	18	18
Male	82	82
Total	100	100
<i>Level of education</i>		
Basic	57	57
High school	14	14
Degree	26	26
Graduate	3	3
Total	100	100
<i>Marital status</i>		
Married	74	74
Single	24	24
Common-law marriage	2	2
Total	100	100
<i>No. employees</i>		
From 0 to 9	83	83
From 10 to 24	10	10
From 25 to 50	5	5
50 or more	2	2
Total	100	100
<i>Company years</i>		
1–years	17	17
6–10 years	25	25
11–20 years old	19	19
21–50 years old	33	33
51 years or older	6	6
Total	100	100

22 items distributed in three dimensions was constructed: (1) Traits of a solidary family business [5, 11, 16, 23]; (2) Sustainable Development [19, 24–26]; and (3) Competitiveness [14–16]. Initial individual and construct reliability and validity tests were performed to ensure the robustness of the instrument.

20.2.3 BootstrapPLS-Fuzzy-Genetic Hybridization

Previous studies have used hybrid empirical, statistical, and fuzzy mathematical techniques to explore the socially sustainable aspect of firms. For example, the combination of Empirical statistical and mathematical/TrIFTOPSIS in India [27] or Muñoz-Pascual et al. [28] in Portugal Structural Equations (SEM) and Qualitative Comparative Analysis of Fuzzy Sets (fsQCA). This work adopts a hybrid strategy, combining PLS-SEM and a FIS-GA. Although uncommon, this approach has proven its worth and compatibility [29, 30], especially when addressing the complexity of non-parametric and fuzzy empirical data on social aspects of Sustainability.

On the one hand, PLS-SEM is helpful for theoretical tests in social sciences focused on prediction in a robust way where conditions of structural complexity, small samples, and abnormalities in the sample distribution are shared and where it is necessary to generate scores of latent variables for subsequent analysis. This is in addition to offering high statistical power within samples in social sciences, i.e., PLS-SEM is more likely to identify significant relationships when they are present in the population [31–33]. In this phase, the measurements are normalized and validated using a non-parametric structural technique of a PLS-based algorithm. The measurement model can be expressed as follows:

$$X_i = \lambda_i \xi_i + \delta_i \tag{20.1}$$

$$Y_j = \lambda_j \eta_j + \epsilon_j \tag{20.2}$$

where:

- X_i : Exogenous Construct Indicators
- Y_j : Indicators of the endogenous construct
- λ_i, λ_j : Loadings
- ξ_i : Exogenous constructs
- η_j : Endogenous constructs
- δ_i, ϵ_j : Measurement errors

Subsequently, the structural model is evaluated for the predictive test of the hypothesis using a bootstrapping of $n = 10,000$. Bootstrapping proves to be a powerful tool to validate statistical significance in the analysis of PLS-SEM results, thanks to its flexibility in not assuming specific distributions and its ability to handle small samples, offering a robust and reliable approach for the statistical testing of predictive hypotheses in the context of PLS-SEM [32, 34]. The structural model can be expressed as follows:

$$\eta_j = \beta_{0j} + \sum_i \beta_{ij} \xi_i + \zeta_j \tag{20.3}$$

where:

- η_j : Endogenous constructs
- ξ_i : Exogenous constructs
- β_{ij} : Regression coefficients
- ζ_j : Structural errors

The following algorithm represents a bootstrap procedure with $n = 10,000$ resampling to estimate the statistic of a sample.

Algorithm 1 Bootstrap for Statistic Estimation

- 1: **Entrada:** Original data X , Number of resamplings $N = 10,000$
 - 2: **Salida:** Estimation of statistics of interest
 - 3: Initialize an array E to store calculated statistics
 - 4: **for** $i = 1$ to N **do**
 - 5: Select a Sample X^* de X with replacement
 - 6: Calculate the Interest Statistic e de X^*
 - 7: Store e en $E[i]$
 - 8: **end for**
 - 9: Calculate the Final Estimate of the Statistic (e.g., mean, median, variance) de E
-

In terms of the parametric mathematical approach, we have implemented a fuzzy inference system of Mamdani [35] for the prediction of the competitiveness of rural firms. In this system, linguistic granularity is defined in terms of low, medium, and high, using triangular shapes for fuzzy sets, which were also used to explore the predictive effects of social variables in the Latin context [29]. The equation for calculating the triangular membership function is based on Pedrycz's [36] formulation [36], providing a more precise and rigorous mathematical structure. The aim is to establish a knowledge base that minimizes the discrepancy between the crisp output of the FIS and the numerical values of the training set. After analyzing the relevant characteristics, we selected the inputs associated with the Supportive Family Business and the attributes of Sustainability. These, in turn, act as independent variables, while competitiveness in firms in a farmer's market in Mexico is considered the dependent (consequential) variable. This approach seeks more accurate and adjusted modeling by applying solid mathematical principles, thus contributing to the robustness and reliability of the FIS-GA application.

The inputs and outputs of the Fuzzy Inference System (FIS) inputs and outputs are carefully validated using the PLS-SEM approach, ensuring that the independent and dependent variables for the FIS are significant (an $R^2 \geq 0.5$). The FIS works using two inputs (independent variables), the constructs of Traits of Social Solidarity Economy and Sustainable development, and one output (dependent variable), the construct of Competitiveness of the farmer's market. The input and output constructs are normalized using their maximum and minimum values to simplify the fuzzy sets computations. The fuzzy sets are triangular membership functions with three sets (low, mid, and high) and use standard parameters in the literature; the triangular membership function parameters are equal as in [37, 38]. In summary, the crisp search space of the fuzzy knowledge on the antecedents and consequent is in Table 20.2.

Table 20.2 Crisp inputs and outputs of the FIS

AND antecedents		Consequent
Traits social-solidarity economy	sustainable development	Competitiveness
Input ∈ [0, 1]	Input ∈ [0, 1]	Output ∈ [0, 1]

Table 20.3 Parameters for the genetic algorithm

	GA
<i>Crossover:</i>	Single-point
<i>Crossover probability:</i>	$p_c = 1.0$
<i>Mutation:</i>	Boundary
<i>Mutation probability:</i>	$p_m = 0.5$
<i>Population size:</i>	100
<i>Generations:</i>	250

In this work, we define the fuzzy operations for the AND operator as in other works using the min operator in Eq. (20.4) The min operator computes the minimum membership of the two inputs; the above membership modifies the consequent fuzzy output. Finally, the fuzzy output is converted to a crisp value using the centroid computation.

$$\mu_{A \cap B} = \min(\mu_A, \mu_b) \tag{20.4}$$

Given the nature of the discrete search levels (low, mid, and high) and exponential search space, genetic algorithms (GAs) are suitable for optimizing the FIS and reducing prediction errors. Thus, we use a regular GA with a single-point crossover, a random boundary mutation, with a probability of crossover p_c of 100%, mutation probability of p_m of 5%, a population size of 100, and 250 generations. The parameters of the GA are in Table 20.3.

The above describes a hybridization between the PLS-SEM and Genetic Algorithm, our core approach in this work. However, the fitness function is relevant to reproduce the experimental results. The fitness function of this work is the Mean Absolute Error (MAE) as in Eq. (20.5)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y - \hat{y}| \tag{20.5}$$

where y is the desired output and \hat{y} is the produced output of the FIS. The database and source code used in this manuscript can be accessed at <https://github.com/Reyna-Castillo/NODESS-FISGA> (accessed on May 18, 2024).

Table 20.4 Measurement quality criteria (PLS algorithm)

Criterion	Condition	Threshold
(1) The reliability of the indicators (λ)	\geq	0.400
(2) The internal consistency of the construct (CR)	\geq	0.700
(3) The reliability of the construct (ρA)	\geq	0.708
(4) The convergent validity of the construct (AVE)	\geq	0.500
(5) Discriminant validity between constructs ($HTMT$)	\leq	0.900
(6) The variance explained (R^2)	\geq	0.100
(7) Standardized path coefficients (β)	\geq	0.200

20.3 Results

20.3.1 PLS-SEM Predictive Algorithm

The first phase of our hybrid technique involved the evaluation and validity of the reflective measurement model (Type A) using the PLS [31] Algorithm. Using the statistical program SmartPLS v. 4.1.0.0 [39], the quality criteria of the model were assessed (Table 20.2).

As seen in Table 20.3, external loads were obtained with expected values ranging from $\lambda = \geq 0.400$. Therefore, the items are relevant.

Likewise, the internal consistency and convergent validity reliability in the constructs required by $\rho A \geq 0.708$, $RC \geq 0.700$, and $AVE \geq 0.500$ presented required quality values. Regarding discriminant validity, tested with the test $HTMT$, it confirms that the constructs differ, with values that remain below the $HTMT$ limit of ≤ 0.900 [40] (Table 20.4).

Table 20.3 Regarding the power of the predictive capacity, it was also relevant, on the one hand, with an explained variance of R^2 of 0.636, exceeding the required threshold of $R^2 \leq 0.100$ and with standardized coefficients β of 0.488 and 0.393, according to the required parameter of $\beta \leq 0.200$ [31]. Since the structural model does not present collinearity problems, with variance inflation factor (VIF) values below 3.3 [31], we proceed to show the results of the hypothesis tests.

After the potential predictive approach, the statistical program SmartPLS v. 4.1.0.0 [39] offers Bootstrapping as an essential tool for evaluating hypotheses based on PLS-SEM. This tool enables the evaluation of the statistical significance of key coefficients such as path, Cronbach's alpha, HTMT, and R^2 values. This non-parametric procedure made it possible to generate subsamples through the random extraction (with replacement) of observations from the original dataset, repeating the process about 10,000 times.

Table 20.5 specifies the criteria used. Using this method, 95% confidence intervals were derived, standard errors were provided, and the importance of each estimate was assessed, ensuring solid and replicable results in the research.

Table 20.5 Reliability and validity of indicators and constructs

Construct/indicator	λ	α	ρ	RC	AVE
1. Traits social solidarity economy (β , 0.488)	–	0.898	0.901	0.917	0.551
1.1. Strong roots with the community	0.714				
1.2. Family values are very important	0.697				
1.3. Important cohesion between partners	0.769				
1.4. Close relationship between stakeholders	0.760				
1.5. Great influence of the founder's opinion	0.760				
1.8. Strong family-business bond	0.652				
1.9. Community is vitally important to activities	0.788				
1.10. Family-Community-Centered Workforce	0.753				
1.11. Collective well-being as a mission	0.778				
2. Sustainable development (β , 0.393)	–	0.807	0.832	0.868	0.572
2.1. Positively impacting the local community	0.685				
2.2. Sustainable development is of great importance	0.596				
2.3. Community economic development is of great importance	0.772				
2.4. Mutual cooperation with the community is of great importance	0.866				
2.5. Improving and not harming the environment is very important	0.830				
3. Competitiveness of the farmers' market (R^2 , 0.636)	–	0.843	0.848	0.884	0.561
3.1. Adaptation to changes in the environment	0.668				
3.2. Intangibles to outperform the competition	0.785				
3.3. Competitive economic performance	0.757				
3.4. Growing local job creation	0.722				
3.5. We maintain an advantage over competitors	0.794				
3.6. We have sustained growth in the company	0.763				

Table 20.6 Discriminant validity between constructs (HTMT)

Constructs	1	2	3
1. Traits social solidarity economy	–	–	–
2. Sustainable development	0.745	–	–
3. Competitiveness of the farmers' market	0.839	0.840	–

The test contained in Table 20.6 shows the significant, direct, and positive results of the prediction hypotheses tested. Where β represents the path coefficients and the estimated relationships in the structural model and correspond to standardized beta coefficients in a regression analysis [32]. Their significance level is based on the parameters of t value ≥ 3.310 ($p \leq 0$)***, ≥ 2.586 ($p \geq 0,01$)**, ≤ 1.965 ($p \geq 0.05$). Since the signs of the proposed hypotheses are positive, consistent with the signs

Table 20.7 Bootstrapping criteria

Criteria	Configuration
Complexity	With the most important thing (faster)
Samples	10000
Confidence interval method	Percentil bootstrap
Level of significance	0.05
Parallel processing	Yes
Save results per sample	No
Seed	Fixed Seed
Type of test	One tail

Table 20.8 Results for hypothesis testing of the model

Hi	Interaction	β	M	STDEV	Statistical <i>t</i>	<i>p</i>	Supported
1	Traits social solidarity economy— Competitiveness of the farmers' market	0.488	0.492	0.095	5.165	0.000	Yes
2	Sustainable development— Competitiveness of the farmers' market	0.393	0.396	0.094	4.179	0.000	Yes

of the resulting path coefficients, the latter being within the acceptable threshold ranging from -1 to $+1$, the values allow the acceptance of hypotheses with a positive and significant relationship, according to the criterion proposed by Chin [41].

20.3.2 Fuzzy Genetic Algorithm

By combinatorics, for two inputs with three possible levels (granularity), gives a total of 9 AND rules to be adjusted. Given the number of rules equals 9 with three granularity levels (low, mid, high), we have a solution space equal to $27^3=19683$, in other words 27 parameters to be adjusted to one specific granularity level, low, mid, or high. Space that would be very difficult to compute systematically or manually, so the use of genetic algorithms is ideal in this situation. We avoid human biases and local optima by resorting to GA-based artificial intelligence.

It is essential to remember that genetic algorithms (GA) are stochastic, and the FIS are deterministic. Since FIS is deterministic, we can calculate the exact value of the relative error produced by the best fuzzy knowledge base in the training phase against the test data. The best fuzzy rules found by the GA are listed in Table 20.7.

Table 20.9 The 9 rules of the fuzzy knowledge base

Rules	AND antecedent	AND antecedent	Consequent
	1. Traits solidarity family business	2. Intellectual property culture	3. Sustainability
1	Low	Low	Low
2	Low	Mid	High
3	Low	High	Mid
4	Mid	Low	Mid
5	Mid	Mid	Mid
6	Mid	High	High
7	High	Low	Mid
8	High	Mid	High
9	High	High	High

Percentage error: 0.1817218212386449

Using the rules, we calculate a percentage error of 18.17%, giving a classification accuracy of 81.82% (Tables 20.8 and 20.9).

However, this absolute error could be the sum of slight differences between values, and for some decision-makers, these differences would be negligible. Therefore, we have a predictive power in Competitiveness of more than 80%, based on aspects related to the Social Solidarity Economy and Sustainable development of the farmers’ market. Our proposed FIS is a powerful tool for decision-makers and policy-makers in organizations.

20.4 Discussion

The objective of this work was to predict under uncertainty the effect of the Social Solidarity Economy and Sustainable Development on Business Competitiveness using the representative case of a rural peasant market in Mexico. On the one hand, (i) to determine the predictive potential of the Social Solidarity Economy traits on Competitiveness, and (ii) to determine the predictive potential of the Sustainable Development traits on Business Competitiveness. To meet our objective, we use a hybridization based on two techniques of a predictive nature: PLS-SEM and FIS-GA.

The overall results of the predictive tests based on PLS-SEM show a valid structural model with significant predictive potential with an explained variance of R^2 0.636 (Fig. 20.1). This result implies that, within the sample analyzed, the indicators of the variables of Aspects of the Solidarity Family Business and Sustainable Development explain business competitiveness above 60%. On the other hand, when determining the FIS-GA performance and calculating the exact value of the relative error produced by the best fuzzy knowledge base, a Percentage Error of 18.17% was obtained, implying a classification accuracy of almost 82%. These results confirm the

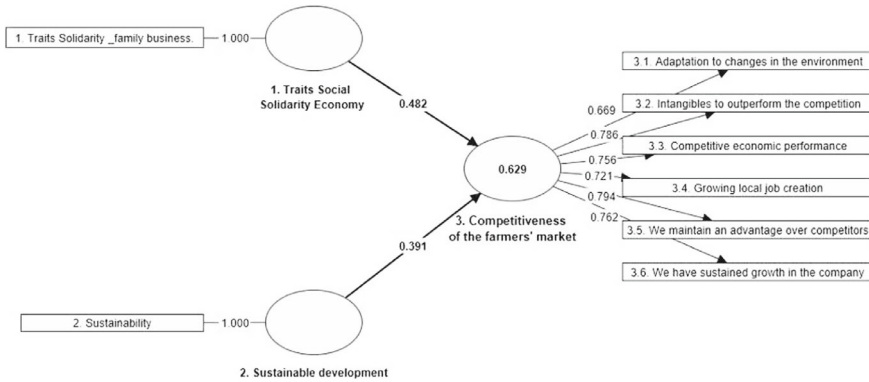


Fig. 20.1 PLS-SEM structural nomogram of specific relationships

predictive assumptions from non-parametric and mathematical statistics, allowing us to affirm that the collaborative features of rural family businesses and the aspects of sustainable development have predictive power on Competitiveness within the representative case of a rural peasant market in Mexico.

In light of the theoretical hypothesis of Barney’s [18] RBV Theory, the results support his central assumption that sustainable competitive performance depends primarily on firms’ internal valuable resources. Considering the Aspects of the Solidarity Family Business and Sustainable Development as internal resources, these are valuable capabilities that generate Business Competitiveness using the representative case of a rural peasant market in Mexico. The findings are in line with previous studies where, from the perspective of the RBV, they demonstrated that solidarity cooperation in rural family businesses is configured as a crucial source of sustainable competitiveness performance [19–21].

20.4.1 Social Solidarity Economy and Business Competitiveness

The potential and specific significance of the relationship between Aspects of the Solidarity Family Business and Business Competitiveness was explored. The hypothesis *Hi1* that states “The attributes of the Social Solidarity Economy predict Business Competitiveness in the case of a peasant market in Mexico” was supported, where significant predictive coefficients were obtained from the PLS-SEM perspective ($\beta = 0.488***$, $p = 0.000$, $R^2 = 0.636$). From a theoretical perspective, aspects of the solidarity family business are valuable resources and a factor of business competitiveness. These results are consistent with previous studies that have found the collaborative economy as an alternative for sustainable development [2, 4]; Md Ajis et al. [9] found that the SSE develops better economic competitiveness initiatives in Malaysian cases,

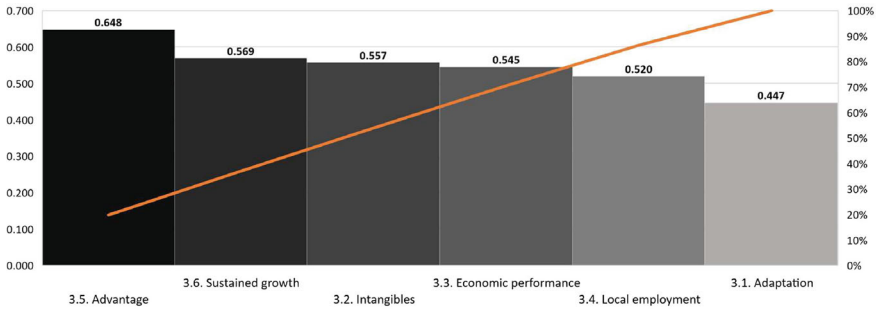


Fig. 20.2 Pareto de correlaciones PLS: Empresa solidaria versus Aspectos de Competitividad

but noted that the government’s information governance significantly conditions such initiatives. Sustainable competitive advantage makes it essential to build innovation networks in the context of collaborative economies [21].

Regarding the specific results, the correlations obtained using the PLS-SEM Algorithm show the most vital relationships of the variable Solidary Family Business on the specific items of the dependent variable of Competitiveness. Figure 20.2 shows the correlation weights of the aspects of the variable “Traits of a solidary family business.” The most affected item of Competitiveness was “3.5. We maintain an advantage over competitors” ($p = 0.648$). Figure 20.2 also shows that the second item of the Competitiveness variable most affected was “3.6. We have a sustained growth of the company” ($p = 0.569$). The specific results demonstrate that entrepreneurship under the social values of collaboration and cohesion offers a differentiator that promotes Competitiveness and growth within the context of the representative cases of a rural farmer’s market in Mexico. The results statistically demonstrate what the cases of the research by Esteves et al. [4] found in the United Kingdom, where companies based on values of “care for the Earth,” “care for people,” and “fair participation” can promote highly productive firms from a financial point of view. Alternatively, in Portugal, how the development of coexistence strategies combined with the change in consumption habits can be aligned with the productive capacity of a regional network [4]. Negash et al. [8] also verified, using the diffuse Delphi method, the driving role of collaboration and innovation in values in Economic Sustainability.

20.4.2 Sustainability and Business Competitiveness

The potential and specific significance of the relationship between Sustainable Development and Business Competitiveness was explored. The hypothesis $H21$ that states “The attributes of Sustainability predict Business Competitiveness in the case of a peasant market in Mexico” was supported, where significant predictive coefficients were obtained from the PLS-SEM perspective ($\beta = 0.393***$, $p = 0.000$, $R^2=0.636$).

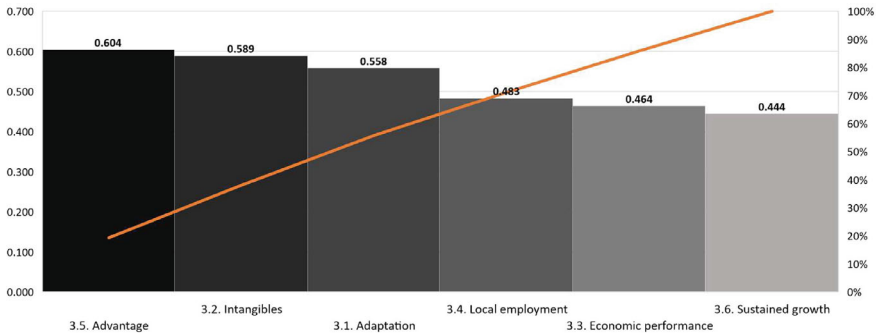


Fig. 20.3 Pareto of PLS correlations: sustainability versus competitiveness aspects

From a theoretical perspective, sustainable development is a valuable resource and a factor in business competitiveness. These results are consistent with previous studies that have found the collaborative economy as an alternative for sustainable development [11, 12].

Regarding the specific results, the correlations obtained using the PLS-SEM Algorithm show the most vital relationships of the Sustainable Development variable on the specific items of the dependent variable of Competitiveness, as shown in Fig. 20.3, which shows the correlation weights. The Competitiveness item most affected was “3.5. We maintain an advantage over competitors” ($p = 0.604$). Figure 20.3 also shows that the second item of the Competitiveness variable most affected was “3.2. We have intangibles to outperform the competition” ($p = 0.589$). The specific results demonstrate that entrepreneurship under the social values of collaboration and cohesion offers a differentiator that promotes Competitiveness and captures intangibles within the context of the representative cases of a rural farmer’s market in Mexico. The results are aligned with findings in emerging economic contexts such as Brazil and Senegal, where sustainable development goals are highly associated with sustained performance, with the intangible of the association being a factor of development between smallholder farmers and workers (Whether or not they own hectares of land), thus achieving the satisfaction of local needs [4]. Alternatively, how intangible aspects such as digitalization enhance the sustainable performance of social enterprises based on the social solidarity economy [42].

20.5 Theoretical Implications

The predictive capacity of the Social Solidarity Economy and Sustainable Development on business competitiveness in a rural peasant market in Mexico is substantial, revealed through a hybrid approach that combines PLS-SEM and the FIS-GA. The results of the hybrid technique indicate an explained variance (R^2) between 60% and

80%, demonstrating that the aspects of the solidary family business and sustainable development significantly explain competitiveness in a significant number of the cases analyzed. This conclusion supports the hypothesis that these attributes have predictive power in the specific context of a rural farmer's market under conditions of uncertainty.

20.6 Policy Implications

The results suggest important implications for policy actions at the state and federal levels in strengthening peasant-rural markets in Mexico. Promoting and supporting the Social Solidarity Economy and Sustainable Development are positioned as effective strategies to boost the Competitiveness of rural family businesses. Policies should focus on fostering collaboration, social cohesion, and sustainable practices, recognizing the crucial role of these elements as valuable internal resources. In addition, the importance of considering information governance and the role of government in maximizing Social Solidarity Economy initiatives is highlighted.

The specific results reveal that the traits of the solidary family business exert a significant influence on business competitiveness, mainly in favoring maintaining an advantage over competitors and ensuring sustained growth within the context of a rural peasant market in Mexico. These specific correlations reinforce the importance of promoting social values of collaboration and cohesion in rural family businesses. Specifically, the Ministry of Economy in Mexico, through its program aimed at entrepreneurs linked to a Node for the NODESS, could strengthen these companies through strategies that promote competitive advantage and sustainable growth, aligning with the specific results of our research.

20.7 Managerial and Entrepreneurial Implications

This study's findings have significant implications for management and entrepreneurship, especially in the context of farmers' markets in Mexico. First, the predictive capacity demonstrated by the Social Solidarity Economy and Sustainable Development on business competitiveness suggests that rural businesses can benefit from promoting collaborative values, social cohesion, and sustainable practices. This implies that business strategies that foster family solidarity, cooperation with the community, and care for the environment can lead to better competitive performance in rural markets.

Moreover, the identification of specific aspects of solidarity economy and sustainable development that influence business competitiveness, such as maintaining an advantage over competitors and ensuring sustained growth, provides actionable guidance for rural enterprises. These findings underscore that enterprises can enhance

their competitiveness by focusing on strengthening their ties with the community, improving their adaptability and focusing on sustainable growth.

From a managerial perspective, these results reinforce the notion that rural enterprises can secure substantial competitive advantages by adopting business practices that promote solidarity, sustainability, and cooperation. Therefore, managers and entrepreneurs in Mexico's farmers' markets are encouraged to consider integrating these values and practices into their business and operational strategies to bolster their competitive position and ensure long-term sustainable growth.

In summary, this study highlights the importance of the social solidarity economy and sustainable development in business competitiveness in Mexico's farmers' markets. It offers managers and entrepreneurs practical guidance on improving their competitive performance by adopting business practices focused on solidarity, sustainability, and community collaboration.

20.8 Conclusions

The study explored the utility of a predictive hybrid Technique that enables a collaboration loop between fuzzy variables from social sciences and mathematical techniques, allowing for robust handling of anomalous data. Utilizing a hybrid methodology that combines PLS-SEM and a Fuzzy Inference System based on Genetic Algorithms, it has been found that these aspects significantly influence the competitiveness of rural enterprises. The predictive capacity of the social solidarity economy and sustainable development in business competitiveness in Mexican rural markets has been demonstrated in the representative case.

These findings contribute to theoretical understanding and support the Resource-Based Vision Theory while providing concrete evidence to inform government policies and economic development strategies to benefit the Nodes for the NODESS. In addition, managerial and business implications highlight the importance of promoting business practices that foster solidarity, sustainability, and cooperation to improve competitiveness and ensure sustainable growth in Mexico's peasant markets. Overall, this study provides an avenue for discussing a new technique and a comprehensive perspective on how rural enterprises can improve their competitive performance by adopting values and practices that promote collaboration and sustainability in their environment.

Acknowledgements Funding is gratefully acknowledged to CONAHCYT under the Postdoctoral Fellowships for Mexico (2021-1) program with application number 2264959.

References

1. Tadesse, M.E., Elsen, S.: The social solidarity economy and the hull-house tradition of social work: keys for unlocking the potential of social work for sustainable social development. *Soc. Sci.* **12**, 189 (2023)
2. Arampatzi, A. Social solidarity economy and urban commoning in post-crisis contexts: Madrid and Athens in a comparative perspective. *J. Urban Aff.* **44**, 1375–1390 (2022)
3. Binns-Hernández, H.: La economía social solidaria como agente de activación económica en tiempos de pandemia del covid-19. *Revista Tecnología en Marcha* (2022)
4. Esteves, A.M., Genus, A., Henfrey, T., Penha-Lopes, G., East, M.: Sustainable entrepreneurship and the sustainable development goals: community-led initiatives, the social solidarity economy and commons ecologies. *Business Strat. Environ.* **30**, 1423–1435 (2021)
5. INAES.: *Nodos de impulso a la economía social y solidaria nodess* (2023)
6. Martínez-Gutiérrez, R., Solís-Quinteros, M.M., Sánchez-Hurtado, C., Carey-Raygoza, C.E.: Challenges for an Observatory of the 2030 Goals, SDG and Social Economy, in Northern Mexico (2021)
7. Maciel, A.S., De la Garza Ramos, M.A.I., Aguilar, J.L.E., Reyna, J.M.S.M.: La sucesión de la empresa familiar; una aproximación teórica. *Cuadernos de Administración.* **31**, 105–137 (2019)
8. Negash, Y.T., Hassan, A.M., Lim, M.K., Tseng, M.-L.: Sustainable supply chain finance enablers under disruption: the causal effect of collaboration value innovation on sustainability performance. *Int. J. Logis. Res. Appl.*, 1–25 (2024)
9. Ajis, A.F.M., Jali, J.M., Baharin, S.H., Johari, M.K., Sani, A.: Information governance derivatives of social solidarity economy initiatives, pp. 29–33. *IEEE* (2019)
10. Parrilla-González, Juan, Ortega-Alonso, Diego: Sustainable development goals in the andalusian olive oil cooperative sector: heritage, innovation, gender perspective and sustainability. *New. Medit.* **21** (2022)
11. Kim, D., Cho, W., Allen, B.: Sustainability of social economy organizations (SEOs): an analysis of the conditions for surviving and thriving. *Social Sci. J.*, 1–17 (2020)
12. Li, J., Zuo, Q., Yu, L., Ma, J.: A multi-dimensional relationship assessment framework for water resources, social economy and eco-environment: a case study of china's largest arid zone. *Environ. Impact Asses. Rev.* **102**, 107221 (2023)
13. Ramos, F.M., Salgado, P.M., Arturo, D., Ortiz, C., Vanessa, M., Ortiz, R.: Influencia del capital relacional en el desempeño organizacional de las instituciones de educación superior tecnológica. *Innovar.* **26**, 35–50 (2016)
14. Baldazo-Molotla, F., Marcelino-Aranda, M., Roberto, L., Aguirre, D., Camacho, A.D.: From the peasant economy to the management of a family microenterprise: a case study in tepotzotlán, state of Mexico. *Text.*, 107–130 (2020)
15. Bikefe, G., Zubairu, U., Araga, S., Maitala, F., Ediuku, E., Anyebe, D.: Corporate social responsibility (CSR) by small and medium enterprises (SMEs): a systematic review. *Small Bus. Int. Rev.* **4**, 16–33 (2020)
16. San-Martín, Durán, J.: *Radiografía de la empresa familiar en México* (2017)
17. Luque-Vílchez, M., Rodríguez-Gutiérrez, P.: *Internacionalización y supervivencia de la pyme agroalimentaria del sur de España. Informacion Tecnica Economica Agraria* (2021)
18. Barney, J.: Firm resources and sustained competitive advantage. *J. Manag.* **17**, 99–120 (1991)
19. Barrios-DelÁngel, Ana-Xóchitl., Reyna-Castillo, Miguel, Bucio-Gutiérrez, Daniel: Activos intangibles y la competitividad sostenible en las empresas familiares. *Revista de Ciencias Sociales* **28**, 94–109 (2022)
20. Siltaloppi, J., Ballardini, R.M.: Promoting systemic collaboration for sustainable innovation through intellectual property rights. *J. Co-operative Organ. Manag.* **11**, 100200 (2023)
21. Sun, Y., Wang, T., Gu, X.: A sustainable development perspective on cooperative culture, knowledge flow, and innovation network governance performance. *Sustain.* **11**, 6126 (2019)
22. Wacker, J.G.: A definition of theory: research guidelines for different theory-building research methods in operations management. *J. Operat. Manag.* **16** (1998)

23. Vélez, L.E.M.: Aportes conceptuales de la economía social y solidaria a la economía circular. *Cuadernos de Administración* **37**, e5010824 (2021)
24. Joo, J-H.: A mediating role of social capital between corporate social responsibility and corporate reputation: perception of local university on CSR of KHNP. *J. Indust. Distrib. Bus.* **11**, 63–71 (2020)
25. Sallah, C.A., Caesar, L.D.: Intangible resources and the growth of women businesses. *J. Entrepreneurship Emerg. Econ.* **12**, 329–355 (2020)
26. do Nascimento, F.S., Calle-Collado, A., Benito, R.M.: Economía social y solidaria y agroecología en cooperativas de agricultura familiar en brasil como forma de desarrollo de una agricultura sostenible. *CIRIEC-España, revista de economía pública, social y cooperativa*, p. 189 (2020)
27. Majumdar, A., Jeevaraj, S., Kaliyan, M., Agrawal, R.: Selection of resilient suppliers in manufacturing industries post-covid-19: implications for economic and social sustainability in emerging economies. *Int. J. Emerg. Mark.* (2021)
28. Muñoz-Pascual, L., Galende, J., Curado, C.: Human resource management contributions to knowledge sharing for a sustainability-oriented performance: a mixed methods approach. *Sustain.* **12**, 161 (2019)
29. Reyna-Castillo, M., Santiago, A., Martínez, S.I., Rocha, J.A.C.: Social sustainability and resilience in supply chains of Latin America on Covid-19 times: classification using evolutionary fuzzy knowledge. *Math.* **10**(2371) (2022)
30. Ringle, C.M., Sarstedt, M., Schlittgen, R.: Genetic algorithm segmentation in partial least squares structural equation modeling. *OR Spect.* **36**, 251–276 (2014)
31. Hair, J.F., Risher, J.J., Sarstedt, M., Ringle, C.M.: When to use and how to report the results of PLS-SEM. *European Bus. Rev.* **31**, 2–24 (2019)
32. Hair, J.F., Jr., Tomas, G., Hult, M., Ringle, C.M., Sarstedt, M., Apraiz, J.C., Carrión, G.A.C., Roldán, J.L.: *Manual de Partial Least Squares Structural Equation Modeling (PLS-SEM) (Segunda Edición)*. *OmnSci.* **7** (2019)
33. Hair, J.F., Jr., Ringle, C.M., Gudergan, S.P., Apraiz, J.C., Carrión, G.A.C., Roldán, J.L.: *Manual avanzado de Partial Least Squares Structural Equation Modeling (PLS-SEM)*. *OmnSci.* **7** (2021)
34. Becker, J.M., Cheah, J.-H., Gholamzade, R., Ringle, C.M., Sarstedt, M.: PLS-SEMS most wanted guidance. *Int. J. Contemp. Hosp. Manag.* **35**, 321–346 (2023)
35. Cordón, O.: A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: designing interpretable genetic fuzzy systems. *Int. J. Approx. Reason.* **52**, 894–913 (2011)
36. Pedrycz, Witold: Why triangular membership functions? *Fuzzy Sets Syst.* **64**(1), 21–30 (1994)
37. Santiago, Alejandro, Dorronsoro, Bernabé, Nebro, Antonio J., Durillo, Juan J., Castillo, Oscar, Fraire, Héctor. J.: A novel multi-objective evolutionary algorithm with fuzzy logic based adaptive selection of operators: fame. *Info. Sci.* **471**, 233–251 (2019)
38. Santiago, Alejandro, Dorronsoro, Bernabé, Fraire, Héctor. J., Ruiz, Patricia: Micro-genetic algorithm with fuzzy selection of operators for multi-objective optimization: μ fame. *Swarm Evolut. Comput.* **61**, 100818 (2021)
39. Ringle, C.M., Wende, S., Becker, J.-M.: *Smartpls 4* (2022). <http://www.smartpls.com>
40. Henseler, C.M.R., Sarstedt, M.: A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Market. Sci.* **43** (2015)
41. Chin, W.W.: *The Partial Least Squares Approach to Structural Equation Modeling*, vol. 295, 2nd edn, 295–336. Psychology Press (1998)
42. Ridley-Duff, R., Bull, M.: Common pool resource institutions: the rise of internet platforms in the social solidarity economy. *Bus. Strat. Environ.* **30**, 1436–1453 (2021)