Hybrid Intelligent Systems

Research in Computing Science

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 148, No. 6,** junio de 2019. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No.: 04-2005-121611550100-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores. Queda prohibida la reproducción total o parcial, por cualquier medio, sin el permiso expreso del editor, excepto para uso personal o de estudio haciendo cita explícita en la primera página de cada documento. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor responsable: *Grigori Sidorov, RFC SIGR651028L69*

Research in Computing Science is published by the Center for Computing Research of IPN. **Volume 148**, **No. 6**, June 2019. The authors are responsible for the contents of their articles. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research.

Hybrid Intelligent Systems

Edgar Gonzalo Cossío Franco
Carlos Alberto Ochoa Ortiz
José Alberto Hernández Aguilar
Julio César Ponce Gallegos (eds.)

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2019

ISSN: 1870-4069

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

The editors and the publisher of this journal have made their best effort in preparing this special issue, but make no warranty of any kind, expressed or implied, with regard to the information contained in this volume.

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Editorial

This volume of the prestigious journal "Research in Computing Science" presents selected papers that discuss Hybrid Intelligent Systems (HIS) and its applications. Papers were carefully chosen by the editorial board on the basis of the at least two blinded reviews by the members of the reviewing committee and additional expert reviewers. The criteria taken into account were: 1) originality, 2) scientific contribution to the field, 3) soundness, and 4) technical quality of the papers. It is worth noting that 50% of received papers for this special issue were rejected.

A smart city is a complex phenomenon that has become critical for companies to reach their development locally and internationally. On the one hand, macro factors and market structure influence on business competitiveness, but also in a regional or sector context. The internal aspects and the use of various business tools contribute to the ability to create value in an organization.

Through the different chapters of this book, the reader will be identifying how different organizations in the context of diverse societies deploy their resources and leverage their capabilities to achieve better performance of their various labor skills, marketing, social responsibility and management capacity.

Since human capital is a source of competitive advantage, the book includes chapters related to the analysis of cultural differences as a tool to enrich the human capital and making processes more efficient, the factors that influence job satisfaction and the social capital value as competitive strategy for achieving organizational productivity and competitiveness.

Being competitive enables to a company, a region or a country the power to advance in different areas, contributing to the benefit of a social group, therefore, and organizations need to make efforts that lead to adding value and generate a competitive advantage. The papers in this special issue show progress and challenges related to the future of Smart Cities, as well as the need for the human capital to achieve systemic and comprehensive competitiveness required in the XXI century.

We divide our special issue in two sections: a) Industrial and Technological Applications in a Smart City and b) Daily Life in a Smart City.

Section A: "Industrial and Technological Applications in a Smart City"

In "Automatic Generation of Programs for Data Tables with Batch Least Square Mamdani Inference Systems: Applied in the AWG table", Martín Montes et al., Describes the better use of a technology implementation to improve the use of specialized software in a Smart City; On the other hand Mariana Martinez-Valencia et al. in “Determining the optimal and ideal helmet for an Italian scooter used in a Smart City considering cranial anthropometry and intelligent data analysis”, they propose a helmet with ideal and optimal characteristics that save lives in a Smart City when used by the now famous Italian Scooter, which have a high density of vehicular accidents by reach. For his part Humberto Velasco-Arellano in "Forward kinematics for 2 DOF planar robot using linear genetic programming" propose an improvement to the use of a robot capable of being used in various aspects of the industry within the Smart Manufacturing concept. In "Modeling a Roof garden to buildings in a Smart City using equation weight to calculate distribution of load live and weight maximum on a roof top", Angel Calam ert al. propose the implementation of an intelligent model capable of determining the optimal characteristics of a roof garden including its distribution. On the other hand in "PETS-IoTmL: A dog personalized services using IoT and machine learning", Sandra Lopez et al. propose to improve the tracking of pets within a Smart City through the use of avant-garde technology. In "Predicting airline customer satisfaction using k-nn ensemble regression models", Vicente Garcia et al. propose to analyze the components associated with customer satisfaction in an international airline used by several Smart Cities. Edgar Cossio et al. propose in "Predictive model as a tool when acquiring to certification for client companies and certifying entities with machine learning" improve the management processes in a technology company. In "Realtime recoloring objects using artificial neural networks through a cellphone", Martín Montes et al. propose the use of recoloring of objects for diverse implementations in the industry, something of value and that it would help to diminish accidents in the automotive industry. In "A hybrid intelligent system to improve a health model associated with cardiovascular disease", Ana Martinez implements an intelligent hybrid system to prevent cardiovascular diseases in its initial phase. Finally, to close this section, in "Development of a graphic user interface focused on multicriteria analysis among a plethora of passive exoskeletons to improve the social inclusion of infants in a Smart City", Jorge Restrepo et al. describe how the use of cutting-edge technology in this case the use of exoskeletons could improve the social inclusion of children with some type of motor disability.

Section B: "Daily Life in a Smart City"

In "Blurring Northeast Mexican Societies: an approach to cultural capital and results of the PISA Test", Mónica Mendirichaga et al. Determine the results of the Pisa Test among societies of Northeast Mexico and specify how the state of Nuevo Leon stands out only in terms of its heavy industry, but in public policies adequate for the educational aspect of a Smat City. While in "Conceptualization of a predictive model to analyze the health outcomes of dust events in a society with Köppen climate classification BW", Estrella Molina-Herrera analyzes with various intelligent techniques the effects over time of the dust storm and how they can represent a future problem. In "Happiness and its socio-demographic determinants analyzed with datamining, the case of a community at the north of the border of Mexico", Erika Donjuan specifies by intelligent analysis the quality of life in a Smart City. In "Interconnection APP: Proposal of Interaction with a virtual agent, animations and Augmented Reality: An easy way to learn the use of sensors in Smart Cities", Cesar Lozano et al. specify how to interact with the existing flora in a Smart City using technology, especially augmented reality. In "Leisure Organization Models of Young People in the North Mexican Border", Aida Reyes et al. propose how is the social representation of fun using data mining. In "Patterns of motivational orientation and its relationship with academic performance in university students", Arely López determines the scholar performance through an Artificial Intelligence technique. While in "Apptojo: behavioral relationship between consumers and food merchants, through a CRM mobile application" by Uriel Cambrón et al., it is proposed to use technology to determine a gastronomic outlet and assess the options available. In "Recognition of Colors through use of a Humanoid Nao Robot in Therapies for Children with Down Syndrome in a Smart City", Martha Jiménez et al. propose to help children with down syndrome in a therapy that allows them their best integration.

While in "Sign language recognition based on EMG signals through a Hibrid Intelligent System", Bernabé Rodríguez-Tapia et al. propose to help people with hearing deficit to communicate better in a Smart City. On the other hand, in "Visual association rules on the psychological connection of students with their studies", Erika Morales et al. propose an innovative perspective of the visualization of data associated with psychological type tests in a university. Finally to close this section, in "Development of a Serious Games for Asperger's Syndrome, based on a bio-inspired algorithm to measure empathy performance", Alejandro Lara et al. propose to improve the situation of empathy in children with Asperger's Syndrome in a Smart City.

The volume also contains the regular paper on construction of historic Arabic dictionary.

We would like to thank the Mexican Society for Artificial Intelligence, and MICAI 2018 Committee for all the support provided for the publication of this special volume. The entire submission, review, and selection process, as well as the preparation of the proceedings, were supported by the EasyChair system (www.easychair.org).

Edgar Gonzalo Cossío Franco
Carlos Alberto Ochoa Ortiz
José Alberto Hernández Aguilar
Julio César Ponce Gallegos
Guest editors
June 2019

ISSN 1870-4069

Table of Contents

ISSN 1870-4069

A Hybrid Intelligent System for Improving a Health Model Associated with Cardiovascular Disease

A. Martínez, C. Alberto Ochoa, Jorge R. Rodríguez

Universidad Autónoma de Ciudad Juárez, Mexico
al164654@alumnos.uacj.mx

Abstract. In some individuals, cardiovascular disease (CVD) is a congenital defect. However, factors such as stress, caffeine, alcohol, tobacco, and certain medications are the prevailing causes of CVDs. These factors are considered by studies such as the Framingham Heart Study, which is a reference point today to determine cardiovascular risk as a preventive measure. However, this study was conducted on US-based individuals, whose genetics, customs and lifestyle are different from the Latin American population. Due to the above, in our research we'll use the PhysioBC database, which contains 114 registries of inhabitants of Mexicali, Baja California. Each registry has its own fact sheet, electrocardiogram (ECG) record, and doctor's diagnosis, with ages ranging from 18 to 68 years. To process data, we employed data mining techniques for extraction and preprocessing (cleaning). To analyze and interpret data, we used the Waikato Environment for Knowledge Analysis (WEKA), and the classification algorithms Naive Bayes, Multilayer Perceptron, and J48. While testing these algorithms, we obtained the best results with the Naive Bayes classifier.

Keywords: cardiovascular risk, data mining, WEKA, J48, naive Bayes, multilayer perceptron.

1 Introduction

Any abnormality or irregularity in the natural heart rhythm can be defined as arrhythmia, and there are several factors that can cause it. Arrhythmias can be found in individuals with cardiovascular diseases (CVD), which are one of the leading causes of death in the entire world, representing about 30% of total deaths from heart disease [1].

CVDs are a global problem. Estimates suggest that in 2020, CVD deaths will increase from 15% to 20%, and by 2030, approximately 23.6 million people will die due to heart attacks and strokes [2]. In Western countries, CVDs are the leading cause of death and an important source of disability, which, cost-wise, means a huge burden for the healthcare sector [3,4]. To speed up the response of the health sector to CVD, the WHO Global Strategy as well as the Pan American Health Organization (PAHO) Regional Strategy, establish that health systems should focus on promotion and primary health care by increasing prevention and improving medical care [5].

In the literature review we found that risk factors for ECVs are widely identified as: age, diabetes, smoking (treated and untreated), systolic blood pressure, total cholesterol, HDL cholesterol, and Body Mass Index (BMI) [6–10]. There are several studies that can determine the risk of CVD, such as the Framingham Heart Study, whose factors are mentioned above; however, this analysis is based on the US population, with a CVD risk and prevalence different than ours[6], as well as a different lifestyle, socioeconomic status and genetics. Although the Framingham Study is one of the most widely used ones to determine risk factors for cardiovascular diseases, there are several studies that analyze the applicability of this model for a different population than the American. [6,10]. The study of genetic variation compares the data obtained through genome sequencing of different individuals, which allows finding genes linked to certain diseases. In this sense, [11] highlights that caution must be taken when applying genetic CVD risk prediction models based on Single Nucleotide Polymorphism (SNP) that do not belong to the group of ancestors from which it derived. Due to the aforementioned, our work analyzed the CVD risk factors of the Baja California population. For this, we used the open-access PhysioBC database [12], which has the registry of 114 individuals, men and women aged between 18 and 68. This database contains ECG records, fact sheets, and diagnosis by a specialist doctor.

Having a CVD affects the quality of life of the individual, and it has both social and economic repercussions. CVDs are considered costly diseases, and the government of Mexico allocates a part of its budget to the Fund for Protection against Catastrophic Health Expenditure (FPGC) in response to the expenses related to this type of disease [13]. In Mexico, the population pyramid determines that 75% of the adults have less than 55 years old, and despite the fact that prevalence of cardiovascular risk factor is higher after 40, a large amount of the carriers of these risk factors are located in an economically active population [2]. In the same manner, these data can be observed in the population pyramid of Baja California, as shown in figure 1, data obtained from the National Population Council (CONAPO), projections of population for 2010-2050 [14].

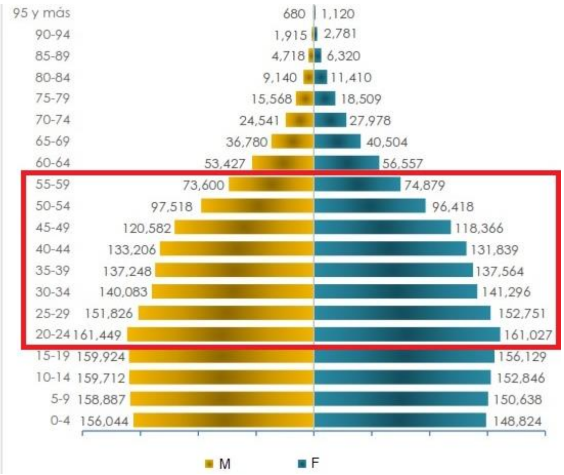


Fig. 1. Population distribution 2017 [14].

At present, there are vast databases that allow us to analyze the trend of the population in regards to public health, in order to find out their current situation and implement strategies to prevent diseases. There are several techniques for processing large amounts of data, among them is Knowledge Discovery in Databases (KDD), a process that allows us to identify valid, novel, potentially useful and easy-to-understand patterns. The process begins with the understanding of the field of study and establishment of specific goals, followed by selection and integration of data from different sources. Due to this, there may be a need to clean up information noise, missing data and other forms of inconsistencies. Once data reprocessing is done, the next step is data mining, in which algorithms are used to find patterns or relationships between databases. The last step of this process consists in interpretation of the obtained information, where knowledge is finally acquired, which allows decision-making analyses [15]. This process can be observed in figure 2.

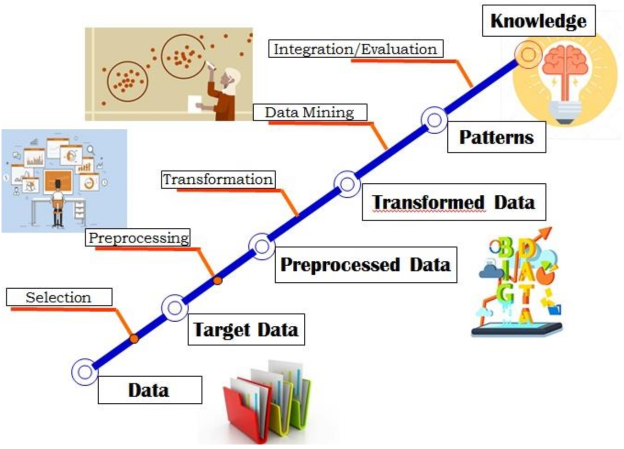


Fig. 2. KDD process partially obtained from: https://mnrva.io/kdd-platform.html

In our study, we used data mining techniques and the open-source software Waikato Environment for Knowledge Analysis (WEKA), written in Java, and developed at the University of Waikato, in New Zealand. WEKA is a collection of machine learning algorithms that contains tools for data preparation, classification, regression, clustering, data mining, association rules mining, and visualization rules [16]. There are several researches that use WEKA for classification, highlighting the use of the Naive Bayes, Multilayer Perceptron and J48 Algorithm for its high efficiency for observing results [17–22].

2 Related Research

In their research, [20] used the J48 Classifier Algorithm of the WEKA platform to predict cancer recurrence. The authors worked with a database of patients that underwent treatment for breast cancer provided by the UCI Machine Learning Repository. The dataset consists of 286 instances and 9 attributes, such as the patient’s age at time of diagnosis, menopause, tumor size, inv-nodes (number of lymph glands that contain metastatic breast cancer), node caps, degree of malignancy, breast (whether left or right is diagnosed with tumor), breast quadrants, and irradiation. The dataset was in an ARFF file format. The selected algorithm was the J48, which analyzes data through decision trees. The study was conducted for experimental purposes, and a decision tree was generated by taking the degree of malignancy as root node, so after interpreting results of the experiment, the authors concluded that patients with a specific value range of this attribute have higher chances of recurrent cancer.

On the other hand, [21] proposed an improved J48 Classification Algorithm to predict risk factors of diabetes. By using an interface between WEKA and MATLAB, they introduced an improved J48 Algorithm. The authors compared the Naive Bayes Classifier Algorithm, the multilayer perceptron (MLP), and the improved J48 Algorithm for the analysis of the same database; and came to the conclusion that the J48 Algorithm performed better at classification, with an accuracy rate of up to 99.87%. The Pima Indians Diabetes Data Set was used for experimental purpose.

For their part, [22] used several classifiers, such as the J48, Decision Tree, Random Trees, Random Forests, and Naive Bayes to analyze the relationship between people with diabetes and the risk of having a heart disease or not. In their experiments, they found that the J48 Algorithm showed the highest accuracy (95%) in comparison to the other classifiers, and that the Naive Bayes Algorithm spent less time in classification (0,00 seconds).

Below we present the results obtained in the application of the Naive Bayes, Multilayer Perceptron and J48 algorithms as an approach for the prediction of CVDs from risk factors obtained from the PhysioBC database.

3 Techniques and Tools Used

In the current work, we conducted experiments by using the PhysioBC database, which contains 114 ECG reports and 91 fact sheets of different patients. From these fact sheets, we generated a database with 89 instances (removing those that had missing, null or insufficient values) and 17 attributes, such as: patient, age, sex, 10-year and 30-year Framingham Risk Score, Body Mass Index (BMI), systolic blood pressure, diastolic blood pressure, smoking, drinking, exercise, diabetes, Arterial Hypertension (AH), CVD diagnosis, treatment, respiratory rate, and heart rate. Of the 89 patients, 50 were women and 39 men, aged between 18 and 68 years. The patients were volunteers from the health sector and the textile manufacturing sector in the municipality of Mexicali, Baja California.

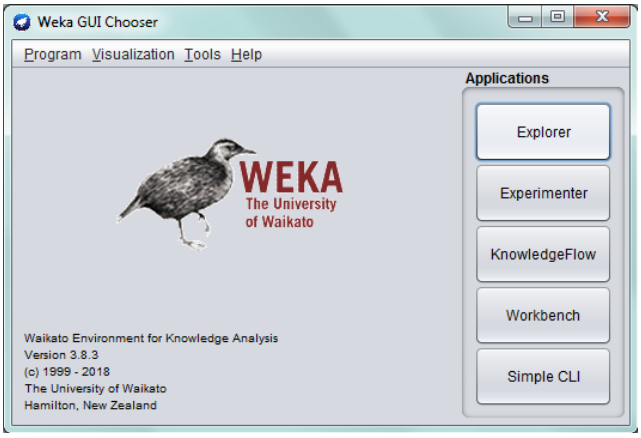
In order to analyze the risk factors of suffering a CVD the database was prepared in a comma-separated values (.csv), making use of the different extensions and formats supported by WEKA, including .arff, .data, .names, .data, and .csv, among others. Figure 3 illustrates WEKA’s Graphical User Interface (GUI), which was used along with the Explorer application shown in the interface menu.

The WEKA platform has several classification algorithms which we work with:

Naive Bayes, which is one of the most widely used classifiers for its simplicity and speed. It is a supervised classification and prediction technique, building models that predict the probability of possible outcomes.

Multilayer Perceptron (MLP), which is a neural network connecting multiple layers in a directed graph, the signal path through the nodes only goes one way. Each node, apart from the input nodes, has a nonlinear activation function. An MLP uses backpropagation as a supervised learning technique, is widely used in research into computational neuroscience and parallel distributed processing.

Algorithm J48, is a version of C4.5 and builds decision trees from a set of training data using the concept of information entropy. At each tree node, choose the attribute that most effectively divides the set of samples into subsets. Its criteria is the normalized for information gain (obtained from the entropy difference) that results in the choice of an attribute to divide the data. The attribute with the highest information gain is chosen as the main node from which the branches are derived.



Waikato Environment for Knowledge Analysis
Version 3.8.3
(c) 1999 - 2018
The University of Waikato
Hamilton, New Zealand

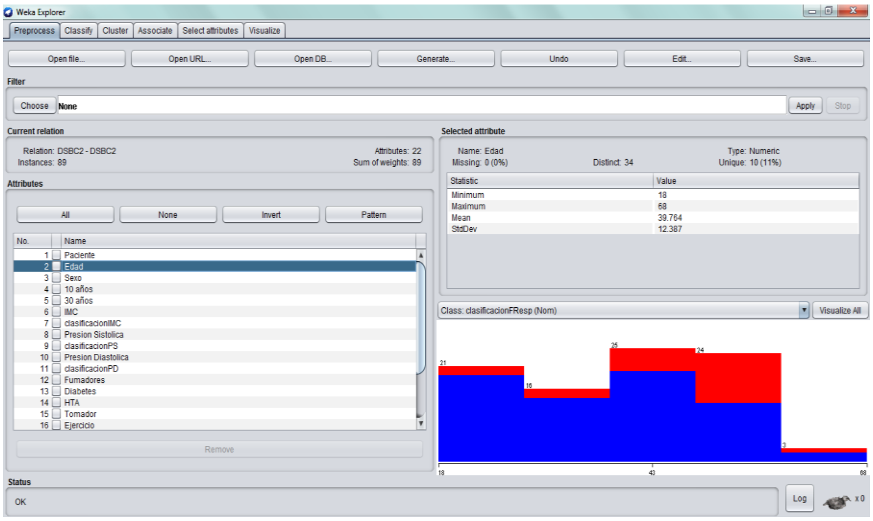
Fig. 3. WEKA graphical user interface (GUI).

Figure 4 shows the GUI of WEKA at the moment of opening the database, in which appears the list of attributes and a graphic representation of the distribution of patients according to the selected attribute, in this case the graph of the attribute "age" is observed.

3.1 Naive Bayes Classifier

In this case, we used WEKA’s ‘Use training set’ option, and obtained 86 instances classified correctly, which represents a 96.62%, where 4 instances were classified as ‘Yes CVD’ and 82 as ‘No CVD’, as observed in the main diagonal of the confusion matrix, shown in figure 5. The Kappa coefficient was 0.7095 which indicates a considerable degree of agreement according to the Landis and Koch scale [23].

To visualize the results in a graphical format, we right-click on the Results list and select the option ‘Visualize Classifier Error’ in the panel. Figure 6 shows the graph obtained, where it can be seen that points in the upper right corner represent instances that had ‘No CVD’ and were classified correctly. In the same way, those that had ‘Yes CVD’ and were classified correctly can be observed in the lower left corner. In this part, we can visualize the attributes of each instance individually by selecting the point in the graph, which makes easier to analyze instances that were not classified correctly. Figure 7 portrays an example of the attributes of an instance that had ‘Yes’ on CVD diagnosis and was not classified correctly, which are the 2 cases observed in the upper left corner of the graph.



Filter: None

Current relation

Relation: DSBC2 - DSBC2 Attributes: 22
Instances: 89 Sum of weights: 89

Attributes

No.		Name
1	□	Paciente
2	□	Edad
3	□	Sexo
4	□	10 años
5	□	30 años
6	□	IMC
7	□	clasificacionIMC
8	□	Presion Sistolica
9	□	clasificacionPS
10	□	Presion Diastolica
11	□	clasificacionPD
12	□	Fumadores
13	□	Diabetes
14	□	HTA
15	□	Tomador
16	□	Ejercicio

Selected attribute

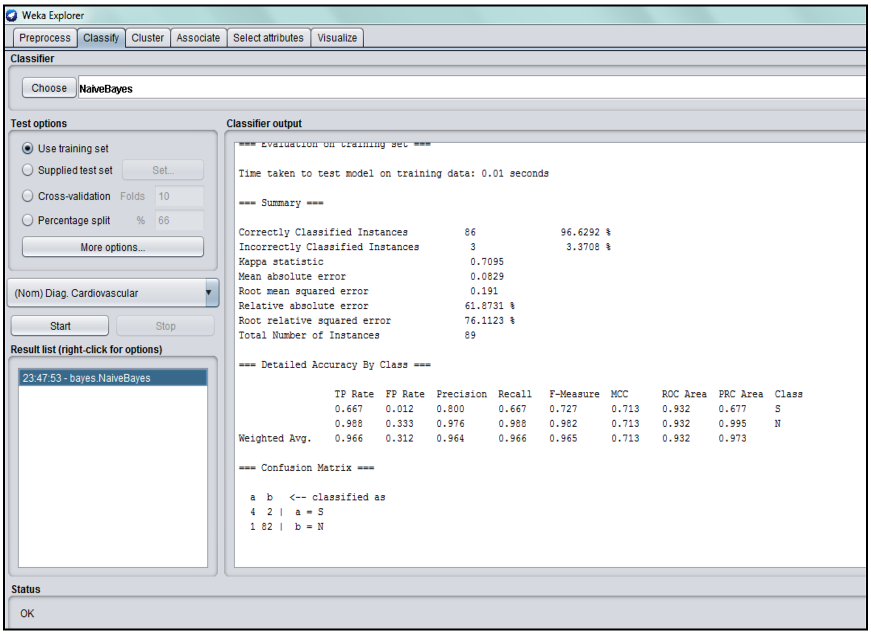
Name: Edad Type: Numeric
Missing: 0 (0%) Distinct: 34 Unique: 10 (11%)

Statistic	Value
Minimum	18
Maximum	68
Mean	39.764
StdDev	12.387

Class: clasificacionFResp (Nom)

Status
OK

Fig. 4. List of attributes shown when opening the database and the graphical representation of the age attribute distribution.



Classifier: NaiveBayes

Test options

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Diag. Cardiovascular

Result list (right-click for options)

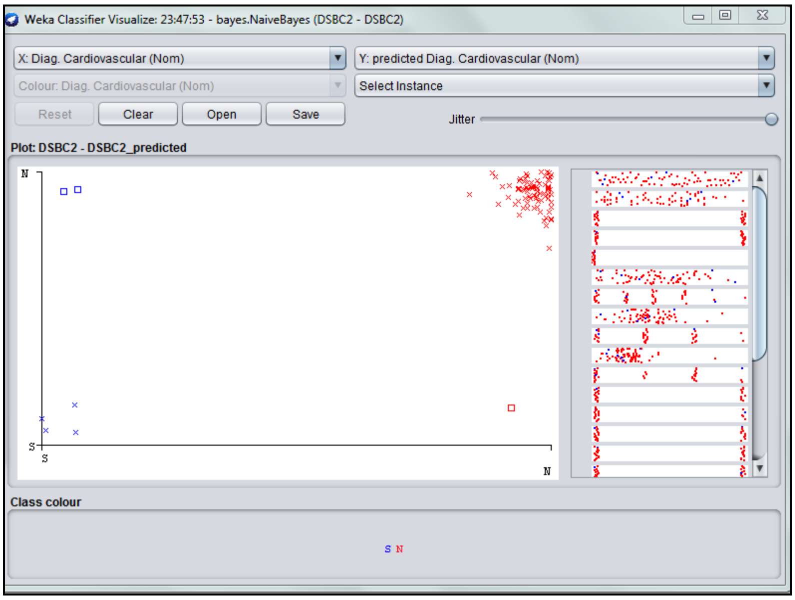
23:47:53 - bayes.NaiveBayes

Classifier output

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          86               96.6292 %
Incorrectly Classified Instances         3                3.3708 %
Kappa statistic                          0.7095
Mean absolute error                      0.0829
Root mean squared error                  0.191
Relative absolute error                 61.8731 %
Root relative squared error             76.1123 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.012    0.800      0.667    0.727      0.713    0.932     0.677     S
                 0.988    0.333    0.976      0.988    0.982      0.713    0.932     0.995     N
Weighted Avg.    0.966    0.312    0.964      0.966    0.965      0.713    0.932     0.973

=== Confusion Matrix ===

  a  b   <-- classified as
  4  2 |  a = S
  1 82 |  b = N
```

Status
OK

Fig. 5. Results of the Naive Bayes classifier.



Weka Classifier Visualize: 23:47:53 - bayes.NaiveBayes (DSBC2 - DSBC2)

X: Diag. Cardiovascular (Nom)

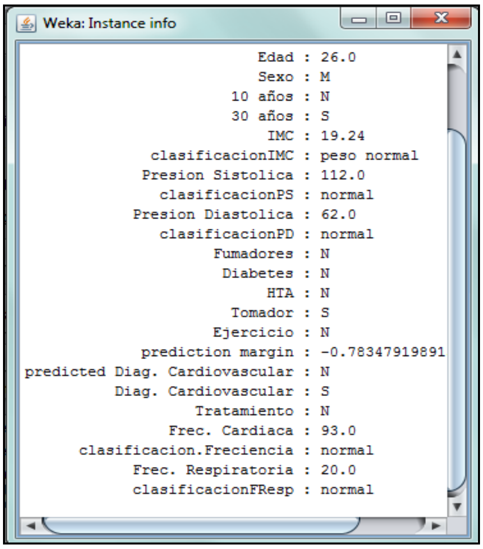
Y: predicted Diag. Cardiovascular (Nom)

Colour: Diag. Cardiovascular (Nom)

Select Instance

Plot: DSBC2 - DSBC2_predicted

Fig. 6. Graphic visualizer of classifier errors with Naive Bayes.



Weka: Instance info

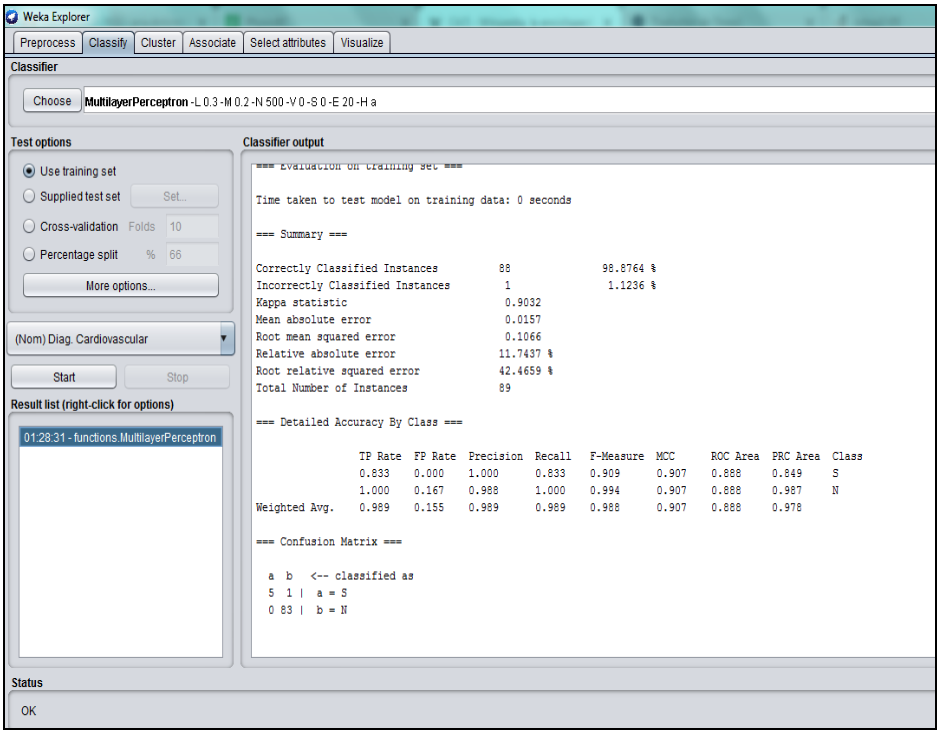
```
                       Edad : 26.0
                       Sexo : M
                   10 años : N
                   30 años : S
                        IMC : 19.24
             clasificacionIMC : peso normal
            Presion Sistolica : 112.0
              clasificacionPS : normal
           Presion Diastolica : 62.0
              clasificacionPD : normal
                   Fumadores : N
                    Diabetes : N
                         HTA : N
                     Tomador : S
                   Ejercicio : N
            prediction margin : -0.78347919891
predicted Diag. Cardiovascular : N
          Diag. Cardiovascular : S
                 Tratamiento : N
               Frec. Cardiaca : 93.0
       clasificacion.Freciencia : normal
            Frec. Respiratoria : 20.0
           clasificacionFResp : normal
```

Fig. 7. Attributes of an instance that had cardiovascular diagnosis (Yes) and was not classified correctly.

3.2 Multilayer Perceptron

In this case we also used the use training set option and obtained 88 instances correctly classified which represents 98.87%, of which 5 instances were classified as Yes

CVD and 83 as No CVD as observed in the main diagonal of the confusion matrix, see figure 8. The Kappa coefficient in this case was 0.9032 which indicates a total or almost perfect concordance according to the Landis and Koch scale [23].



MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) Diag. Cardiovascular

01:28:31 - functions.MultilayerPerceptron

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          88               98.8764 %
Incorrectly Classified Instances         1                1.1236 %
Kappa statistic                          0.9032
Mean absolute error                      0.0157
Root mean squared error                  0.1066
Relative absolute error                 11.7437 %
Root relative squared error             42.4659 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.833    0.000    1.000      0.833    0.909      0.907    0.888     0.849     S
                 1.000    0.167    0.988      1.000    0.994      0.907    0.888     0.987     N
Weighted Avg.    0.989    0.155    0.989      0.989    0.988      0.907    0.888     0.978

=== Confusion Matrix ===

  a  b   <-- classified as
  5  1 |  a = S
  0 83 |  b = N
```

OK

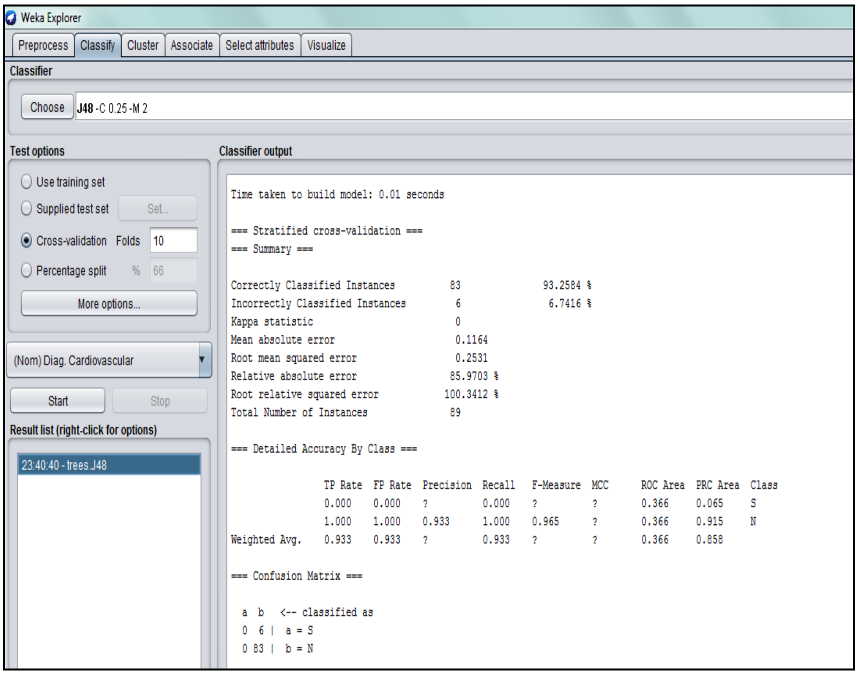
Fig. 8. Results of the Multilayer Perceptron classifier.

To visualize the results in a graphical form, we right-click on the results list and select the option "Visualize the errors" in the panel. Figure 9 shows the graph obtained, in this case only one instance was not classified correctly, as observed in the upper left corner of the graph.

3.3 J48 Algorithm

In this case we also used the ´use training set´ option and obtained 83 instances classified correctly, which represents 93.25%, of which 0 instances were classified as ´Yes CVD´ and 83 as ´No CVD´, as observed in the main diagonal of the confusion matrix in figure 10. Here, the Kappa coefficient was 0.0 which indicates a poor concordance according to the Landis and Koch scale [23].

To visualize the results in a graphical format, we followed the same procedure as in other cases. Figure 11 illustrates the graph obtained. In this experiment, only those instances that had 'No CVD' were classified correctly, while the algorithm failed to classify instances that had 'Yes CVD'. The J48 algorithm is a decision tree classifier, and there is an option to visualize results as a tree; however, in this case, a clear tree was not obtained to determine whether a patient had a CVD or not.



J48 -C 0.25 -M 2

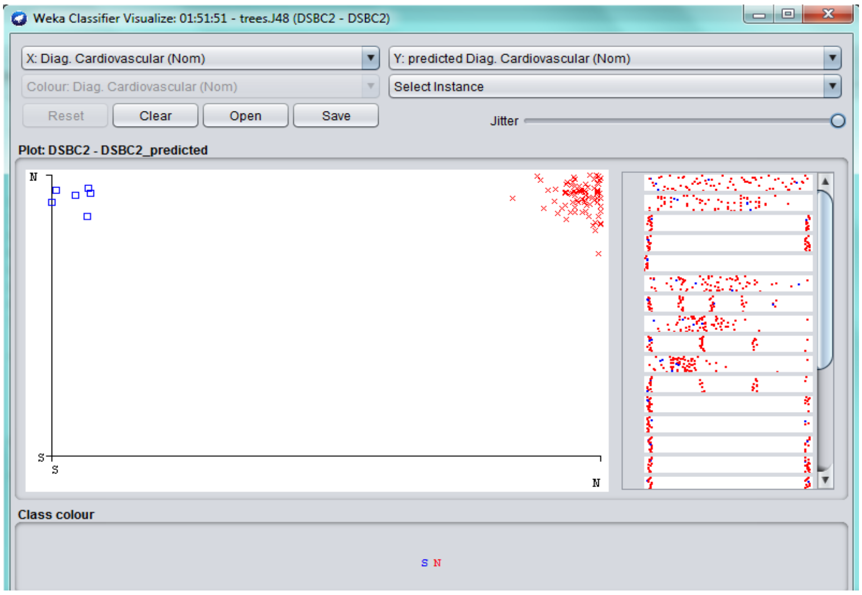
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Diag. Cardiovascular

23:40:40 - trees.J48

```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          83               93.2584 %
Incorrectly Classified Instances         6                6.7416 %
Kappa statistic                          0
Mean absolute error                      0.1164
Root mean squared error                  0.2531
Relative absolute error                 85.9703 %
Root relative squared error            100.3412 %
Total Number of Instances               89

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.366     0.065     S
                 1.000    1.000    0.933      1.000    0.965      ?        0.366     0.915     N
Weighted Avg.    0.933    0.933    ?          0.933    ?          ?        0.366     0.858

=== Confusion Matrix ===

  a  b   <-- classified as
  0  6 |  a = S
  0 83 |  b = N
```

Fig. 10. Results of the J48 classifier.


Weka Classifier Visualize: 01:51:51 - trees.J48 (DSBC2 - DSBC2)
Reset
Clear
Open
Save
Jitter
N
S
S
N
Class colour
S N

X: Diag. Cardiovascular (Nom)

Y: predicted Diag. Cardiovascular (Nom)

Colour: Diag. Cardiovascular (Nom)

Select Instance

Plot: DSBC2 - DSBC2_predicted

Fig. 11. Graphic visualizer of classifier errors with J48.

4 Conclusions

In this work, the Naive Bayes, Multilayer Perceptron and J48 classifiers were used as tools to determine the levels of accuracy in classification and prediction that a patient has a CVD by studying the presence or absence of widely identified risk factors, such as obesity (through BMI), diabetes, HA, smoking, alcoholism, among others. In the results obtained with each of these classifiers we found that the Multilayer Perceptron had better results, with 98.87% of the instances classified correctly, a Kappa coefficient of 0.9032, showing a complete or almost perfect concordance and a response time of 0.0 seconds. On the other hand, the results from the J48 classifier we obtained 93.25% of instances classified correctly, but only from those patients who did not present CVD; additionally, the Kappa coefficient was 0.0, which indicates a poor concordance. After these results, we came to the conclusion that for the J48 algorithm requires more instances that have CVD in order to make predictions correctly, since this data set only had 6 instances with diagnosed CVD. In the analysis of data to predict the risk factor of suffering or not a high risk disease such as CVDs there is still much to do, it is convenient to work on the creation of large databases that can identify the important aspects of a given population, in addition to making them available for use in research this would allow the development of accessible systems that support prevention plans in diseases such as Cancer, Diabetes and Cardiovascular Diseases.

Acknowledgments. We are very grateful to Dr. R. L. Avitia and N. Flores, members of the Department of Bioengineering and Environmental Health of the Autonomous University of Baja California, who kindly provided us with information from the database used in this work.

References

1. WHO: “WHO | Cardiovascular diseases (CVDs),” *WHO* (2016)
2. Sanchez, A., Bobadilla, M., Altamirano, B., Ortega, M., Gonzalez, G.: Enfermedad cardiovascular: primera causa de morbilidad en un hospital de tercer nivel Heart. Rev. Mex. Cardiol. 27(s3), pp. 98–102 (2016)
3. Hidalgo, M.M.: Nuevos modelos multivariantes en la medición del riesgo cardiovascular. Universidad de Salamanca, Departamanto de Estadística (2015)
4. Schnabel, R.B. *et al.*: “50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: A cohort study,” *Lancet*, vol. 386, no. 9989, pp. 154–162 (2015)
5. Gómez, L.A.: Las enfermedades cardiovasculares: un problema de salud pública y un reto global. Biomédica 31(4), (2011)
6. Jiménez-Corona, A., López-Ridaura, R., Williams, K., González-Villalpando, M.E., Simón, J., González-Villalpando, C.: Applicability of Framingham risk equations for studying a low-income Mexican population. Salud Publica Mex. 51(4), pp. 298–305 (2009)
7. Rosas-Peralta, M. *et al.*: Cardiovascular risk reduction : Past, present and future in Mexico. pp. 38–47 (2018)
8. de León, G.P. y P., Campoy, U.R., Bravo, A.C., Witrón, J.J.: Factores de riesgo cardiovascular y la percepción del estado de salud en profesores de tiempo completo de la

UABC, campus Mexicali / Cardiovascular disease risk factors and the perception of health in full professors of the UABC, campus Mexicali. RICS Rev. Iberoam. las Ciencias la Salud 5(10), pp. 98–120 (2016)

9. María, J., Cortes, M., Aragonés, N., Godoy, P., José, M., Moros, S.: Las enfermedades crónicas como prioridad de la vigilancia de la salud pública en España. 30(2), pp. 154–157 (2016)
10. Alvarez, A.:Las tablas de riesgo cardiovascular. Una revisión crítica. Medifam 11(3), pp. 122–139 (2001)
11. Carlson C.S. *et al.*: Generalization and Dilution of Association Results from European GWAS in Populations of Non-European Ancestry: The PAGE Study. PLoS Biol. 11(9), (2013)
12. Flores, N.: PhysioBC. [Online]. Available: http://www.physiobc.org/. [Accessed: 14-Nov-2018]
13. Comisión Nacional de Protección Social en Salud: INTERVENCIONES DEL FONDO DE PROTECCIÓN CONTRA GASTOS CATASTRÓFICOS 2018 (2018)
14. CONAPO: Projections of the Population of Mexico 2010-2050. p. 15 (2010)
15. Minerva Data Mining: KDD: Knowledge Discovery in Databases | Minerva Data. [Online]. Available: https://mnrva.io/kdd-platform.html. [Accessed: 16-Nov-2018].
16. The University of Waikato (NZ): Weka 3 - Data Mining with Open Source Machine Learning Software in Java."[Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 16-Nov-2018].
17. Bhargava, Sharma: Decision Tree Analysis on J48 Algorithm for Data Mining. IJARCSSE 3(6), pp. 1114–1119 (2013)
18. Arora, R.: Pxc3882492 54(13), pp. 21–25 (2012)
19. Kau,r H., Raghava, G.P.S.:A neural-network based method for prediction of γ-turns in proteins from multiple sequence alignment. Protein Sci. 12(5), pp. 923–929 (2003)
20. Sharma, S., Purohit, R., Rathore, P.S.:Prediction of Recurrence Cancer using J48 Algorithm.In:. Icces 2017, pp. 386–390 (2017)
21. Kaur G., Chhabra, A.: Improved J48 Classification Algorithm for the Prediction of Diabetes. Int. J. Comput. Appl. 98(22), pp. 13–17 (2014)
22. Gokilam G.G., hanthi, K.: Performance Analysis of Various Data mining Classification Algorithms on Diabetes Heart dataset. Compusoft 5(3), pp. 2074–2079 (2016)
23. Landis, J.R. *et al.*: Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. Acta Trop. 33(1), pp. 54–58 (2015)

ISSN 1870-4069

Apptojo: Behavioral Relationship between Consumers and Food Merchants through a CRM Mobile Application

Uriel Cambrón Hernández, Mariela Chávez Marcial

Instituto Tecnológico Superior de Ciudad Hidalgo
Ingeniería en Sistemas Computacionales, Mexico
urcambron@gmail.com, marielawiroma@hotmail.com

Abstract. Emerging markets become part of a digital economy [1], in which the marketing of food is one of the most important supply lines in the topics of human consumption, the MiPymes of the State of Michoacán, Mexico, have a clear disadvantage to the large fast food chains, which have technologies (CRM) for the distribution and marketing of their products. The use of mobile technologies allows MSMEs the opportunity to compete with large chains, while creating a consumer experience, giving rise to a set of data that can de- termine the behavior of customers and companies. Apptojo emerges as a tool that provides a solution to link and make sales between customers and compa- nies, generating proposals of dishes visible to the customer and ordered by a dynamic system of geolocation and randomness of most consumed dishes.

Keywords: marketing, android, food, trends, CRM.

1 Introduction

Electronic commerce systems serve as a media pillar in emerging markets. Approximately every minute $ 862,823.00 dollars are spent on online purchases [7]. Of this figure, a substantial percentage goes to electronics, clothing, furniture, and digital content; However, the percentage that goes to electronic food stores is almost imperceptible.

In the state of Michoacán, Mexico, the businesses of the food sector (MiPymes of food sales), have a clear problem when it comes to marketing their products through the Internet, resulting in a sinuous and sometimes counterproductive process in the investment aspect -benefit that can be obtained; subtracting competitive value with large food chains whose processes and marketing lines are territorially and logistically speaking, superior.

Information and communication technologies (TIC's) are a tool that offers the possibility of competing against the marketing and distribution systems of large food chains. In order to provide an agile tool for the MSMEs of the food sector, the mobile app "Apptojo" was created, which allows the food business (entrepreneur), offer their dishes, promote their business and specify orders with the final consumer (customer); by means of a geolocation system the client is offered the possibility of searching for

pp. 25–37; rec. 2018-09-05; acc. 2018-10-07

dishes by categories, names, businesses, and throwing as a result the ones that are closest, at the same time suggesting dynamic dishes in order to generate consumer mobility in the future choices of the customer.

For the development of the mobile application "Apptojo" was based on a hermeneutical methodology, where the study background was the mobile applications of Uber Eats and Without Apron, because they are the most used in the country, but without reflecting an exponential growth as It was planned by the companies that developed them.

2 What is in the Current Market?

There is a variety of similar projects in the market, of which projects in the state of Michoacán can be highlighted, generated by engineering students in systems; the projects are omitted since none has been shown with a scientific or research approach of user behavior, joined to the disappearance of most of these projects attributed to mismanagement of distribution and marketing strategies.

For the purposes of the investigation, some projects were analyzed that, although they are not identical to the project, meet the expectations of the goals by individual sought. The project "FoodGo", intended to generate an increase in the information of a dish through situated analytics [13]; however, the results were not assertive in terms of a good market outcome, in addition to the fact that the application remained a prototype that, although it had a high expectation on the part of the users who used it, could not go to market, and therefore, it did not generate a commercial impact that could be quantifiable.

The NYAM project, part of the expectation of generating a link of recreational and necessity consumption in the end user, allowing him to find businesses near his location in real time, and filtering the best food businesses based on the opinion of other consumers [14], the project generated favorable results, a good market acceptance (for the area where it was applied), however the graphic quality of the application has not favored its expansion in the market, besides having a great potential for growth, if the application is submitted to a BPMN model [11] to be able to restructure its GUI and generate a fidelity segment in its users.

A project that was very useful to support the principles of user behavior, is "Differences in perceptions about food delivery apps between single-person and multi-person households", since it is an SEM that allows to determine the behavior of the relationships between multiple consumption structures (customers-businesses) [15]. The project has promising results that allow scalability in the medium term, however, the application was generated as an evaluation of the hypothesis and cannot be quantified or compared directly with Apptojo; although the hypothesis model is highlighted and recommended for future scope of the project.

3 Methods

The development of a mobile application requires an in-depth study of What? And for what? We want to do, considered as principles of innovation in quality models such as

CMMI or Babok [11]. In order to comply with the previous premise, the agile development model SCRUM [8] was adopted, given that this model allowed working with a reduced work team (3 members) in charge of the application's development, algorithm, logic and software engineering, in addition to a team external to the development area, made up of professionals in the areas of graphic design, user experience, digital marketing and finance, with both teams approved in one, the application could be developed without major inconvenience.

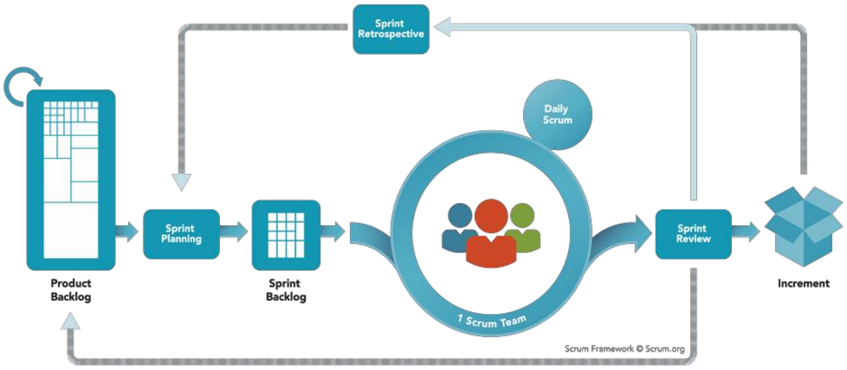


Fig. 1. SCRUM Methodology [8].

In the CMMI Dev Model v1.3 it is explained: "In Agile environments, the needs and ideas of the client are iteratively produced, elaborated, analyzed and validated. The requirements are documented in forms such as: user stories, scenarios, use cases, product backlog and iteration results (code under development in the case of software). What requirements will be addressed in a given iteration are determined by an assessment of the risk and by the priori- ties associated with the requirements left in the product's Product Backlog. What details of the requirements (and other artifacts) to document are deter- mined by the need for coordination (between team members, teams and sub- sequent iterations) and the risk of losing what has been learned. When the client is on the team, there may still be a need to separate the client and product documentation to allow multiple solutions to be explored. While the solution arises, the responsibilities of the derived requirements are assigned to the appropriate equipment" [12], so it is suggested the implementation of agile and simple acquisition techniques for the work teams that require it.

In order to replicate the development of mobile applications and CRM to generate an experience between customers and consumers, it is suggested to use the techniques listed below, with the aim of achieving results based on creativity and the conceptualization of ideas that generate added value in models technologies for electronic commerce.

3.1 Workshops

It is an effective technique to obtain information quickly from several people.

- It is advisable to have a predefined agenda and pre-select the participants, following good practices for effective meetings.
- A neutral facilitator and a transcriptionist (other than the facilitator) can be used.
- You can use a common material on which to focus attention and talk, for example, a presentation with a breakdown of the process being studied or a flow chart.

They can be combined with other techniques such as interviews and questionnaires.

3.2 Observation

- Consists of studying the work environment of users, customers and project stakeholders (Stakeholders) [5].
- It is a useful technique when documenting the current situation of business processes.
- It can be of two types, passive or active.
- In passive observation, the observer does not ask questions, limited only to taking notes and not interfering in the normal performance of operations.
- In active observation, the observer can talk with the user.

3.3 Brain Storm

- It is a structured work session oriented to obtain as many ideas as possible.
- It is advisable to limit them in time, use visual aids and designate a facilitator.
- Rules are important, for example, the criteria for evaluating ideas and assigning a score, not allowing criticism of ideas and limiting discussion time.
- In the first phase, the largest number of ideas must be identified and then evaluated. All ideas must be considered and an idea must be limited to drowning or criticizing before having time to develop it [6].

In previous lines mentioned the axes that supported the development of the application: What? And for what?, which represents a challenge at the time of generating the means for development and the capture of requirements, especially in a mobile application that seeks to implement in a community where technology is not the determining factor in the models of business and marketing of the products, representing the challenge when generating projects that have a high level of innovation, creativity, disruptivity and that are candidates to generate a new experience scheme and facilitate the observation of the behavior of the agents they are part of the user experience. In the case of the mobile application, the following models were generated that fulfilled the expectations of the project:

The fulfillment of the Why? It was given to implement a model where the needs of the entrepreneur are functional and satisfied through the use of the application. The issues to solve in this method are the validity and verification of the expectations of the client, as well as the capacity of technological adoption by the employer; where the human factor plays an important role to assimilate the idea that an application is a profitable tool, sustainable and scalable to new technologies that facilitate the empowerment of the entrepreneur. The key in this section is to ask whether the objective set by the client [5], entrepreneur and the development team is satisfied for all involved, generating a harmony between the technological, commercial and marketing aspects.

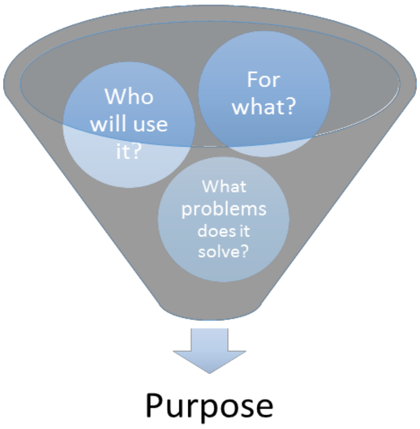


Fig. 2. Purpose delimitation model.

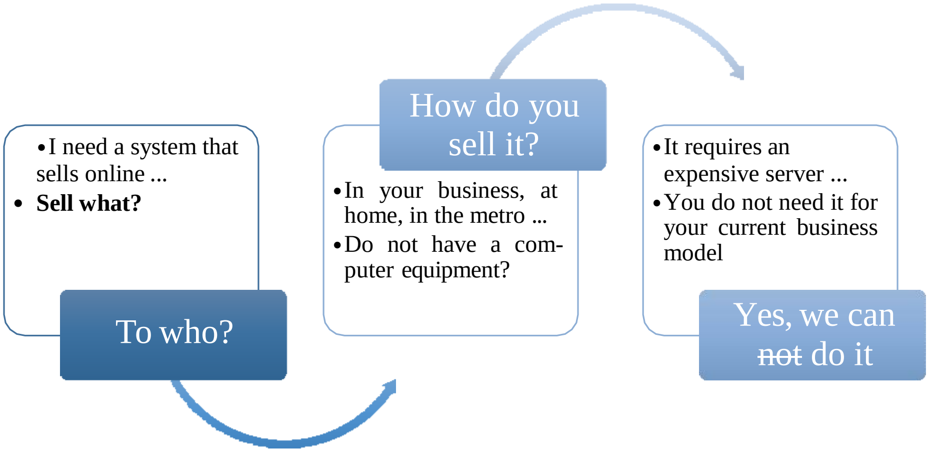

I need a system that sells online ...
Sell what?
To who?
How do you sell it?
In your business, at home, in the metro ...
Do not have a computer equipment?
It requires an expensive server ...
You do not need it for your current business model
Yes, we can ~~not~~ do it

Fig. 3. Model of Why? (verification and validation).

When talking about the development of the application, technologies for mobile development were used that allowed to speed up the development in order to be able to focus on the conceptualization and engineering of project requirements, for this reason technologies such as Android Studio were used to codify the project using the Java and Kotlin programming languages.

As a backend system, Firebase was adopted, which is a real-time database system, messaging, mail server, test lab, among other services that were not used at the time.

For the graphics processing, the Glide library was used, and the Openpay REST API was used as the payment system.

The complete system is based on a hybrid model of its own REST API, which allows the generation of payments by means of a credit / debit card only for the subscription of the business in the application.

4 Results

The generation and publication of the application was successfully achieved obtaining a plausible audience in the first days of organic dissemination in social networks of the application. Below are the graphics of the application in Play Store, as well as in operation.

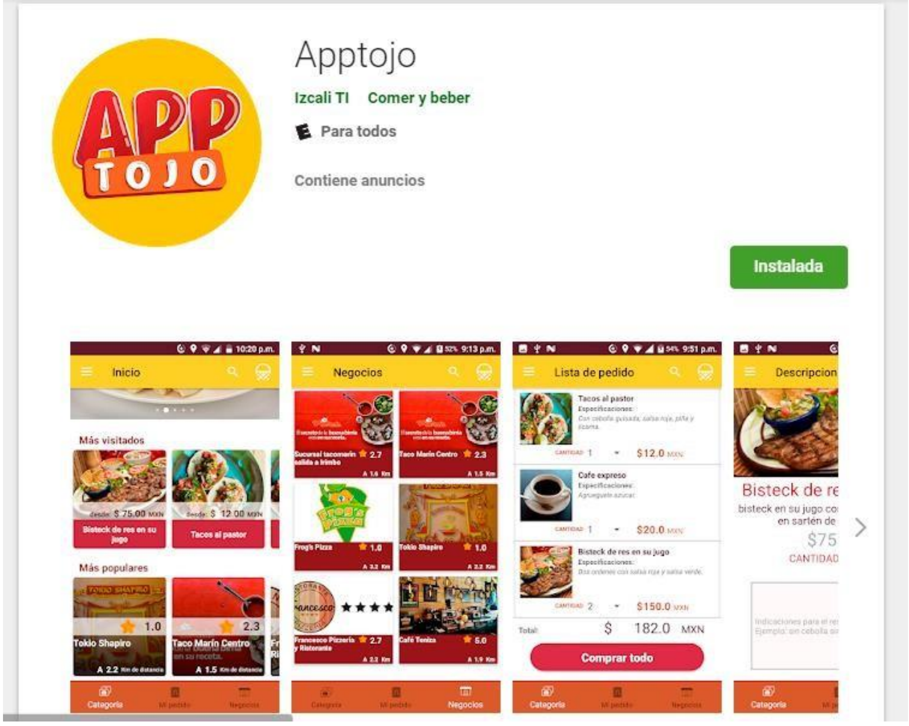
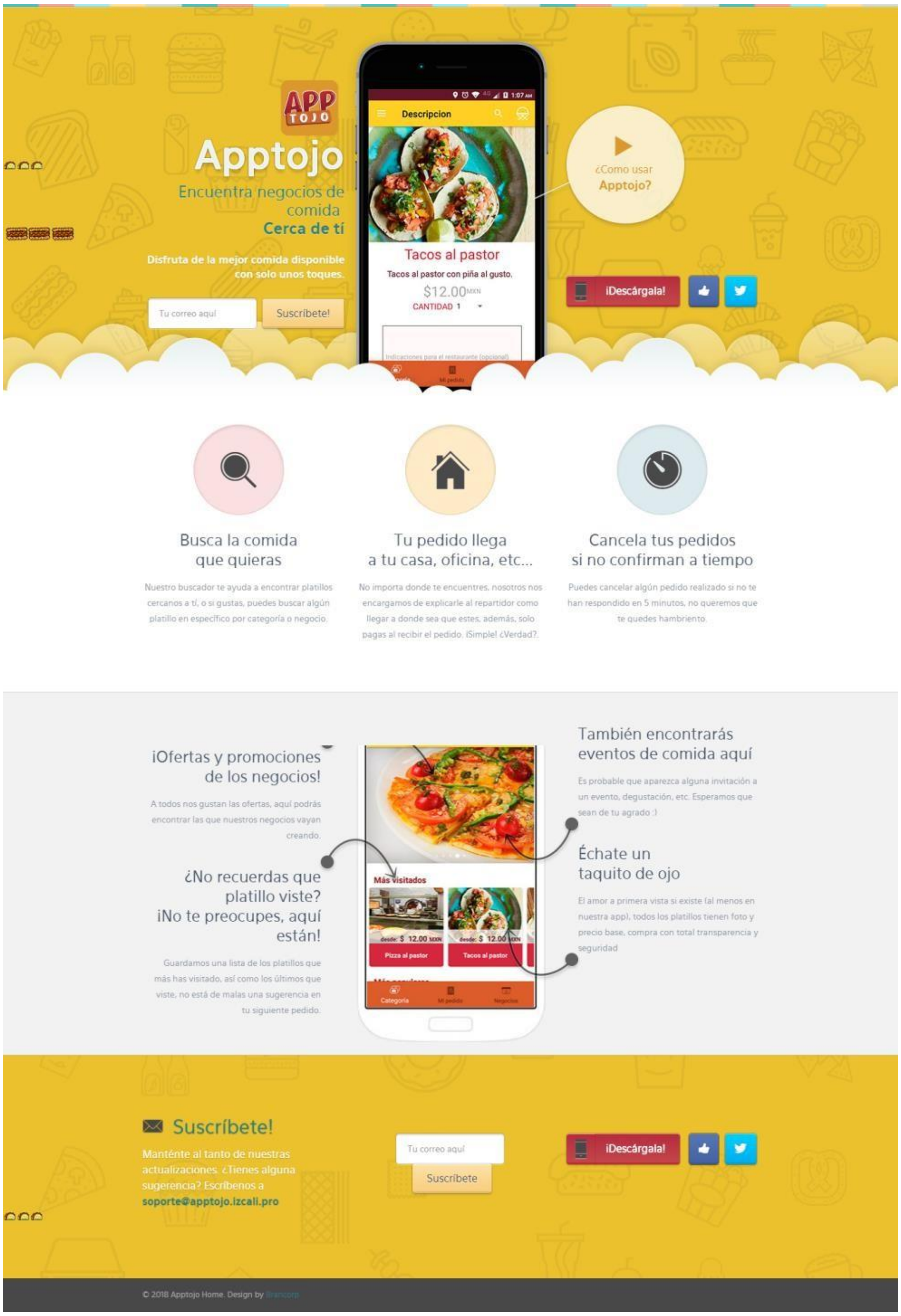


Fig. 4. Play Store view of the Customer version.

The generation of a launch page is essential to form a network of SEO positioning, then the view of the official website of the project is presented.

In the illustration 4 you can see the website of the Apptojo application; This site allows to obtain general information about the project, as well as to provide a download link for the application in Playstore.



Apptojo

Encuentra negocios de comida

Cerca de tí

Disfruta de la mejor comida disponible con solo unos toques.

Tacos al pastor

Tacos al pastor con piña al gusto.

$12.00 MXN

CANTIDAD 1

Busca la comida que quieras

Nuestro buscador te ayuda a encontrar platillos cercanos a tí, o si gustas, puedes buscar algún platillo en específico por categoría o negocio.

Tu pedido llega a tu casa, oficina, etc...

No importa donde te encuentres, nosotros nos encargamos de explicarle al repartidor como llegar a donde sea que estes, además, solo pagas al recibir el pedido. ¡Simple! ¿Verdad?.

Cancela tus pedidos si no confirman a tiempo

Puedes cancelar algún pedido realizado si no te han respondido en 5 minutos, no queremos que te quedes hambriento.

¡Ofertas y promociones de los negocios!

A todos nos gustan las ofertas, aquí podrás encontrar las que nuestros negocios vayan creando.

¿No recuerdas que platillo viste? ¡No te preocupes, aquí están!

Guardamos una lista de los platillos que más has visitado, así como los últimos que viste, no está de malas una sugerencia en tu siguiente pedido.

También encontrarás eventos de comida aquí

Es probable que aparezca alguna invitación a un evento, degustación, etc. Esperamos que sean de tu agrado :)

Échate un taquito de ojo

El amor a primera vista si existe (al menos en nuestra app), todos los platillos tienen foto y precio base, compra con total transparencia y seguridad

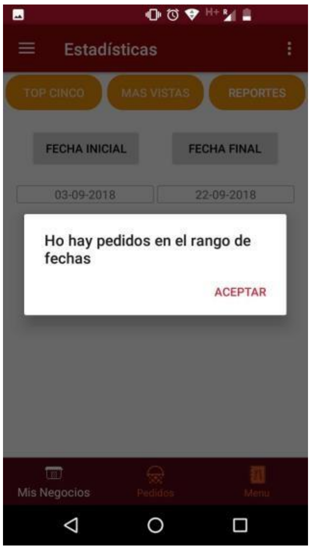
Suscríbete!

Manténte al tanto de nuestras actualizaciones. ¿Tienes alguna sugerencia? Escríbenos a **soporte@apptojo.izcali.pro**

© 2018 Apptojo Home. Design by Brancom

Fig. 5. Landing Page by Apptojo.

Experience of the mobile application for entrepreneurs:

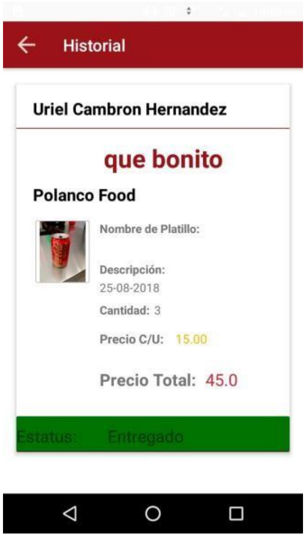


03-09-2018

22-09-2018

Ho hay pedidos en el rango de fechas

Fig. 6. Apptojo View for Companies.



Uriel Cambron Hernandez

que bonito

Polanco Food

Nombre de Platillo:

Descripción:
25-08-2018

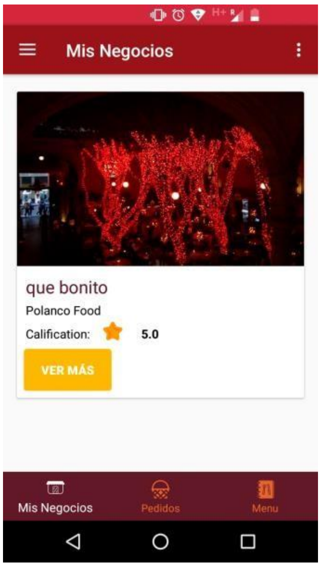
Cantidad: 3

Precio C/U: 15.00

Precio Total: 45.0

Estatus: Entregado

Fig. 7. View of Businessman's History

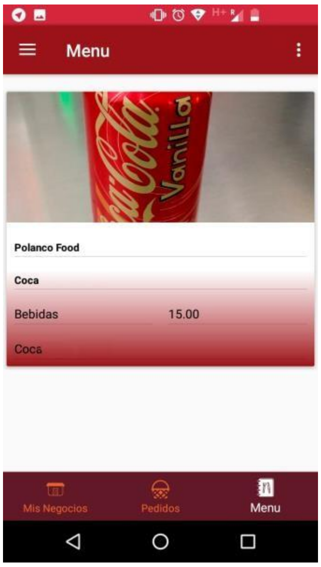


que bonito

Polanco Food

Calification: 5.0

Fig. 8. Business view.



Polanco Food

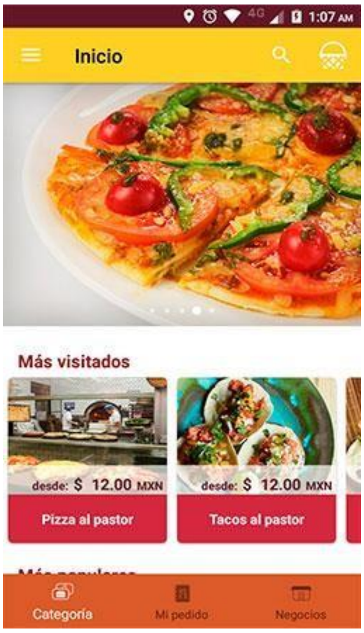
Coca

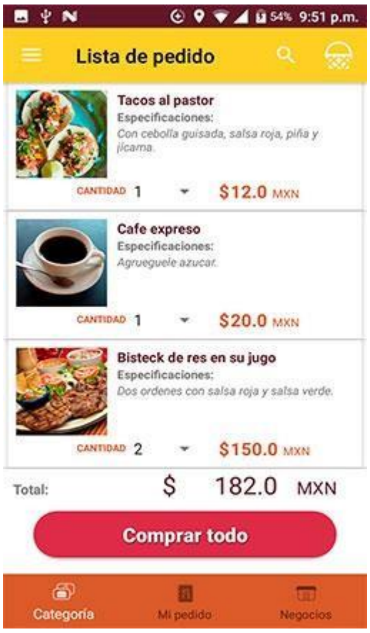
Bebidas 15.00

Coca

Fig. 9. Menu View.

The client application seeks to encourage you to consume a dish of those present in the list, for this a combination of graphic elements that generate appetite from the first view of the application was sought:





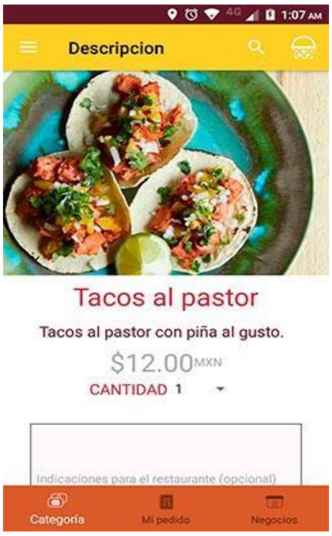
Tacos al pastor
Especificaciones:
Con cebolla guisada, salsa roja, piña y jícama.
CANTIDAD 1 $12.0 MXN

Cafe expreso
Especificaciones:
Agrueguele azucar.
CANTIDAD 1 $20.0 MXN

Bisteck de res en su jugo
Especificaciones:
Dos ordenes con salsa roja y salsa verde.
CANTIDAD 2 $150.0 MXN

Total: $ 182.0 MXN

Fig. 10. Main view of the client application. **Fig. 11.** List of the customer's order.



Tacos al pastor
Tacos al pastor con piña al gusto.
$12.00MXN
CANTIDAD 1
Indicaciones para el restaurante (opcional)

Fig. 12. Saucer view.

4.1 Behavior Results Between Clients and Businesses

By analyzing the behaviors among the different types of users, the following results were obtained:

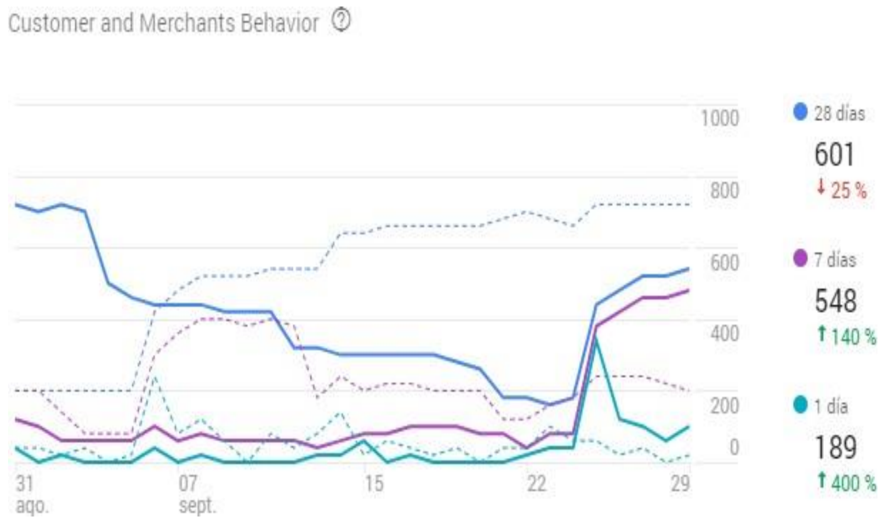
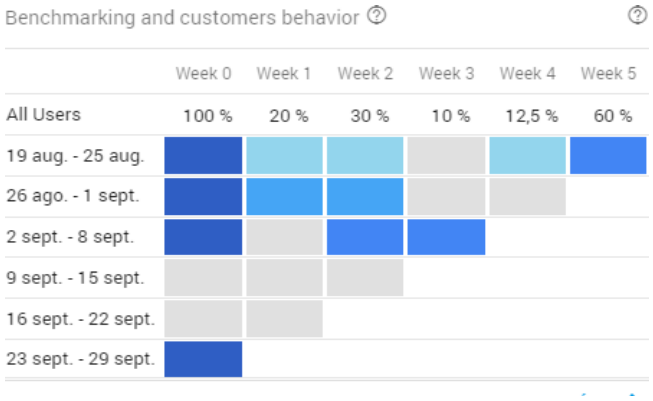


Fig. 13. Behavior between Customers and Merchants.

Through a growth rate of 140% in the last 7 days of evaluation of the application at the time of the cut (September 29), you can see the increase of 140% interaction in contrast to the audience captured 28 days ago, at the time of publishing the first version of the application in Play Store.



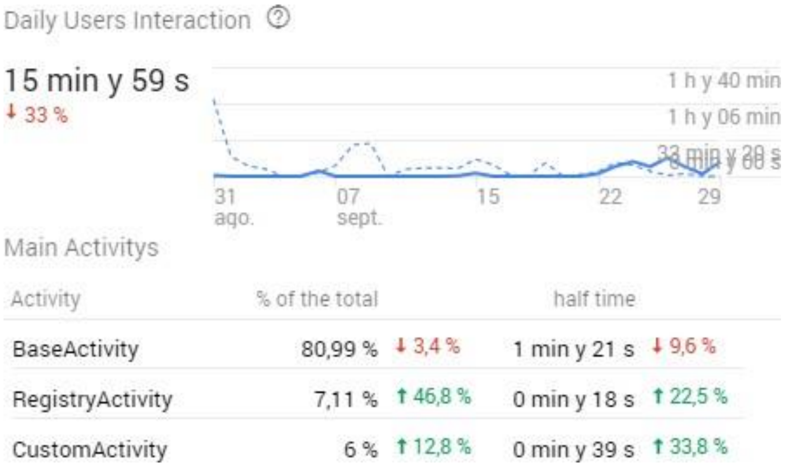
Benchmarking and customers behavior

	Week 0	Week 1	Week 2	Week 3	Week 4	Week 5
All Users	100 %	20 %	30 %	10 %	12,5 %	60 %
19 aug. - 25 aug.	🟦	🟦	🟦	⬜	🟦	🟦
26 ago. - 1 sept.	🟦	🟦	🟦	⬜	⬜	
2 sept. - 8 sept.	🟦	⬜	🟦	🟦		
9 sept. - 15 sept.	⬜	⬜	⬜			
16 sept. - 22 sept.	⬜	⬜				
23 sept. - 29 sept.	🟦					

Fig. 14. Retention and Customer Behavior for Weeks.

4.2 Times Used

The interaction in the application requires less than 5 minutes to be able to generate an order (on the client's side), and approximately 20 minutes on the business side to be able to add an order. The average time in the application is 15 min and 59 s, the above arises as a result of the interaction segments that lead to events, in which users generate interactions between them, confirming or canceling orders, as well as generating reports of sales, history and qualifications of orders.



Main Activitys

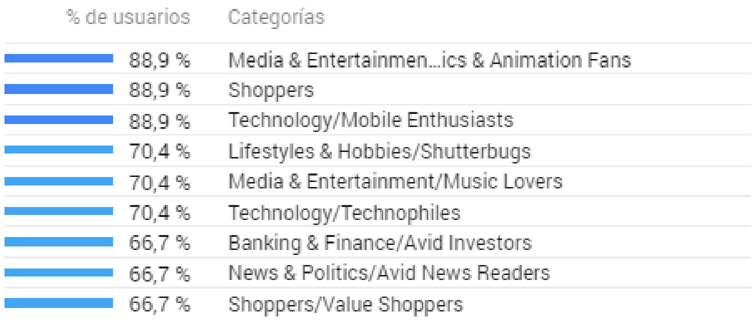
Activity	% of the total		half time	
BaseActivity	80,99 %	↓ 3,4 %	1 min y 21 s	↓ 9,6 %
RegistryActivity	7,11 %	↑ 46,8 %	0 min y 18 s	↑ 22,5 %
CustomActivity	6 %	↑ 12,8 %	0 min y 39 s	↑ 33,8 %

Fig. 15. Daily Users Interaction.

4.3 Behavior of Users Who Downloaded the Application

The behavior of the users (Table 1) was variable in terms of trends and the area from which the application was downloaded, along with the category as the application was found in the Play Store. Most of the audience was found in the category of Media and Entertainment, as well as purchases, technology, and lifestyle, with this goal is achieved to bring people looking for a style of digital consumption, with the MiPymes of Hidalgo, Michoacán, Mexico.

Table 1. Trends Behavior of Customers.

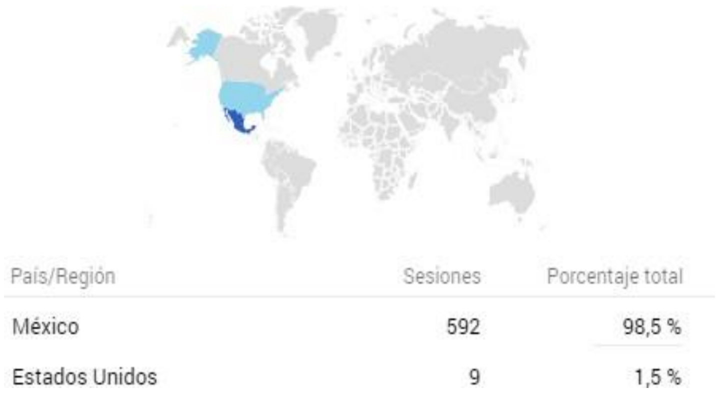


% de usuarios	Categorías
88,9 %	Media & Entertainmen...ics & Animation Fans
88,9 %	Shoppers
88,9 %	Technology/Mobile Enthusiasts
70,4 %	Lifestyles & Hobbies/Shutterbugs
70,4 %	Media & Entertainment/Music Lovers
70,4 %	Technology/Technophiles
66,7 %	Banking & Finance/Avid Investors
66,7 %	News & Politics/Avid News Readers
66,7 %	Shoppers/Value Shoppers

4.4 Scope and Relevance of the Application

It was satisfying to find that the application was a great taste for the geo- graphic segment where the project began, however, the unexpected in the United States was unexpected, from which no scope was expected. In a saucer purchase report, a business reported an order generated with a telephone and location in the United States, with this the project's scalability and market vi- ability are foreseen to expand in a few years to the rest of the country.

The total number of visits generated in the last days was 621 customers, (the business data is kept in a confidential phase); Of which, 98.5% were in Mexico, while 1.5% were in the United States.



País/Región	Sesiones	Porcentaje total
México	592	98,5 %
Estados Unidos	9	1,5 %

Fig. 16. Customer Location.

5 Conclusions and Future Work

The generation of projects and electronic marketing systems are a necessity to generate growth in global economic markets, allowing Mexico to incentivize the use of digital currencies to obtain goods and services.

Apptojo's application allows, among other things, the following:

- Food marketing,
- Promotion of local businesses,
- Consolidation of CRM systems for MiPymes [2],
- Exponential increase in sales in MiPymes,
- Facilitate the customer to obtain dishes close to the location,
- Generation of a Big Data system for the analysis of user behavior (which consumes, where, in what season).

The models of mobile applications that innovate in the conceptualization of CRM, should be a priority in the stack of projects of universities, approved to the private initiative, you can obtain significant results that are an engine of change for the benefit of society in which involves the development and implementation of a technologicalproject.

We thank the honorable committee of HIS 2018 for the opportunity to present the progress and results of this project, wishing success in the future research work carried out, in order to promote scientific and technological development that allows the transcendence of knowledge in humanity.

References

1. del Barrio, L.: Del Business al e-Business en tiempos de crisis. Gestión 2000 (2003)
2. Escobar, M.: El Comercio Electrónico, perspectiva presente y futuro en España. Fundación AUNA (2000)
3. Terceiro, J.B., Matías, G.: Digitalismo, el nuevo horizonte sociocultural. Taurus (2001)
4. García-Valcárcel, I., Munilla, E.: e-Business Colaborativo. FC Editorial (2003)
5. Eisenmann, T.R.: Internet business models: text and cases. McGraw-Hill (2002)
6. Chaffey, D.: E-business and E-commerce management, 2nd edition. Prentice Hall (2003)
7. Ontiveros, E.: La economía en la Red, Nueva Economía, nuevas finanzas. Taurus (2001)
8. Schwaber, K., Sutherland, J.: The Scrum Guide™ (2017)
9. Javier-Maestre, J., Almeida, C.: La Ley de Internet: regimen jurídico de los servicios de la Sociedad de la Información y el comercio electrónico. Servidoc (2002)
10. Vázquez, C.: Comercio electrónico, firma electrónica y servidores: Comentarios y anexo Legislativo. Editorial Dijusa (2002)
11. International Institute of Business Analysis (IIBA): Babok A guide to the business Analysis Body of Knowledge (2015)
12. Equipo del Producto CMMI.: CMMI para Desarrollo, Versión 1.3 CMMI-DEV, V1.3. Software Engineering Institute (2010)
13. Ochoa, A., Ruiz-Jaimes, M., Leon, S., Toledo, Y., Ramírez, I.: Decreased Business Uncertainty by Using Bayesian Networks for the Paradigm Shift in Business Simulator. Research in computing science (2016)
14. Rodríguez-Díaz, A., Juárez-Martínez, U., Peláez-Camarena, G., Muñoz-Contreras, H., Olivares-Zepahua B.A.: Intercesión en invocaciones con la reflexión de Java. Research in computing science (2016)
15. Urbina-Delgadillo, M.L., Figueroa, M.A.A., Peláez-Camarena, G., Alor-Hernández, G., Sánchez-García A.I.: Mixing Scrum-PSP: Combinación de Scrum y PSP para mejorar la calidad del proceso de software. Research in Computing Science (2016)
16. Piña-García, C.A., Gu, D., Gershenson, C., J. Siqueiros-Garca, M., Robles-Belmont, E.: Exploring Dynamic Environments Using Stochastic Search Strategies. Research in computing Science (2016)
17. Safi, H., Jaoua, M., Belguith-Hadrich, L.: Learning-to-Rank for Hybrid User Profiles. Research in Computing Science (2017)
18. Abaoa, R.P., Malabananb, C.V., Galidoc, A.P.: Design and Development of FoodGo: A Mobile Application using Situated Analytics to Augment Product Information. Procedia Computer Science (2018)
19. Cho, M., Bonn, M.A., Li, J.: Differences in perceptions about food delivery apps between single-person and multi-person households. International Journal of Hospitality Management (2018)
20. Isabela, E., Drona, J., Fadhilah, N., Tanoto, D.F., Harefa, J., Prajena, G., Chowanda, A.A.: NYAM: An Android Based Aplication for Food Finding Using GPS Procedia Computer Science (2018)

ISSN 1870-4069

Automatic Generation of Programs for Data Tables with Batch Least Square Mamdani Inference Systems: Application in the AWG Table

Martín Montes Rivera[1], Alberto Ochoa Zezzatti[2], Christian Alejandro Mejía Ramírez[1]

[1] Universidad Politécnica de Aguascalientes, Mexico
[2] Universidad Juárez Autónoma de Tabasco, Mexico
martin.montes@upa.edu.mx, alberto.ochoa@uacj.mx

Abstract. Tables are used in several areas where is required to show value responses, relationships, scores, percentages, and statistical results, among others. Tables increase accessing speed to information, but they need space for be presented and saving space could make that all the required values be not always included. Computer programs can replace tables, so that space not be required when presenting information, this has been explored since computers were developed presenting several algorithms. Fuzzy systems could be an alternative to generate programs for replace tables, more over this allow to get an expert system that knows the data in the table, but fuzzy systems require several numerical parameters to be tuned. In this paper is proposed the use of Batch Least Square Mamdani system for mimic tables in computer programs specifically applied in the AWG table which is a very common table used by engineers and electrical technicians.

Keywords: tables to programs, system identification, batch least square Mamdani systems.

1 Introduction

Arrays of rows and columns for presenting information are called tables, they are used in several areas where is required to show value responses, relationships, scores, percentages, and statistical results, among others. Measurements obtained from input-output models and the evaluation of functions are information that could be also expressed using tables by describing the behavior of dependent variables in function of independent variables; i.e., its relation can be expressed as a mathematical function [1].

Tables increase the speed when accessing to information, but they need space so that tables be presented and this could make that all the required values be not always included forcing a person to perform interpolation or extrapolation operations for determining the missing values [1].

Computer programs can replace tables, especially when it is possible to stablish specific ranges for determining values in tables. More over this have been documented since computers were developed leading to the creation of specific algorithms like those proposed in [2].

An alternative way for converting tables to programs is explored in [1] in this work tables are returned to equations by using Genetic Programming (GP) an optimizing algorithm for determining unknown structures using natural selection principles. But this technique uses a high complexity algorithm, which could demand long time for determining a suitable equation solution, depending from the data.

Fuzzy inference systems are based on Zadeh proposed fuzzy logic, which is the same kind of logic used when constructing inference rules, i.e. if then rules, but with fuzzy limits in range [0,1] or not binary concluding limits (crisp logic) [3].

Fuzzy systems are based on expert knowledge which is knowledge given by an expert, i.e., the input data for training them is given as correct and then systems are heuristically trained for following the intrinsic logic in the input data allowing them to be applied in several areas like modeling human reasoning, recommendation systems, systems identification, automatic control among others [4].

Input data for training inference systems it is always in pairs because its required a relational input-output map containing the expert knowledge, since the system must react correctly for an input premise with an output consequence (If input is A then output is B) [3].

When fuzzy systems are correctly trained they become an available expert to assist persons selecting the correct resource or action depending from the application, in this case use the AWG gauge table would generate a system that automatically recommend the best wire for a specific application.

Inference systems are always trained using numerical data (expert knowledge) like the information in a table for example in the AWG table (Table 1) there are several columns that describe information for 33 kinds of wires with different proprieties. If the AWG table is adjusted as input data for training an inference system, this would require a single input for the AWG gauge and 9 outputs for giving the 9 columns information about every specific class of wire (Fig. 1), when this is manually adjusted fuzzy sets are tuned by setting its numerical parameters, supposing that there are 3 fuzzy sets per input and output with 3 parameters each one, then would be at least 30 numerical parameters that must be adjusted and this does not guarantee that all the data in the table would be correctly reproduced [5].

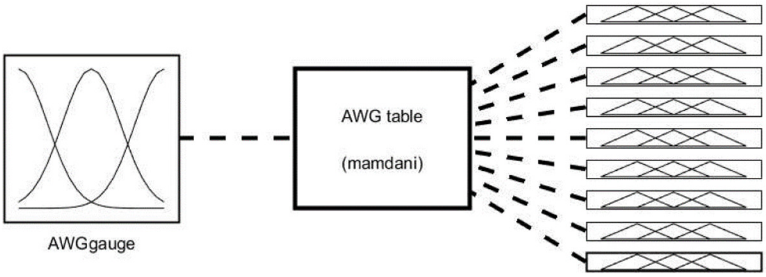


Fig. 1. Inference system required for mimic AWG table.

To reduce the complexity of the system for tuning the parameters the fuzzy system could be divided into 9 sub-systems with a single input and a single output (Fig. 2) and there would be 6 numerical parameters per sub-system that must be optimized, but this would require 54 total numerical parameters to be adjusted.

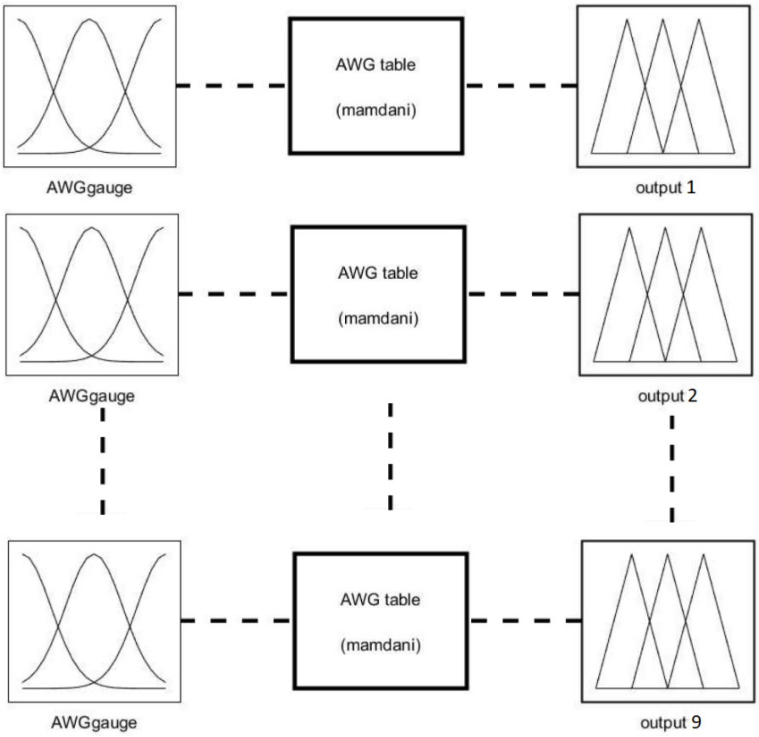


Fig. 2. Divided Inference System for AWG table.

There is an alternative that can compute the required fuzzy sets and its numerical parameters for the AWG table. The Batch Least Square Mamdani (LSM) is a technique used in fuzzy control and system identification that can automatically tune an inference system by calculating the inverse of its input parameters the same way that it is done in the Least Square Estimator (LSE).

Fuzzy inference systems have been spread in several areas since its very beginning in 1965, today there still appearing several applications because of its capability for mimic human reasoning using expert knowledge. There are several papers where fuzzy systems have been applied in the recent years, they have been used in image enhancement in [6], image compression analysis in [7], risk evaluation and periodization in industries in [8], estimating the length-weight relationship of fishes in [9], for failure analyzing in automobile industry in [10], in the assessment of power transformers in [11], among other several applications.

In this paper it is proposed the use of LSM for mimic tables in computer programs specifically applied in the AWG table and the user can select an acceptable error in the operation so that the system complexity be related with this error.

Table 1. AWG gauge table [5].

AWG gauge	Conductor Diameter (in)	Conductor Diameter (mm)	Conductor cross section in mm^2	Ohms per 1000 ft.	Ohms per km	Max. Amps for chassis wiring	Max. Amps for power transmission	Maximum frequency for 100% skin depth for solid conductor copper	Breaking force soft annealed Cu 37000 PSI (lbs)
0	0.3249	8.2525	53.5	0.0983	0.322424	245	150	250 Hz	3060
1	0.2893	7.3482	42.4	0.1239	0.406392	211	119	325 Hz	2430
2	0.2576	6.5430	33.6	0.1563	0.512664	181	94	410 Hz	1930
3	0.2294	5.8268	26.7	0.197	0.64616	158	75	500 Hz	1530
4	0.2043	5.1892	21.1	0.2485	0.81508	135	60	650 Hz	1210
5	0.1819	4.6203	16.8	0.3133	1.02436	118	47	810 Hz	960
6	0.162	4.1148	13.3	0.3951	1.29393	101	37	1100 Hz	760
7	0.1443	3.6652	10.6	0.4982	1.63898	89	30	1300 Hz	605
8	0.1285	3.2639	8.37	0.6282	2.05607	73	24	1650 Hz	480
9	0.1144	2.9058	6.63	0.7921	2.59567	64	19	2050 Hz	380
10	0.1019	2.5883	5.26	0.9989	3.2772	55	15	2600 Hz	314
11	0.0907	2.3038	4.17	1.26	4.1339	47	12	3200 Hz	249
12	0.0808	2.0523	3.31	1.588	5.21	41	9.3	4150 Hz	197
13	0.072	1.8288	2.63	2.003	6.572	35	7.4	5300 Hz	150
14	0.0641	1.6281	2.08	2.525	8.284	32	5.9	6700 Hz	119
15	0.0571	1.4503	1.65	3.184	10.45	28	4.7	8250 Hz	94
16	0.0508	1.2903	1.31	4.016	13.18	22	3.7	11 k Hz	75
17	0.0453	1.1506	1.04	5.064	16.614	19	2.9	13 k Hz	59
18	0.0403	1.0236	0.823	6.385	20.948	16	2.3	17 kHz	47
19	0.0359	0.91186	0.653	8.051	26.414	14	1.8	21 kHz	37
20	0.032	0.8128	0.519	10.15	33.301	11	1.5	27 kHz	29
21	0.0285	0.7239	0.412	12.8	41.995	9	1.2	33 kHz	23
22	0.0253	0.64516	0.327	16.14	52.953	7	0.92	42 kHz	18
23	0.0226	0.57404	0.259	20.36	66.798	4.7	0.729	53 kHz	14.5
24	0.0201	0.51054	0.205	25.67	84.219	3.5	0.577	68 kHz	11.5
25	0.0179	0.45466	0.162	32.37	106.201	2.7	0.457	85 kHz	9
26	0.0159	0.40386	0.128	40.81	133.891	2.2	0.361	107 kHz	7.2
27	0.0142	0.36068	0.102	51.47	168.865	1.7	0.288	130 kHz	5.5
28	0.0126	0.32004	0.080	64.9	212.927	1.4	0.226	170 kHz	4.5
29	0.0113	0.28702	0.0647	81.83	268.471	1.2	0.182	210 kHz	3.6
30	0.01	0.254	0.0507	103.2	338.583	0.86	0.142	270 kHz	2.75
31	0.0089	0.22606	0.0401	130.1	426.837	0.7	0.113	340 kHz	2.25
32	0.008	0.2032	0.0324	164.1	538.386	0.53	0.091	430 kHz	1.8

2 Theoretical Framework

Fuzzy inference systems describe systems that reason using inference rules i.e. IF-THEN rules, and its general structure is given in Fig. 3.

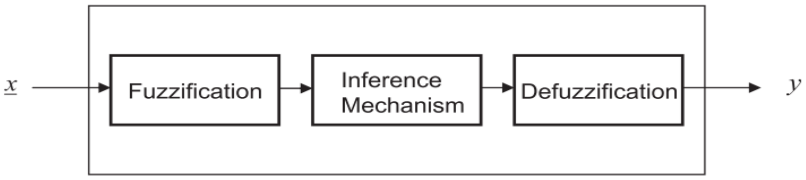


Fig. 3. General structure of fuzzy systems [3].

Fuzzification is the part where a scalar input data is transformed to fuzzy data by identifying its corresponding membership value (μ) in range $[0,1]$ for the input sets. Fuzzy input sets are obtained using mathematical functions called membership functions, which determine the membership grade for a certain premise with the input data, for example in the LSM Gaussian membership functions are used with equation (1), corresponding with the shape shown in Fig. 4 [3]:

$$\mu(x) = e^{-\frac{1}{2}\left(\frac{x-c}{\sigma}\right)^2} . \tag{1}$$

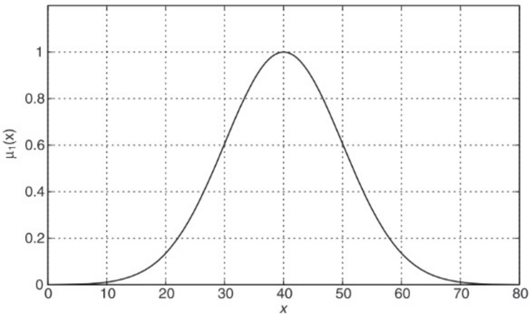


Fig. 4. Gaussian membership function.

The Gaussian membership function require 2 numerical parameters that must be selected for adjusting the fuzzy set according to given problem, the parameter c variates the function center while the parameter σ variates its spread [3].

Mamdani fuzzy inference system connects al premises using max and min operations related to OR and AND connectors, these operations are performed with the membership values of the fuzzy sets [3].

Mamdani fuzzy inference system conclude using as mechanism of inference min operation between the input sets and the output sets according to equation (2):

$$\min(\mu_i(x), \mu_o(y)) , \tag{2}$$

where $\mu_i(x)$ is the membership of the input set with $x \in X$ the input universe and $\mu_i(y)$ is the membership of the output set with $y \in Y$ the output universe, depending from the number of inputs and outputs it is possible to have more than one input and output in their respective universes.

Defuzzification of Mamdani inference system is performed with Center of Gravity (COG), which is performed according to equation (3):

$$y = \frac{\sum_{i=1}^{n} y \cdot \mu(y)}{\sum_{i=1}^{n} \mu(\mathrm{y})}, \tag{3}$$

where n is the number of output elements in the output set, determined with the aggregation operation which in Mamdani systems is the max operation.

Least square estimation finds the parameters that minimize the square difference between estimated and obtained parameters but applied to an inference system there could be expressed as the contribution of the input sets to the output sets, equation (4):

$$y(j) = \phi_1(j)\theta_1 + \phi_2(j)\theta_2 + \phi_3(j)\theta_3 + \ldots + \phi_R(j)\theta_R, \tag{4}$$

where ϕ_R are the R parameters known corresponding to fuzzy sets and θ_R are the R constant parameters unknown that are adjusted for guarantee a desired response so if the system can be written as

$$y(j) = \Phi^T(j)\Theta, \tag{5}$$

with Θ containing al the unknown θ constant parameters that optimize the fuzzy inference system for its desired response.

But with Batch Least Square can be computed like described in [3] with equation (6):

$$\Theta = \left(\Phi^T \Phi\right)^{-1} \Phi^T Y, \tag{6}$$

where Y contain all the desired responses or values in a table for the proposed application in this work.

Following Mamdani inference systems defuzzification can be obtained with equation (3), but if all parameters are separated before summing its numerator with the same denominator then a new expression can be generalized as in equation (7):

$$\xi_i(x) = \frac{\mu_i(x)}{\sum_{i=1}^{R} \mu_i(x)}, \tag{7}$$

where ξ_i is the fuzzy basis function of the i rule like described in [3], so $\Phi^T = [\xi_1(\mathrm{x})\ldots\xi_R(\mathrm{x})]$ and then will compute the desired system by obtaining the Θ parameters [3].

3 Methodology

The AWG table is separated in fuzzy systems like in Fig. 2 then every fuzzy system is trained using the LSM formulation where Θ parameters are determined and the number of required parameters for satisfying a specific desired error in the generation of the program.

After that table and systems are separated then the desired values for the output systems are assigned to Y for a given column like in equation (8), with i describing de corresponding column and system:

$$Y_i = c_i \; . \tag{8}$$

Then all systems compute the Φ_i^T matrix using equation (7) and the constant parameters required for mimic the columns in the AWG table using equation (6).

To modify the complexity of the system i.e. the quantity of required input fuzzy sets the new output response is calculated with equation (5) and then relative error is calculated with equation (9):

$$e = \frac{\hat{Y} - Y}{Y} \; , \tag{9}$$

where $\hat{Y}$ is the obtained response with the Mamdani fuzzy system adjusted with LSM.

The system complexity it is initialized in 1and then is increased until $e < e_d$ with e_d given as the desired error.

4 Results

In this section are presented the results obtained after use LSM for determining every fuzzy system related to its corresponding column so that the union of all 9 fuzzy systems replace the AWG table.

The AWG table can be extended to more values but in this program was used with the 33 different wires shown in the Table 1. The obtained Θ_i matrixes with the coefficients for each fuzzy set in the i system are shown in equations 10 to 18, all of then satisfying $e < 0.01$ i.e. a relative error of 1%.

$$\begin{aligned}\Theta_1 = [&0.3330, 0.2531, 0.2039, 0.1605, 0.1278, 0.1011, 0.0803, 0.0637\\ &0.0505, 0.0400, 0.0318, 0.0251, 0.0200, 0.0158, 0.0125, 0.0100, 0.0077]\end{aligned} \tag{10}$$

$$\begin{aligned}\Theta_2 = [&8.4579, 6.4276, 5.1797, 4.0775, 3.2459, 2.5688, 2.0390, 1.6177\\ &1.2822, 1.0171, 0.8070, 0.6410, 0.5072, 0.4015, 0.3186, 0.2536, 0.1964]\end{aligned} \tag{11}$$

$$\begin{aligned}\Theta_3 = [&55.8069, 31.6841, 20.9408, 12.8395, 8.2121, 5.0898, 3.2347, 2.0220\\ &1.2767, 0.8010, 0.5052, 0.3186, 0.1994, 0.1246, \; 0.0784, 0.0499, 0.0299]\end{aligned} \tag{12}$$

$$\begin{aligned}\Theta_4 = [&0.0912, 0.1532, 0.2415, 0.3847, 0.6114, 0.9723, 1.5467, 2.4565, 3.9121,\\ &6.2076, \; 9.9024, 15.6516, 25.1509, 39.2994, 64.3206, 97.2561, 171.1798]\end{aligned} \tag{13}$$

$$\begin{aligned}\Theta_5 = [&0.2989, 0.5034, 0.7900, 1.2624, 2.0032, 3.1890, 5.0750, 8.0596, 12.8403\\ &20.3640, 32.4893, 51.3503, 82.5162, 128.9349, 211.0261, 319.0815, 561.6136]'\end{aligned} \tag{14}$$

$$\begin{aligned}\Theta_6 = \; [&252.1789, 193.0706, 159.1440, 128.1148, 103.1033, 84.9516, 64.1592, 54.0400, 42.3556, 34.3010\\ &30.8227, \; 21.5508, 17.5140, 14.1156, 10.0950, 7.4230, 4.0581, 2.8224, 2.1002, 1.4242, 1.1967, 0.7189, 0.5163]'\end{aligned} \tag{15}$$

$$\begin{aligned}\Theta_7 &= [156.6304, 88.4998, 59.6106, 35.3790, 23.8245, 14.5016,\ 9.1445, 5.7173\\ &3.6311, 2.1873, 1.4711, 0.9014, 0.5585, 0.3530, 0.2210, 0.1397, 0.0843]'\end{aligned} \tag{16}$$

$$\begin{aligned}\Theta_8 &= [235.51,353.17,461.54,632.14,889.72,1238.0,1615.4,2191.3,2952.7,4065.0,5738.2,7455.9,10985.0\\ &13757.0,19507.0,26842.0,34652.0,48893.0,66938.0,91482.0,119144.0,168411.0,223377.0,309644.0,445755.0]'\end{aligned} \tag{17}$$

$$\begin{aligned}\Theta_9 &= [\ 3191.1, 1821.7, 1198.2, 733.9, 466.6, 304.7, 192.0\\ &113.0, 73.7, 45.3, 28.4, 17.5, 11.3, 6.9, 4.4, 2.7, 1.7]'\end{aligned} \tag{18}$$

The structures variates per column according to the number of Gaussians required per inference machine, as shown in table 2.

Table 2. Fuzzy structure characteristics after adapting to AWG table.

Name of column in table	Number of Gaussian functions	Desired Error	Obtained Error	Computing time (seconds)
Conductor Diameter (in)	17	1%	0.1%	0.084453
Conductor Diameter (mm)	17	1%	0.1%	0.078887
Conductor cross section in mm^2	17	1%	0.5%	0.073799
Ohms per 1000 ft.	17	1%	0.5%	0.075804
Ohms per km	17	1%	0.5%	0.081492
Max. Amps for chassis wiring	23	1%	0.9%	0.170864
Max. Amps for power transmission	17	1%	0.7%	0.175134
Maximum frequency for 100% skin depth for solid conductor copper	25	1%	0.9%	0.202492
Breaking force soft annealed Cu 37000 PSI	17	1%	0.9%	0.077608

Figures 5 to 13 show the comparison between the plot of real data in the AWG table and the obtained using the fuzzy system trained with LSM and the obtained matrixes $A_{1\ldots9}$.

The results were computed in a maximum time of 0.277147 seconds for the column Maximum amps for power transmission (A) for AWG gauge and in a minimum time of 0.095222 seconds for the column Conductor diameter (mm) for AWG gauge.

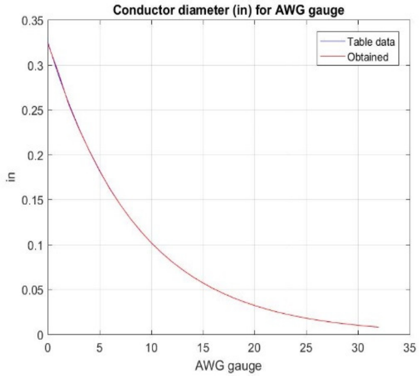


Fig. 5. Fuzzy system vs table column 1.

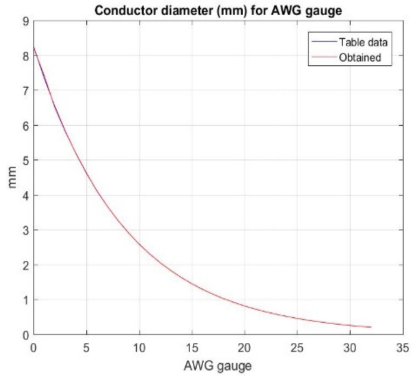


Fig. 6. Fuzzy system vs table column 2.

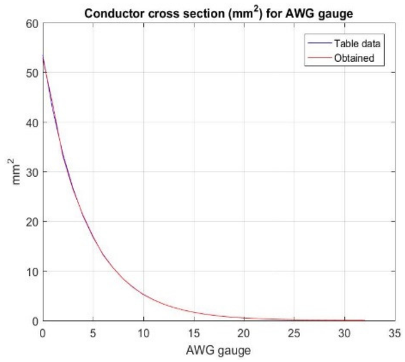


Fig. 7. Fuzzy system vs table column 3.

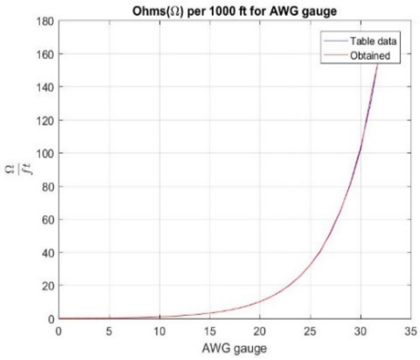


Fig. 8. Fuzzy system vs table column 4.

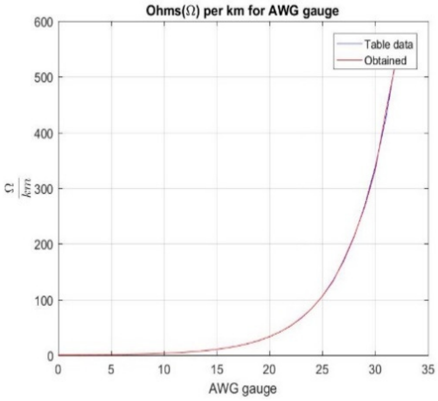


Fig. 9. Fuzzy system vs table column 5.

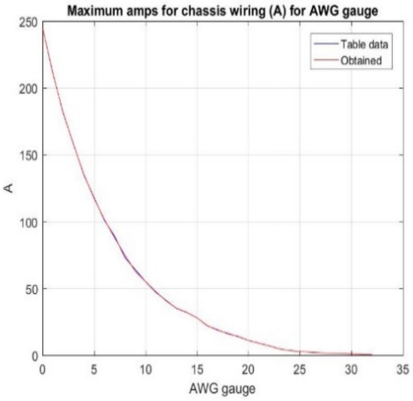


Fig. 10. Fuzzy system vs table column 6.

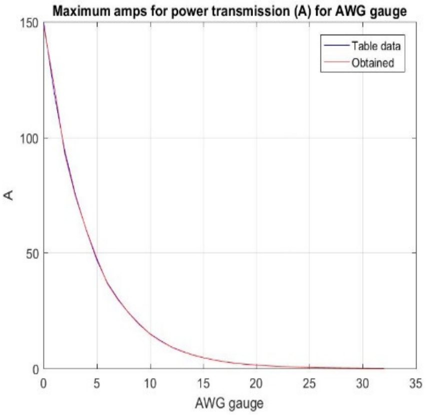


Fig. 11. Fuzzy system vs table column 7.

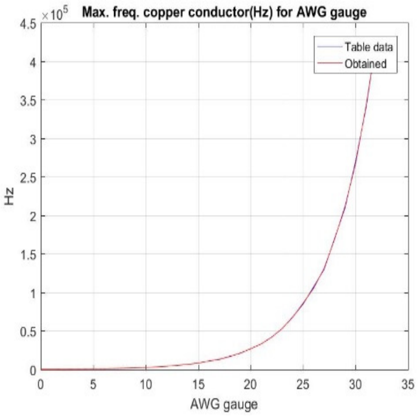


Fig. 12. Fuzzy system vs table column 8.

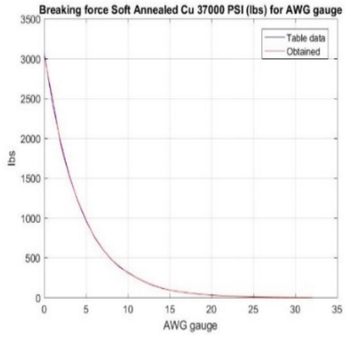


Fig. 13. Fuzzy system vs table column 9.

5 Conclusions

In this paper is presented an alternative to generate programs automatically from huge tables using Batch Least Square Mamdani Inference Systems with the desired error specifically applied in the AWG table, where time required for computing is less than 0.202492 seconds per column in tables, depending mainly from the spent time in the inverse equation (6). These results are achieved with a relative error minor than 1%.

References

1. Montes Rivera, M., Aguilar Justo, M.O., Ochoa Zezzatti, A.: (2016). Equations for Describing Behavior Tables in Thermodynamics Using Genetic Programming: Synthesizing the Saturated Water and Steam Table. Research in Computing Science 122, 9–23 (2016)

2. Pollack, S.L.: Conversion of Limitedl-Entry Decision Tables to Computer Programs. Comunications of the ACM, 677–682 (1965)
3. Lilly, J.H.: Fuzzy Control and Identification. Hoboken, New Jersey: John Wiley & Sons (2010)
4. Passino, K.M.: Fuzzy Control. Menlo Park, California: Addison-Wesley (1998)
5. Lund Instrument Engineering: Power Stream . Retrieved from Wire Gauge and Current Limits Including Skin Depth and Strength: http://www.powerstream.com/Wire_Size.htm (2018)
6. Chithrakshi, T.H.: Image Compression Using Fuzzy Enhancement. International Journal of Engineering and Advanced Technology, 348–351 (2014)
7. Kumar Gangwar, R., Kumar, M., Jaiswal, A., Saxena, R.: Performance Analysis of Image Compression Using Fuzzy Logic Algorithm. Signal & Image Processing: An International Journal, 73–80 (2014)
8. Wen Kerk, Y., Meng Tay, K., Peng Lim, C.: An Analytical Interval Fuzzy Inference System for Risk Evaluation and Prioritization in Failure Mode and Effect Analysis. IEEE Systems Journal, 1589–1600 (2017)
9. Bitar, S.D., Campos, C.P., Freitas, C.E.: (2016). Applying fuzzy logic to estimate the parameters of the length-weight relationship. Braz. J. Biol, 611–618 (2016)
10. Geramian, A., Mehregan, M.R., Garousi Mokhtarzadeh, N., Hemmati, M.: Fuzzy inference system application for failure analyzing in automobile industry. International Journal of Quality & Reliability Management, 1493–1507 (2017)
11. Chacón, D.P., Lata, J.P., Medina, R.D. (2017). Health Index Assessment for Power Transformers with Thermal Upgraded Paper up to 230kV, Using Fuzzy Inference Part II: A Sensibility Analysis. In: International Caribbean Conference on Devices, Circuits and Systems (ICCDCS), 109–112 (2017)

ISSN 1870-4069

Blurring Northeast Mexican Societies: An Approach to Cultural Capital and Results of the PISA Test

Mónica Mendirichaga Pérez-Maldonado[1], Alberto Ochoa-Zezzatti[1],
Aida Yarira Reyes Escalante[1], Roberto Mendirichaga-Dalzell[2], Sandra Bustillos[1]

[1] Universidad Autónoma de Ciudad Juárez (UACJ), Mexico
[2] Universidad de Monterrey (UDEM), Mexico

Abstract. The present research shows, from the results of the PISA Test of the years 2012-2015, the difference that exists in the societies of the northeast, comparing the state of Nuevo León with the states of Coahuila, San Luis Potosí and Tamaulipas. It seeks to evaluate the field of reading in each of these regions, coupled with the results of mathematics and science. Taking from the PISA Test 38,251 records, those of these four states are included, giving a total of 4,764 records. For the above, it is basic what is related to the reading process and the existence of libraries, as well as other exogenous variables, all this done through the software of the Weka 3.9.2 program.

Keywords: cultural capital, PISA test, northeastern Mexico, incentive of reading, use of libraries.

1 Introduction

Understanding adequately the representation of knowledge in a group of societies where children of 15 years old need to have a job, is basic to establish educational policies associated with this group of consolidated economies that make up the OECD (Organization for Economic Cooperation and Development). When we talk about knowledge, it refers to the most important aspect or component of the current civilization, because it is more than natural resources, money and other goods. Nowadays, whoever possesses the knowledge has the key of power, hence this key factor for the future of Mexico and the world is included. The PISA test evaluates not only the 43 countries that make up the OECD, but also political entities from other countries, such as: Macao, Wales or even the Basque Country; that is to say, that the developed countries can not disassociate themselves from the underdeveloped ones. In fact, today new economic-political alliances have been formed that no longer obey the traditional definitions of groups and alliances.

Data Mining allows you to identify hidden patterns in large groups of information. It is a new technology, which makes a great contribution to the social and economic-administrative sciences. There are several studies conducted using data mining such as the study of: Data Mining Applied for the Identification of Risk Factors in Students [1], Evaluation of CENEVAL Admission Surveyed Parameters for Students that are Candidates to Enter the Higher Education, ITP Study Case [2], Modeling Students' Dropout in Mexican Universities [3].

Through the Weka program 3.9.2. the data about the PISA Test are analyzed and compared to arrive at specific conclusions. The methodology that is carried out in the present research is applied to the analysis of the results of the PISA Test of the years 2012-2015 in the northeast societies, comparing the state of Nuevo León with the states of Coahuila, San Luis Potosí and Tamaulipas. The Pygmalion effect consists in the change of behavior of a group that shares similar induced characteristics, by an expectation of one of them; this phenomenon has been documented since the sixties, but it appears here with some details of the investigation based on accurate information, because it is by means of the information obtained in the evaluation of the PISA Test that it is possible to arrive at the required information. [4].

2 The Northeast Mexican

It is known as the Mexican northeast societies (mns), the integration of the states of San Luis Potosí, Coahuila, Tamaulipas and Nuevo León. Before the Treaty of Guadalupe-Hidalgo in 1848, Texas was part of this subregion and, although politically it is no longer part of the American Union, in some way it is from the cultural point of view part of Mexico. San Luis Potosí is the most consolidated entity geographically, culturally and politically. It had a great miscegenation.

During the viceroyalty, it played an important role in advancing the New Spain model to the north, through evangelization, military development, economy and education. Presents several areas, such as: The Altiplano, the Middle region and the Huasteca. In the Altiplano region, a large part of it´s history "seems to be linked to the discovery and exploitation of subsurface resources", while the Huasteca and a good part of the Middle region are mainly agricultural regions. At "El Ébano" the first oil well of Mexico was located. Much of the problem of emigration to more distant population centers or even abroad, is due to the hoarding of the means of production and the use and possession of the land, in addition to the problem of water and desertification [5]. From San Luis Potosí and Zacatecas, culture passed to Nueva Extremadura or Coahuila. Five subregions are located here: Saltillo, Monclova, Parras-La Laguna, the Carboniferous Region and Guerrero, each with it´s particular vocation and uniqueness. Its demographic growth was slow and gradual. It´s culture can not be understood without the Tlaxcalan presence. "As part of the modernization of the Coahuila society, since the second half of the 19th century education was one of the aspects that were most encouraged" [6].

It should not be forgotten that there existed at the end of this century and at the beginning of the 20th century a fundamental economic axis that came from Chihuahua, passed through Durango, reached La Laguna, then followed Saltillo, and continued in Monterrey, to finish in Laredo. Tamaulipas, long territory that goes from the border with the United States to Veracruz, was founded in 1700 by José de Escandón and the first settlers, who came mostly from the New Kingdom of León. Its native population were: olives, reeds, janambres and huastecos. Until the Independence, the Nueva Santander or Tamaulipas depended militarily and economically on the intendency of San Luis Potosí. The Franciscan missions contributed to the pacification. From the 19th century, the main population centers were: Aguayo (later Ciudad Victoria), El Mante, Tampico, Tula, Camargo, Matamoros and Reynosa. Filibusterism, rustling and concentration of capital hindered integration. Foreign investment in the field and oil explotation were factors that disrupted traditional sources of wealth. Then came the Revolution, which was factional fighting. In education and culture, Tamaulipas saw it´s awakening with the Obregonist reconstruction and the program of José Vasconcelos in SEP. "With the support of Portes Gil, President Cárdenas began the stage of institutional incorporation [...]".

They excelled in secondary high education, the San Juan Institute, in Matamoros, and the Scientific and Literary Institute, in Ciudad Victoria. The demographic dispersion and the ruggedness of the Sierra Madre made literacy difficult. At the end of the 20th century, the maquila industry strongly influenced the economic takeoff of Tamaulipas, which in some way stopped emigration (*ibid* p.62). Finally, the old New Kingdom of León, then Nuevo León after Independence, is the fourth entity that integrates this northeast. It had a primary economy based on agriculture, livestock and trade, which was modified by the industrialization of the mid-twentieth century, the result of the Reyista era and postrevolutionary governments. Formal education was present with institutions such as the College-Seminar, the State Civil College, the Escuela Normal para Profesores and, already in the twentieth century, with the University of Nuevo León (UNL, in Spanish, and now UANL), the Tecnológico of Monterrey (ITESM, in spanish) and other universities.

The particularity of this entity is that it concentrates most of its population in the capital city, not registering other notable poles of development. It turns out to be the most important entity in the north, basing most of its economy on services. It maintains an organic education system that is close to that of its neighbor Texas and other European and Asian countries [7].

3 Cultural Capital: Reproduction and PISA Test

By *cultural capital* is usually understood the intangible that occurs through generations in a particular country society; the result of many years of work, study, savings and effort. In the case of México, this cultural capital is integrated with the rich Prehispanic Past; the three centuries of the Viceroyalty; what is produced in Independence; the addition in the Reforma and the Porfiriato; what was generated in the Revolution and Reconstruction, plus the whole maturity process of the 20th century, to enter a 21th century full of challenges, but also of opportunities. Annette Lareau and Jessica McCrory, in the chapter entitled “Class, cultural capital and institutions: the case of families and schools”, warn that in the United States, “Americans are convinced that by the power of individualism and only through it they forge the opportunities and the options reached”.

The authors conclude that, in the case of the United States, the most part that comes from the cultural capital is that "a better education is a universally accepted value, in relation to parents and the education of their children [8]. According to the aforementioned authors, one should consider the approach of Pierre Bourdieu, who "uses the concept of *cultural capital* to explore not only how the basic social class produces individuals with different interactional resources, but how these resources can be used to produce different social benefits [9]." Bourdieau's approach, carried out together with

Claude Passeron at the beginning of the fifties of the last century, with something that they called *reproduction,* derived from Marxist theory in it´s concept of surplus value and appropiation, has been worked by many authors and applied to education, reading, art, culture, and other aspects of social life. Sandrine García and Franck Poupeau, in the chapter: "The measurement of 'school democratization'. Notes on the sociological uses of statistical indicators", which is part of the book Pierre Boudieau. *Symbolic capital and social magic,* by Isabel Jiménez, coordinator, possess a series of elements that have to do tangentially with our work on reading process, school, the PISA test and community libraries.

For García and Poupeau, the statistical indicators of school democratization require epistemological surveillance. Taking about the case of four French business schools, "the great prestigious schools have opened very little to the popular classes", but from the above can not easily conclude a direct relationship between social class and professional success. We should also consider what is related to the pedagogical teams and the effectiveness of the school or establishment, as indicated by the aforementioned authors [10]. It is necessary to see the result of the PISA test in México during the last years, and specifically in the northeast; to try to establish if there is any direct relationship between reading and academic success: between the effectiveness of reading done in community libraries and school success; between social class and professional success; between the parents' economic income and the children's school results, which may be far from the objective of this paper.

4 Incentive of Reading Process and Using Libraries

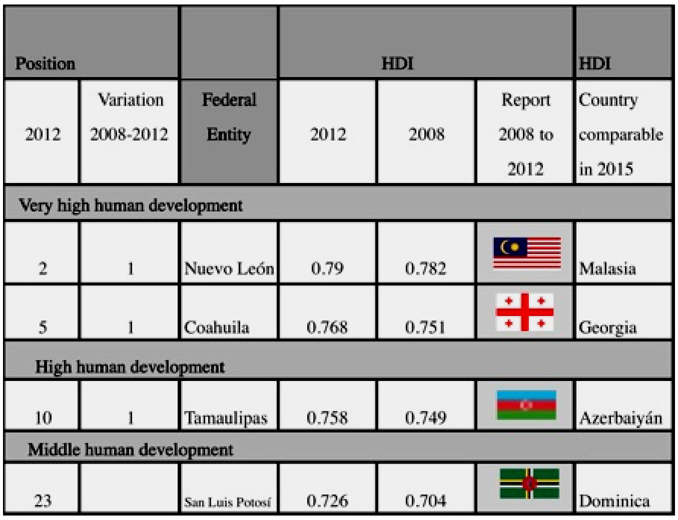
It can be read in many different ways: in a traditional way, in the printed book, in the newspaper, in the magazine, in the comic; and now, also on the ipad, the internet, the cell phone or the kindle. What percentage of Mexicans do it in one way or another is something that has yet to be clarified, since the research work on it does not define it completely. But if we ask to the current university students what percentage of what they read do it in a conventional or traditional way, it will hardly exceed 20 points what is read in print. Juan Domingo Argüelles considers that "[...] they read books on the screen who also read them on paper, and others do not read them on any medium" [11]. What should be a desideratum is that our children, youth, adults and the elderly read the best of universal literature, in the fields of science, arts and humanities. So, if this seems to be the reality, where the most powerful sources of information are television and social networks, why return to the subject of libraries and, specifically, community libraries? Because we consider that these represent a very good option for the intellectual, ethical and aesthetic elevation of large masses of unemployed, workers, housewives, students and children who find it very difficult to acquire books or enter to the traditional internet or its databases, which have become unreachable in their price or difficult to locate.

In her Master's Thesis about libraries, Mónica Mendirichaga Pérez-Maldonado says that it would be desirable for there to have an active participation from government and civil society. Also, in terms of facilities, that they adequate; that there was an open and ready staff to meet the needs of users, texts that would invite reading and specialized materials; likewise, computers that had access to the information superhighway [12], as benefits the vision of what Juan José Arreola, once referred to as schools, would describe as "true centers of exciting experiences"[13]. That is to say, the library must be a cultural and artistic center, where the librarian becomes its main promoter, given that these libraries are located in rural areas or in depressed urban areas. According to what we have been able to analyze in this work, the PISA test has a direct relationship with the cultural capital of the region, with the socioeconomic and cultural level of the parents of these children, and with the possibilities of economic and social advancement. Graduates of technical schools and university centers in the country, which brings us to the next point of this research.

5 PISA Test (Program for International Student Assessment)

The Human Development Index (HDI) is an indicator of the achievements obtained in the fundamental dimensions of human development, which are: a long and healthy life, acquiring knowledge and living a dignified life. They are the indices of each of the three dimensions. It can be said that the HDI of human development is very high in Nuevo León, followed by Coahuila, Tamaulipas is high in human development and in the middle human development is San Luis Potosí. (See Table 1.)

Table 1. Own elaboration. Inspired by the HDI Table (Human Development Index) of the states of Nuevo León, Coahuila, Tamaulipas and San Luis Potosí [14].

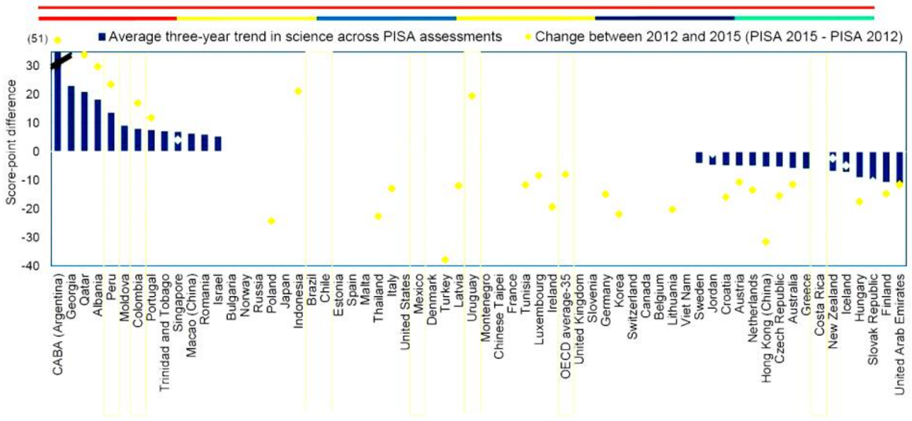


Position			HDI			HDI
2012	Variation 2008-2012	Federal Entity	2012	2008	Report 2008 to 2012	Country comparable in 2015
Very high human development						
2	1	Nuevo León	0.79	0.782		Malasia
5	1	Coahuila	0.768	0.751		Georgia
High human development						
10	1	Tamaulipas	0.758	0.749		Azerbaiyán
Middle human development						
23		San Luis Potosí	0.726	0.704		Dominica

In the results of the 2012 PISA educational evaluation (in points), Mexico in the area of reading has 424 (No. 52), then in science 415 (No. 55) and finally in mathematics 413 (No. 53). In the first three places in the world in reading is Shanghai, China 570 (No. 1), Hong Kong 545 (No. 2) and Singapore 542 (No. 3). In science is Shanghai, China 580 (No. 1), Hong Kong, China 555 (No. 2), Singapore 551 (No. 3). In mathematics there are Shanghai, China 613 (No. 1), Singapore 573 (No. 2) and Hong Kong, China 561 (No. 3). [15]. (See Table 2.)

Average trend of performance in science since 2006: Mexico is at zero, while the United Arab Emirates are those who are located in the highest place.

Table 2. Average performance trend in science since 2006. [16]



5.1 Questions of Great Relevance in the PISA Test

Within the International Program for the Evaluation of the OECD Student (Organization for Economic Cooperation and Development) PISA 2009, the Student Questionnaire includes those considered most important. The sections chosen for this topic are those related to: family environment, your reading activities, the time dedicated to learning, school environment time, your Spanish classes, Libraries, your reading strategies and text comprehension and your educational career.

Section 2: Your family environment. It's about your family and your house. Some of the questions refer to the father, the mother, or the people who act as equivalents; for example, sponsors, uncles, brothers, stepparents, adoptive parents, in anothers

Question 11. Does your mother have any of the following certificates?

Answer: (a) PhD; (b) bachelor's degree; university degree; technological degree; specialization or expertise; (c) superior technician.

Section 3: Your reading activities. Reading activities outside of your school

Question 23. How much time do you spend reading for pleasure?

Answer: I do not read for entertainment; 30 minutes or less a day; More than 30 minutes, but less than 60 minutes a day; From 1 to 2 hours a day; More than 2 hours a day.

Section 4: Time dedicated to learn

Question 31. What type of classes do you currently attend outside of school hours? (This question refers only to the classes of the subjects you study at school and those you attend outside of school hours). Selecting: yes or no.

Answer: (a) High performance classes in Spanish (Reading and Writing Workshop, Language and Literature); (b) High performance classes in Mathematics; (c) High performance classes in Science (Biology, Physics, Chemistry or Earth Sciences); (d) High performance classes in other subjects; (e) Regularization classes in Spanish (Reading and Writing Workshop, Language and Literature); (f) Regularization classes in Mathematics; (g) Regularization classes in Sciences (Biology, Physics, Chemistry or Earth Sciences); (h) Regularization classes in other subjects; (i) Classes to improve your ability to study.

Section 5: School environment time

Question 34. How much do you agree or disagree with each of the following statements about teachers in your school? (Mark only one option of each row). Selecting: totally agree, agree, disagree and totally disagree.

Answer: (a) I get on well with most teachers; (b) Most teachers are interested in my well-being; (c) Most of my professors really listen to what I have to say; (d) If I need extra help, I receive it from my teachers; (e) Most of my teachers treat me fairly.

Section 6: Your Spanish classes

Question 37. In your Spanish classes, (Reading and Writing Workshop, Language and Literature, etc.) how often does the teacher do the following things? (Mark only one option in each row). Selected: in all classes; in most classes; in some classes; never or almost never.

Answer: (a) The teacher asks his students to explain the meaning of a text; (b) The teacher's questions stimulate students to gain a better understanding of what they read; (e) The teacher encourages students to express their opinion about a text; (f) The teacher helps the students relate the stories they read with their lives; (g) The teacher shows that the information in the texts is constructed from what the students know.

Section 7: Libraries

(In this section we will ask about the libraries, it could be the school library or those outside of your school). Selecting: several times a week; sometimes in the month; approx. once a month; sometimes a month and never.

Question 39. For the following activities, how often do you visit the library at your school?

Answer: (a) Ask borrowed books to read for pleasure; (b) Request borrowed books to do school work; (c) Work on assignments, course assignments or research papers; (d) Read magazines or newspapers; (e) Read books for pleasure; (f) Learning things that are not related to the course, such as: sports, hobbies, people or music; (g) Use the internet.

Section 8: Your reading strategies and text comprehension

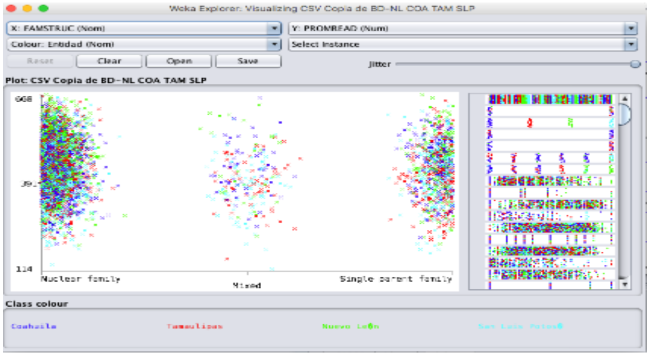
Question 42. Purpose of the reading: You have just read a two-page long and a little difficult text about fluctuations in the water level of a lake in Africa. You have to write a summary. (How do you rate the usefulness of the following strategies for writing a summary of this two-page text?). Selecting from: (6) very useful to (1) for nothing useful.

Answer: (a) I write a summary. Then I make sure that each paragraph is covered in the summary, because the content of each paragraph must be included; (b) I try to copy exactly as many sentences as possible; (c) Before writing the summary, I read the text as many times as possible; (d) I check carefully that the most important data of the text are included in the summary; (e) I read the entire text, underlining the most important sentences. Then I write them in my own words in summary form.

Section 9: Your educational career. (In this section you will be asked about the different aspects related to your school experience)
Question 47. Which of the following educational levels do you expect to finish? (Mark all that apply to your case).
Answer: (a) Junior High (escuela secundaria, in spanish); (b) Technical Professional; (c) High School (Bachillerato in spanish), ColBach (Bachelors), Vocational (Technical Education, post-secondary, Polytechnic Institute), Conalep (Technical Studies), etc.); (d) Higher University Technician; (e) Bachelor, Master Degree or Doctorate (PhD Degree).

6 PISA Test (Program for International Student Assessment)

Test 1
X: FAMSTRUC (Family structure)
Y: PROMREAD (Reading Average)
-Entity (Coahuila, Tamaulipas, Nuevo León and San Luis Potosí)
According to the analysis carried out, it has been possible to know the importance of the student having a nuclear family, since it will make better use of reading and this will be reflected in the average; secondly, we have the children who are cared for either by their mother or by their father; and finally there are students who are part of a mixed family. As for the results on the use of reading, we have first Nuevo León, then Coahuila, followed by Tamaulipas and, finally, San Luis Potosí.



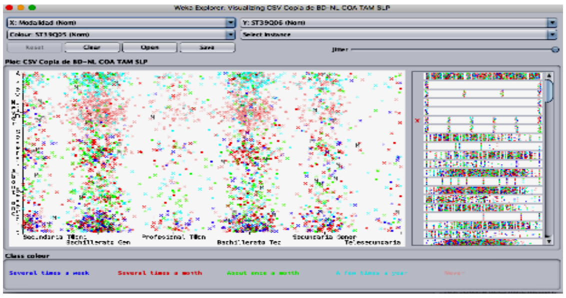
Weka Explorer: Visualizing CSV Copia de BD-NL COA TAM SLP

X: FAMSTRUC (Nom)
Y: PROMREAD (Num)
Colour: Entidad (Nom)
Select Instance

Plot: CSV Copia de BD-NL COA TAM SLP

Test 2
X: Modality (Technical Secondary, General Baccalaureate, Technical Professional, Technological Baccalaureate, General Secondary and Telesecundaria (Distance Education).

Y: ST30Q06 (Learn things that are not related to the course such as sports, hobbies, people or music). Question No. 39.
-ST39Q05 (Read books for pleasure). Question No. 30.
In the Baccalaureate level it has been possible to perceive that it is there where the student acquires greater knowledge and skills that are not necessarily related to the course. Such is the case of independent reading of school texts; the practice of a sport, playing a musical instrument, being part of a theater or dance group; to practice a hobby: to go to the cinema, to the theater, to concerts or dance shows, among a host of other possibilities.

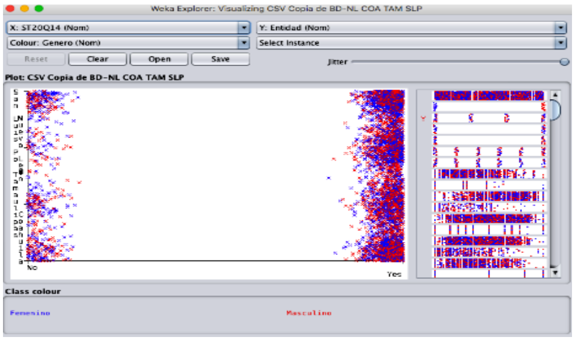


Weka Explorer: Visualizing CSV Copia de BD-NL COA TAM SLP

X: Modalidad (Nom)
Y: ST39Q06 (Nom)
Colour: ST39Q05 (Nom)
Select Instance

Plot: CSV Copia de BD-NL COA TAM SLP

Test 3
X: ST20Q14 (Among the following levels of study, what is the highest level of education your father arrived at?). Question No. 14.
And: Entity (Coahuila, Tamaulipas, Nuevo León and San Luis Potosí).
-Gender (Female and Male).
Following the review of the results of the PISA test in relation to this issue, it has been found that there is a direct influence between the following two variables: if the male child has a father who studied, whether primary, secondary, high school, career, master or doctorate, this will lead to the likelihood that your child will aspire to study to the same level as his father did. Regarding the above, it is known that "the example drags" and that the child will have greater motivation if he expects to become equal or better than his parent. For this reason, a higher level of studies of the father, greater literacy skills will develop in the child, who in turn will impact on a high vocabulary, a higher capacity for synthesis and greater ease to perform processes related to the reflection and problem solving.



Weka Explorer: Visualizing CSV Copia de BD-NL COA TAM SLP

X: ST20Q14 (Nom)
Y: Entidad (Nom)
Colour: Genero (Nom)
Select Instance

Plot: CSV Copia de BD-NL COA TAM SLP

7 How Things Happen and How Can They Be Improved?

A public policy that specifies substantial changes in the educational paradigm of children and young people in the region should be channeled to increase the use of reading as part of the curriculum and also as an instrument for expanding knowledge and reinforcing attitudes and habits. Therefore, it can already be inferred, even if it is preliminary, that there are obstacles and barriers to overcome in terms of reading and writing, science learning, arts and deepening in humanities. Some will say that this occurs as a current phenomenon in all societies, from all continents and without differentiating the social stratum. But this does not deal with the root of the problem, nor does it contribute much to elucidating where we are standing.

Everything lies in a poor appreciation of the value of reading and how a great change in people can be achieved, in order to effect this transformation at the society level. It is thought that reading is unique to the humanities area and away from the sciences. That is why the ideal is that you can include reading and writing within all levels of study. What has been found so far is that the PISA test has obtained different scores in the four entities of the northeast of Mexico, together with other exogenous factors such as libraries, the mass media and social networks.

Finally, we propose a Model based on Structural Equations as is shown in Figure 1.

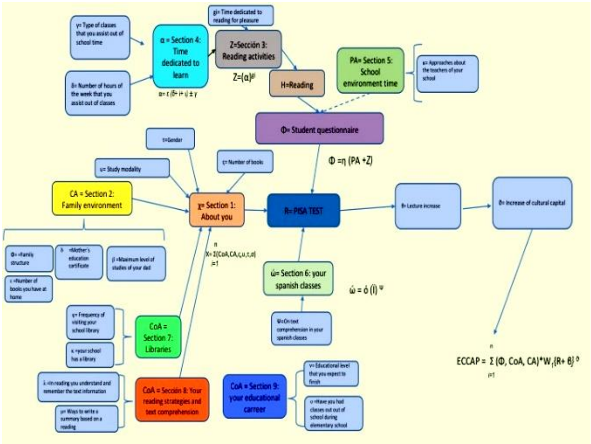

gi= Time dedicated to reading for pleasure
γ= Type of classes that you assist out of school time
α = Section 4: Time dedicated to learn
δ= Number of hours of the week that you assist out of classes
α= ε (δ+ i+ ι) ± γ
Z=Sección 3: Reading activities
Z=(α)^gi
H=Reading
PA= Section 5: School environment time
κ= Approaches about the teachers of your school
Φ= Student questionnaire
t=Gender
u= Study modality
ς= Number of books
Φ =η (PA +Z)
CA = Section 2: Family environment
χ= Section 1: About you
R= PISA TEST
θ= Lecture increase
ϑ= Increase of cultural capital
Φ= =Family structure
δ =Mother's education certificate
β =Maximum level of studies of your dad
ι =Number of books you have at home
X= Σ(CoA,CA,ς,u,t,α)
ώ= Section 6: your spanish classes
ώ = ό (ί) ^Ψ
φ= Frequency of visiting your school library
κ =your school has a library
CoA = Section 7: Libraries
Ψ=On text comprehension in your spanish classes
λ =In reading you understand and remember the text information
μ= Ways to write a summary based on a reading
CoA = Secciòn 8: Your reading strategies and text comprehension
CoA = Section 9: your educational carreer
ν= Educational level that you expect to finish
ο =Have you had classes out out of school during elementary school
ECCAP = Σ (Φ, CoA, CA)*W1(R+ θ)^ϑ

Fig. 1. Model associated with our research. (Source own).

8 Conclusions

A relevant aspect of this research is the use of Social Data Mining, in order to describe the best possible way to educational edges of the Northeast Mexican region and how this can determine the possibility of finding significant changes in future generations, this represented by means of software Weka 3.9.2., by means of the analysis of the results of the PISA test of the years 2012-2015 in the societies of the northeast, comparing the state of Nuevo León, with the states of Coahuila, San Luis Potosí and Tamaulipas. It is considered that reading, especially at early age, will be of great benefit so that in the future a 15-year-old adolescent, such as the one we are analyzing in this paper, will result in a better ability to write and read texts that are aligned to personal interests, or to the search for the truth. But what more do we aspire to find in a text, more than answers, ideas, new ways of seeing the world? We have to put all our effort into achieving an approach to reading, which forms more aware citizens interested in what has happened to us as a country and of what we want to project for the future, given that we have a strong cultural capital. That is why, taking advantage of the results of the PISA test will allow us not to make the same mistakes in educational fields but, rather, correct the way; take the helm again to build a society that, although it has much more access to information than previous generations, does not know how to find the truth. It is there where the libraries are the key piece.

We are inundated by *fake news*; for information that does not have reliable sources, by hundreds of leaders who want to move public opinion one way or the other. We need children and young people awake to learning, to the construction of knowledge and sensitive to the big world problems: inequality, extreme poverty, care of ecology and, above all, we need to believe in the value of communication face to face. Those of us who have had the great opportunity of having had an early childhood, where the evening was accompanied by a good story, a fascinating fairytale, an inspiring tale, we are, or should be, committed to the regeneration of a society that cries for the basics. That is why this research can gain relevance, not only from the sociological aspect, but above all through data mining (social data mining).

We rely on this transversality of knowledge to integrate the exact sciences, statistics, with something as essential as knowing where are we failing to get away from countries such as Shanghai, China, which has the first place in the world in reading. There is much to be done, but we hope that this research brings us closer and closer to achieving the results we deserve as a country. Mexico is a rich country in culture, traditions and values, and reading is the opportunity to break paradigms or judgments that do not allow us to grow as a society.

References

1. Reyes-Nava, A., Flores-Fuentes, A., Alejo, R., Rendón-Lara, E.: Minería de datos aplicada para la identificación de factores de riesgo en alumnos. Research Computer Science 139, pp. 177–189 (2017)
2. Rodríguez-Maya, N. E., Lara-Álvarez, C., May-Tzuc, O., Suárez-Carranza, B. A.: Modeling Students' Dropout in Mexican Universities. Research in Computing Science 139, pp. 163–175 (2017)
3. Gonzalez-Marron, D., Enciso-Gonzalez, A., Hernandez-Gonzalez, A.K., Gutierrez-Franco, D., Guizar-Barrera, B., Marquez-Callejas, A.: Evaluación de parámetros de encuesta de ingreso del CENEVAL para alumnos candidatos a ingresar al nivel superior, caso de estudio ITP. Research in Computing Science 139, pp. 135–147 (2017)
4. Ochoa, A., Tcherassi, A., Shingareva, I., Padméterakiris, A., Gyllenhaale, J., Hernández, J., Italianitá, A.: Discovering a Pygmalion effect on Italian communities using data mining. Advances in Computer Science and Engineering 57 (2006)
5. Monroy. M.I., Calvillo, T.: Breve historia de San Luis Potosí. Colmex-FCE, Ciudad de México, pp. 16 y 36-39 (1997)
6. Santoscoy, M.A., Gutiérrez, L., Rodríguez, M., Cepeda, F.: Breve historia de Coahuila. Colmex-FCE, Ciudad de México, p. 313 (2000)

7. Herrera, O.: Breve historia de Tamaulipas. Colmex-FCE, Ciudad de México, pp. 83, 161-163, 208-212 y 284 (1999)
8. Cavazos-Garza, I.: Breve historia de Nuevo León. Primera edición, Colmex-FCE, Ciudad de México, pp. 113, 156, 168-173-176, 200 y 204-213 (1994)
9. Lareau, A., McCrory, J.: Class, cultural capital, and institutiones: the case of families and schools, capítulo cuarto en Facing social class. En: Fiske, S.T., Markus, H.R. (editors). Rusell Sage Foundation, New York, pp. 61-86, (2012)
10. García, S., Poupeau, F.: La medición de la democratización escolar. Notas sobre los usos sociológicos de los indicadores estadísticos. En: Bourdieu, P.: Capitalismo simbólico y magia social. Segunda edición. Siglo XXI Editores, Ciudad de México, pp. 205-235 (2014)
11. Argüelles, J.D.: Por una universidad lectora. Laberinto-UJAT, p. 61 (2015)
12. Pérez-Maldonado, M.M.: Análisis de las bibliotecas comunitarias del estado de Nuevo León. Tesis de la Maestría en Gestión de Servicios Informativos de la UACJ. DOI: 10.13140/RG.2.2.33746.04801 [recuperado en junio 2018]
13. Roberto, M.: Tres breves lecciones de literatura, por Juan José Arreola. En Armas y Letras Núm. 35, marzo-abril, pp. 49-50 (2002)
14. Elaboración propia. Inspirada en la Tabla IDH (Índice de Desarrollo Humano) de los estados de Nuevo León, Coahuila, Tamaulipas y San Luis Potosí. Anexo: Entidades federativas de México por IDH. [Recuperado de: https://es.wikipedia.org/wiki/Anexo:Entidades_federativas_de_M%C3%A9xico_por_IDH]
15. Vargas, A.M.: Resultados del examen de evaluación educativa PISA 2012 (en puntos). Matemáticas, Lectura y Ciencias (2012) [Recuperado de: https://www.google.com/search?q=resultados+pisa&source=lnms&tbm=isch&sa=X&ved=0ahUKEwim86uSodHcAhUDUK0KHcNYC1wQ_AUICigB&biw=1354&bih=671#imgrc=odQHg7msqhG06M]
16. Tendencia promedio de Prueba PISA sobre el desempeño en ciencia desde 2006. [Recuperado de: https://www.google.com/search?biw=1354 &bih=671&tbm=isch&sa=1&ei=kHBkW-2OA8S8sQXqnL6ICg&q=tendencia+promedio+de+desempe%C3%B1o+pisa+2006&oq=tendencia+promedio+de+desempe%C3%B1o+pisa+2006&gs_l=img.3...18183.23325.0.23946.25.23.0.0.0.0.173.2316.14j8.22.0....0...1c.1.64.img..11.9.850...35i39k1.0.a92lJaD2Bss#imgrc=vxOwJmR9m67x5M]

ISSN 1870-4069

Conceptualization of a Predictive Model for Analysis of the Health Outcomes of Dust Events in a Society with Köppen Climate Classification BW

Estrella Molina-Herrera[1], Alberto Ochoa[2,3], Thomas Gill[1],
Gabriel Ibarra-Mejia[1], Carlos Herrera[1]

[1] University of Texas at El Paso, El Paso, TX, USA
[2] The Autonomous University of Ciudad Juarez, Chihuahua, Mexico
[3] El Paso VA Health Care System, TX, USA*

Abstract. High concentrations of particulate matter (PM) in the air during Dust Events (DEs) are silently impacting the health of people without their awareness. It has been demonstrated that exposure to increased levels of PM can increase the susceptibility to respiratory, circulatory, mental and other diseases due to inflammation. In addition, living in a city with Köppen climate classification type BW (arid) and subsequently with frequent high levels of PM could have a negative impact on the population's health. There are very few studies available in the southwestern United States pertaining to the associations between exposure to atmospheric aerosol after DEs and hospitalizations. Therefore, we will do a conceptualization of a predictive model to analyze the health effects of DEs in a society with Köppen climate classification type BW. We will do a representation of a system in order to understand how the DEs, hospital admissions, elevated PM levels, socioeconomic status (SES), and demographic factors work together. Preliminary results indicate that there are more admissions in all primary diagnoses during a DE than in a regular day.

Keywords: ecological data mining, multivariable analysis, pattern recognition, structural equation, long term health effects, oxidative stress, inflammatory responses, social economic data.

1 Introduction

The Southwestern region has been identified as one of the most persistent dust producing regions of North America (Orgill and Sehmel, 1976; Prospero et al., 2002). Exposure to inhalable particulate matter of 10 micrometers or less in diameter (PM_{10}) originating from desertic landscape during DEs can reach toxic levels (Song et al, 2007). El Paso's ambient air has reached hazardous levels of PM10 above 4000 $\mu g/m^3$ with near zero visibility due to these natural events (Rivera et al., 2010), thus exceeding the primary and secondary 24-hour standard of 150 $\mu g/m^3$. According to the National Ambient Air Quality Standards (NAAQS), this standard should not be exceeded more than once per year based on an average of 3 years (EPA, 2018). In El Paso, TX, DEs occur on average 14.5 times per year (Novlan, D., Hardiman, M., & Gill, T., 2007), which are conditions that resemble those of the Dust Bowl during the 1930's at the Southern Plains of Texas (Lee and Gill, 2015). Deadly respiratory health problems were prevalent during that period (Alexander, Nugent and Nugent, 2018).

Recent literature has shown that exposure to desert-related particles during DEs is associated with increased hospitalizations due to respiratory or circulatory-related problems (Zhang et al. 2016). However, not much is known about the possible effects of exposure to desert-related particles during DEs on mental and neurological-related health problems. Because it has been shown that inhaled particles induce an inflammatory response that starts in the lung, spills into the circulatory system, and ultimately can reach the brain, I suspect that exposure to very high levels of particles from natural sources during DEs might increase hospitalizations due to mental and neurological-related health problems. Understanding the impact of environmental exposures on these types of health problems is important as depression, Parkinson's disease and Alzheimer's disease are the three most prevalent and costly mental and neurodegenerative diseases in the U.S. (Weintraub, Karlawish & Siderowf, 2007). Furthermore, socially disadvantaged individuals, such as those of low socio-economic status (SES) or those who are frequently exposed to discrimination and isolation (e.g., racial and ethnic minorities) tend to be more susceptible to the health effects of air pollution exposure (Grineski et al., 2015; Halonen at al., 2016). In addition, evidence suggests that factors such as age, gender, and ethnicity might affect the association between exposure to particles and health problems, but this mediating role is not clear (Howard, Peace & Howard, 2014).

During DEs -particularly in arid regions- particles from deserted landscapes get lifted into ambient air by high wind speeds where they combine with particles emitted by urban sources (e.g., vehicles, industry source components that are in the air or settled on the roads) as well as with biological particles in nature (e.g., spores, fungi) (Fuzzi et al., 2015). Currently, there is little understanding of the health effects induced by exposure to the above-mentioned particle mixtures, which during DEs can reach high, unsafe concentrations. Exposure to DEs is more frequently experienced by populations that live in arid and semiarid regions of the world. In the United States, DEs are frequent within and around the Chihuahua Desert of Texas, which is where the proposed study will focus on. Addressing the associations between DEs and hospitalizations in these arid regions of the US will greatly inform the scientific community, habitants, and the environmental and social authorities who are responsible for implementing the proper adjustments. The following sections will provide a review of the relevant literature to and identify gaps that illustrate the significance of the proposed study.

2 Literature Review

2.1 Description of the Chihuahuan Desert (EL PASO)

This study will focus on parts of the Chihuahuan Desert (Texas) which is the most persistent dust producing regions of North America (Lee et al., 2009; Novlan et al. 2007; Rivera et al. 2009 and 2010) (see Figure 1). The high frequency of dust storms in these regions are due to large-scale dry climate (climate type -according to the Köppen climate classification system- cold desert (BWk), hot desert (BWh) (Lee et al., 2012; Lee and Tchakerian, 1995; Rivera et al., 2009; Bernier et al., 1998; Li et al. 2018).

The Chihuahuan Desert is the one of the most significant sources of dust in the Western Hemisphere (Prospero et al., 2002). In this region, agricultural lands, ephemeral lakes, and dry river beds have been identified as the main sources of the dust from this desert that is blown into El Paso, Texas (Lee et al., 2009). Within the Chihuahuan Desert, I will focus specifically in dust events occurring in El Paso, Texas, which is the largest city in the US that is located in the central part of the Chihuahuan Desert (see Figure 1). In El Paso, dust events have been identified as important environmental hazardous events. Based on data collected at the El Paso International Airport from 1932 through 2005, dust events in El Paso occur on average 15 times a year and last an average of 2 hours each (Novlan et al., 2007). In this region, dust storms occur most commonly during the months of December through May when ambient air is dry and winds can reach high speeds (>25 mph)), blowing primarily on strong westerly and southwesterly winds (Novlan et al., 2007). At wind speeds greater than 25 mph, dust can be raised into the atmosphere and/or transported for long distances by synoptic-scale weather systems (horizontal length scale of the order of 1000 kilometers or more), which results in widespread exposure to ambient air particle mixtures (Lee et al., 2009).

2.2 Air Pollution

Particles, also called atmospheric aerosols, that are less than 10 μm in diameter (PM_{10}) have very low sedimentation speeds under gravity and may remain in the air for days before eventually being washed out by rain or impacted out onto vegetation or buildings, but they can be re-suspended from surfaces during a DE. These particles are a regulated environmental pollutant, being responsible for reducing visual range, soiling surfaces, and negatively impacting human health (Colls, 2002). PM_{10} concentrations can reach very high levels during DE, particularly in desert environments or near agricultural fields or unpaved roads where high wind speeds can lift surface particles (Jacob, et al., 2009). Ambient levels of PM_{10} in the US are regulated by the US Environmental Protection Agency (US EPA). Standards for PM_{10} consist of 150 $\mu g/m^3$ during 24-hour periods and are not to be exceeded more than once per year on average over 3 years (visit https://www.epa.gov/criteria-air-pollutants/naaqs-table). Peak hourly concentration of PM_{10} in El Paso during a DE has reached 1,955.2 $\mu g/m^3$.

Aerosol content in the atmosphere depends on its origin (urban, rural, marine, desertic or combined), as well as physical properties and chemical composition, all of which induces different health effects within each environment (Carvalho-Oliveira, 2015). Aerosols may have either a primary or secondary origin, be solid or liquid, and come from biological or inorganic sources. Primary sources of particles include industrial processes, transport-related processes, unpaved roads, fields, fires, wood combustion, marine aerosol, and mineral dust aerosol (MDA- principal component from all the atmospheric aerosol in the planet) (Fuzzi et al., 2015). Secondary particles result from complicated reactions of chemicals in the atmosphere from compounds such as sulfur dioxides and nitrogen oxides which are typically emitted from power plants, industrial processes, and automobiles (EPA, 2018b).

Globally, it is estimated that the main sources of particulate matter contributing to urban air pollution are: 25% by traffic, 15% by industrial activities, 20% by domestic fuel burning, 22% from unspecified sources of human origin, and 18% from natural dust and salt (Karagulian et al., 2015). However, in a dusty arid region such as in El Paso, these percentages are likely very different. At El Paso, 35% of the total mass concentrations in the PM_{10} fraction accounted for Major elements from geologic sources, indicating that geologic sources in the area are the dominant PM sources through the year (Li et al., 2001).

2.3 Characterization of Dust Storms

Within the Southwestern US, DEs are caused by synoptic-scale Pacific cold fronts moving across the desert from west to east, and cyclones developing and intensifying to the northeast (Rivas et al., 2014). All these factors create the conditions for DEs, which is defined as an event with PM_{10} above 150 µg/m^3 while wind speeds can have gusts above 10 m/s (see figure 1) (Hosiokangas et al., 2004; Lee et al., 2009; Rivera et al., 2009). Low wind conditions can also lead to elevated levels of pollutants and particulates in the air. Nevertheless, per a study conducted in El Paso (Grineski et al., 2011) and one in Lubbock (Lee and Tchakerian, 1995), low wind conditions are not or rarely associated with dust events.

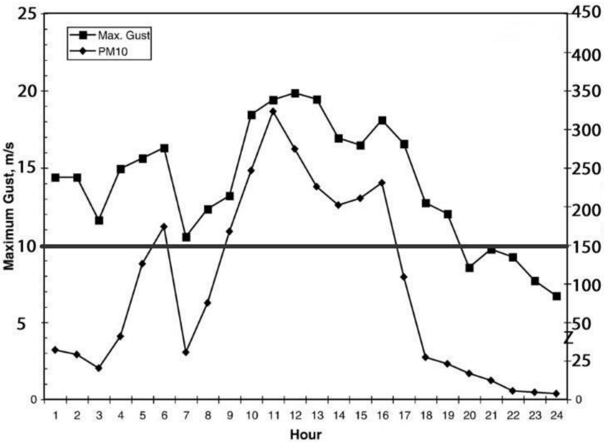


Fig. 1. Definition of a Dust Event for this study; Wind speed with gusts above 10m/s and PM_{10} above 150 µg/m^3.

Desert dust can be transported across the world by arid and semi-arid regions where loose soil can easily be lifted during high wind speeds (Lim & Chun, 2006). For instance, dust from the Sahara Desert can be transported across the Atlantic Ocean and reach northeastern South America, the Caribbean, Central America, and southeastern United States (Kanatani Et al., 2010). This transportation to distant regions by DEs is generated when strong surface winds lift up fine grained dust particles into the air and strong turbulence or convection diffuses the dust, particulate material, biological aerosols and pollutants (Shao, 2008; Zhang et al., 2016). It is estimated that 75% of the global dust emissions is due to natural origin, while 25% are related to anthropogenic (primarily agricultural) emissions (Ginoux et al., 2012), with the Sahara Desert as the largest source of natural mineral dust aerosol (Karanasiou et al., 2012).

It is estimated that the total dust deposition rate during a DE at El Paso, TX is approximately 195.5 g/m^2/yr, where values are elevated in comparison to dust deposition elsewhere in the region, and closer to other global desert areas (Rivas et al., 2014). The principal size class of deposited sediment during DEs is sand (86.81%) followed by 9.25% of PM_{10} and 3.94% of $PM_{2.5}$. An air monitoring station near the study area at the same times indicated peak hourly PM_{10} values of 1955.2 μg/m^3 and for $PM_{2.5}$ 288.33 μg/m^3 (Rivas et al., 2014).

The mineralogy of DE particles at El Paso, TX is dominated by quartz (silicon dioxide) with the presence of other common minerals such as plagioclase, gypsum, and calcite (Rivas et al., 2014). In addition to the inorganic particulate matter contained in the dust during a DE (contained in the PM), there are substantial quantities of foreign microorganisms derived from the downwind atmosphere, terrestrial, and aquatic environments (Zhang, Zhao & Tong, 2016). Significant increases in the concentration of bacteria and fungi are commonly detected in dust clouds during sandstorm events (Tang et al. 2018). DE are known as one of the most far-reaching vehicles for transport of highly stress resistant and potentially invasive/pathogenic microorganisms across the globe (Weil et al. 2017).

2.4 Dust, Fugitive Dust, Aerosols and their Health Effects

In the Southern High Plains, the dominant aerosol elemental content during DE includes Al, Si, S, Cl, K, Ca, Ti, Mn, Fe, and Zn, with minor and trace elements (Cr, Ni, Cu, Rb, Zr, and Pb) (Gill, Stout and Peinado, 2009). On the other hand, Garcia et al. (2004) found that the elements in El Paso's dust-emitting soil are largely the same elements found in the Southern High Plains (Al, Ca, K, Zn, Cr, Ni, Cu, Pb and Mn), plus Na, Ag, As, Cd, Mo, Sb, Ba, Co, and Be (Li et al., 2001), which are fugitive dust sources that might increase during dust events. A recent study near Las Vegas, NV during a DE showed that accumulated particles on the road are re-suspended. These suspended particles are composed of a more complex mixture of elements, including Al, V, Cr, Mn, Fe, Co, Cu, Zn, As, Sr, Cs, Pb, U, and others (Keil et al., 2016), This fugitive dust are disease precursor with hazardous effects on human health (e.g. carcinogenic and non-carcinogenic) (Khan, & Strand, 2018; Kioumourtzoglou et al., 2015). Furthermore, Huang et al. (2014) found house air-conditioner dust to be more hazardous than road dust; within these particles lead was the most abundant element, followed by arsenic.

Several studies have hinted that exposure to particle air pollution during dust events could have a direct impact on human health (Anderson et al., 2013). This is because the PM<10 μm can penetrate into the lungs and exposures are based upon respirable dust (≤5 μm) (Bhagia, 2012; Middleton, 2017). For example, the size fractions of silica in ambient dust is in the range of 2.5-15 μm and PM<2.5 μm can penetrate into deep lung tissue (Bhagia, 2012). Besides the composition of particles, and the size and surface area of breathable particles, air pollution has been found to affect the degree of oxidative stress and the release of cytokines, accelerating inflammation in the body (Dostert et al., 2008; Ghio et al., 2004) (see Figure 2). Systemic inflammation, endothelial activation, and low-grade inflammation caused by inhaled traffic-related PM (Li et al., 2015; Chiarelli et al., 2011), has been hypothesized as a key factor in the pathway leading to detrimental structural and cognitive effects, as well as neurodegenerative and mental illnesses (Calderón-Garciduеñas et al., 2015; Heusinkveld et al., 2016).

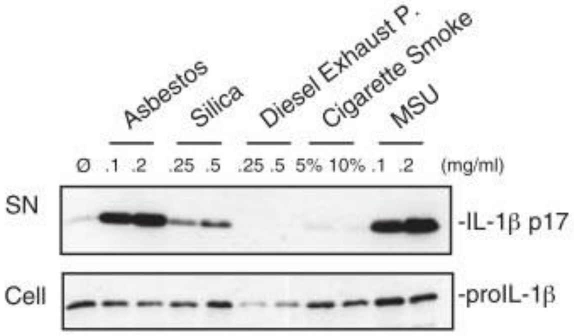


Fig. 2. Silica activate IL-1B secretion in human macrophages. Analyzed in media supernatants (SN) and in cell extracts (Cell). From Dostert et al. 2008.

In addition, recent studies have shown that particle air pollution during DE increase hospitalizations for expected causes such as respiratory and cardiovascular disorders (Khaniabadi et al., 2017; Yu, Chien, and Yang, 2012). Even more, recent studies suggest that silica dust influence brain function and aggravates spinal cord injury. Exposure to silica dust increases epithelial permeability in patients with silicosis who smoke (Nery et al., 1993). Keil et al. (2018) performed an exposure study to dust at the southwest USA with a PM median diameter of 4.6, 3.1, and 4.4 µm. Results showed an overall reduction in the immune response rather than a direct effect of dust samples on neuronal protein-specific antibody production but neurotoxicity cannot be ruled out as a concern. Also, increased levels of serum creatinine -a marker for kidney function-were found. A previous study (Keil et al., 2016) showed that brain CD3+ T cells were decreased in number after dust exposure with silica and heavy metals present in the southwest soil.

Hospitalizations after a DE have been reported to have a prolonged effect on the day of the DE and on the week after the DE (Chien, Yang and Yu et al., 2012). Therefore, in this study, hospitalizations will be under particular scrutiny during those day(s) when a dust storm event is taking place, as well as all throughout the following week.

2.5 Biological Aerosol Particles and its Health Effects

Sandstorms from the Sahara Deserts transmit roughly a billion tons of dust across the atmosphere, and the region is considered one of the major sources of the intercontinental dust transport (Griffin 2007). The Gobi and Taklamakan Deserts in Asia are the second largest sources (Zhang et al., 2016). These dust plumes can reach as far as the Americas (Husar et al., 2001), transporting trillions upon trillions of microbes into the air and downwind destinations along their intermediate path which are added to the own desert microbiome (Behzad, Mineta & Golobori, 2018). By some estimate, a cubic meter of air contains hundreds of thousands of microorganisms (Prussin et al., 2015; Brodie et al., 2007), with an extensive diversity of taxa (Franzetti et al., 2011).

Mineral dust aerosol (MDA) contains primary biological aerosol particles (PBAPs) and has a large range of different biological components, including microorganisms (bacteria, archaea, algae and fungi) and dispersal material (pollen, fungal spores, viruses and biological fragments) (Fuzzi et al., 2015). Furthermore, large deserts create their own Dust Storm Derived Microbiota (DSM) (Griffin, 2007). This microbiota includes highly stress-resistant microorganisms (bacteria and fungi) that are capable of thriving in harsh environmental conditions with restricted water and nutrient availability, extreme temperatures, and UV irradiation (Chan et al. 2013; Etemadifar et al., 2016). Viruses on the other hand can undergo degradation by atmospheric processes and can experience a possible loss of their toxic effects in the source regions as they are carried away (Despres et al., 2012). This large-scale transmission of highly resistant microbial contaminants raises concerns with regards to human health (Chung and Sobsey, 1993 and Cox, 1995).

It has been proven that viruses present during DE are taxonomically diverse (Zablocki et al., 2016) and are transported by the dust across long distances (Chien et al., 2012; Chung and Sobsey, 1993). This movement leads to significantly higher cases of Influenza A virus, typhus, cholera, malaria, dengue and West Nile virus infection than is typically observed during normal non-DE days (Griffin, 2007). Examples of influenza outbreaks type A virus and H5N1 avian influenza occurred in Taiwan, Japan and South Korea during the Asian Dust Storms (ADS) that originated in the deserts of Mongolia and China (Chen et al. 2010).

Bacterial epidemics are strongly linked to DEs. Bacterial meningitis is associated with DEs, which is a major predictor of the timing of meningitis epidemics (Agier et al., 2013). In 1935, Kansas experienced a severe measles epidemic during the Dust Bowl. Hospital admissions were largely for acute respiratory infections such as pneumonia, sinusitis, laryngitis and bronchitis (Brown et al., 1935). Similar cases of respiratory infections due to DE can be found in Western China (Ma et al., 2017). The epidemics of pulmonary tuberculosis was similarly linked to ADS in China (Wang et al., 2016). ADS were also positively associated with diabetes in women (Chan et al., 2018).

Another infectious disease presumably caused by fungi during a DE is the Valley Fever, whose fungal causative agents (Coccidioides immitis and Coccidioides posadasii) are primarily found in hot and arid desert soil (Kirkland and Fierer, 1996). The outbreaks of Kawasaki disease (a serious heart complication acquired in childhood) was linked to a fungal Candida species found in DE from China (Rodo et al. 2014; Tong et al. 2017).

2.6 Inflammatory Response Pathway

The inflammatory response helps the body fight and clear infection, remove damaging chemicals, and repair damaged tissue. However, frustrated phagocytosis (an action where a phagocyte fails to engulf its target and toxic agents can be released) can have a harmful effect on the body (Dostert et al., 2008). At its worst, inflammation can provoke cancer (Tili et al., 2011). There are two pathways that link PM air pollution (gases, ultrafine particles, and nanoparticles present in the particulate matter like silica from the dust) to adverse health outcomes (Shrey et al., 2011).

The first is a direct pathway, which consists of the local oxidative stress/inflammation effects of pollutants on the cardiovascular system, blood, and lung receptors (Garcia et al., 2015). This direct pathway involves the direct translocation (the dominant method of trapping and processing particles in the lung tissue) of inhaled fine particles present in the PM into the circulatory system causing intracellular oxidative stress releasing cytokines and chemokines (Nemmar et al., 2010). Particles can readily cross the pulmonary epithelium or the lung–blood barrier due to their particle size, charge, chemical composition, and propensity to form aggregates (Oberdörster et al., 2004). Once such particles like silica are in circulation, they lead to further deleterious effects such as local oxidative stress and inflammation (Brook et al., 2010). The mechanism starts with local inflammation in the upper and lower respiratory tract resulting in increased levels of pro-inflammatory mediators (e.g., IL-6, IL-8, and of tumor necrosis factor alpha (TNF-α) following into the circulatory system inducing low-grade peripheral inflammation (see Figure 3) (Olvera et al., 2018). An example of this direct pathway is that in rats, a three-hour PM2.5 exposure has been shown to lead to a rapid increase of reactive oxygen species (ROS) generation in the heart and lungs (Gurgueira et al., 2002; Li et la., 2015).

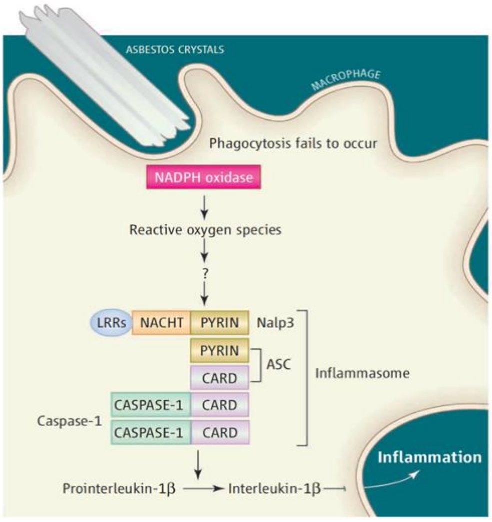


Fig. 3. Systemic inflammation mechanism due to DE. Asbestos crystals or silica are too large to be phagocytosed by macrophages and so are subject to "frustrated" phagocytosis. This leads to activation of NADPH oxidase and the generation of reactive oxygen species. This event activates the Nalp3 and ASC inflammasome promoting the processing and release of the potent proinflammatory molecule interleukin-1B. From O'Neill et al., 2008.

The second pathway is the classical pathway, which explains the indirect effects mediated through pulmonary oxidative stress and inflammatory responses (Nemmar et al., 2003; Tonne et al., 2016). It begins when inhaled traffic-related PM enters the body through the airway to the lungs and causes a local inflammatory response at the bronchial epithelial cells and from alveolar macrophages (Bai and Sun, 2015). Bronchial epithelial cells and alveolar macrophages are in prolonged contact with the inhaled particulates when clearing them from the lung, which can initiate and sustain inflammatory responses (Dostert et al., 2008). Silica are sensed by the Nalp3 inflammasome, whose subsequent activation leads to interleukin-1b secretion. The onset of this inflammatory response, at a cellular level, is triggered by the release of TNF-α and IL-1β which regulate the expression of various secondary cytokines and chemokines, including IL-6 and IL-8 (Schwarze et al., 2013; Morman and Plumlee, 2013).

2.7 Health Disparities

Health disparities are health differences that adversely affect socially disadvantaged groups (Krieger, 2016). Health disparities are systematic, reasonably avoidable health differences according to race/ethnicity, skin color, religion, or nationality; socioeconomic resources or position (reflected by, e.g., income, wealth, education level, or occupation); gender, sexual orientation, gender identity; age, geography, disability, illness, political or other affiliation; or other characteristics associated with discrimination or marginalization. These categories reflect social advantages or disadvantages when they determine an individual’s or group’s position in a social hierarchy (Braveman et al., 2011). Furthermore, inequities in social determinants of health, including neighborhood poverty, crime rates, and reduced access to high-earning jobs, housing, transportation, and healthy foods significantly contribute to these disparities (Cooper et al., 2016). Disparities in health and its determinants are the metric for assessing health equity (Gee, Walsemann, & Brondolo, 2012). An example of health disparities is that being overweight is negatively associated with income, education level, and occupation at the municipality level (Kinge et al., 2016).

Moreover, social factors (e.g., stress, health disparities, low access to resources) may induce intrinsic vulnerability to the effects of air pollution, including a pro-inflammatory phenotype that results in increased inflammatory reactivity to air pollution exposure that may be heritable (Wu et al., 2016; Heusinkveld et al., 2016). An example of this is the impact of PM2.5 on markers of systemic inflammation and oxidation in those with multiple pre-existing cardiovascular diseases with elements of metabolic syndrome (e.g. obesity, diabetes, hypertension and smokers) (Aguilar et al., 2015).

Opposite to health disparities, gender, age and genetics are a natural disorder cause. For example, a study by Reynolds et al. (2016) found that women experienced a significantly greater decrease in incidence of myocardial infarctions compared with men. Other investigators suggest that cumulative stress may result in affecting biological processes, such as shortening telomere length. The length of telomeres on chromosomes declines with age and may be an indicator of remaining life expectancy. Some evidence suggests that there is a systematic relationship between educational attainment and the length of telomeres (Adler et al., 2013; Kaplan, 2014).

2.8 Dust Storm Projections

For the last 50 years, an acceleration of changes on the average climate conditions (IPCC, 2007a) has been observed. The average global temperature has increased by 0.7 °C and it is expected to increase between 1.8 and 4.0 °C by the year 2100 (IPCC, 2007b; Hansen et al., 2006). The frequency of dust storms has increased during the last decade and forecasts suggest that this will continue to rise in response to anthropogenic activities and climate change (Schweitzer et al., 2018). El Paso del Norte is the region that has the highest probability for DE (Rivera, Rivera et al., 2009).

Climate change poses unprecedented threats to human by impacts on health, food and water security, heat waves and droughts, dust storms, and infectious diseases; whether or not humanity will successfully adapt is not yet known (Barrett et al., 2015). Some infectious diseases and their animal vectors are influenced by climate changes, resulting in higher risk of typhus, cholera, malaria, dengue and West Nile virus infection which are carried by DE (Franchini & Mannucci, 2015). Moreover, climate drivers (increase of temperatures, changes in precipitations patters, extreme weather effects), environmental changes (changes in pollutant exposure, changes in allergens production, timing and distribution), urban landscapes, emission patters), and social and behavioral context (income, education, sensitivity, adaptive capacity and housing quality) can affect an individual's or a community's health vulnerability over the time (Global Change, 2017).

3 Methodology

3.1 Data Sources

Hospital admissions: Five years of data were obtained from the Texas Hospital Inpatient Research Data files (RDF) from the Texas Department of State Health Services (TDSHS) for years 2010 through 2014 for El Paso TX. The data included the following five variables: the date of admission, census block group of the patient, the patient's age, gender, ethnicity, and the principal diagnostic code from the International Classification of Diseases, Ninth Revision (ICD)-9 (see Table 1). The principal diagnostic code was preferred over other diagnostic codes because it better captures the exacerbations of disease as opposed to other diagnostics due to existing diseases.

PM and wind speed data: Hourly averages of PM_{10} concentrations, wind speed (m/s), relative humidity, and mean, minimum, and maximum temperature (∘ F) measured at Continuous Air Monitoring Stations (CAMS) located in El Paso, Lubbock, Midland and Amarillo, listed in Table 1, from 2010-2014 will be downloaded from the Texas Commission on Environmental Quality (TCEQ) website. PM_{10} and wind speed missing data will be interpolated using a temporal linear method in cases where the data were missing for three consecutive hours or less; days with data missing for four or more consecutive hours will be excluded from the analysis. It is expected that of the total dataset, about 1% of all analyzed days would require missing data interpolation; after interpolation, the dataset will be over 99.7% complete.

Socio-economic data: Economic characteristics, including income, level median income, poverty, occupation and education for each patient address census block group will be obtained from the U.S. Census Bureau's American Community Survey for the 2010-2014 period. It will help us to identify susceptible individuals. This information will be connected with the RDF's Address Census Block Group code of each hospitalized patient in the CDT and HPWT.

Demographic data: Population increase or decrease (in millions) data between 2010-2014 will be obtained from the census data of statistics in the county of El Paso, Texas, to remove these non-environmental confounding elements (population increase or decrease).

Table 1 Codes from the International Classification of Diseases, Ninth Revision (ICD)-9.

Code Range	**Diagnosis**	**Code Range**	**Diagnosis**
1 001-139	Infectious and parasitic diseases	**10** 580-629	Diseases of the genitourinary system
2 140-239	Neoplasms	**11** 630-679	Complications of pregnancy, childbirth, and the puerperium
3 240-279	Endocrine, nutritional and metabolic diseases, and immunity disorders	**12** 680-709	Diseases of the skin and subcutaneous tissue
4 280-289	Diseases of the blood and blood-forming organs	**13** 710-739	Diseases of the musculoskeletal system and connective tissue
5 290-319	Mental disorders	**14** 740-759	Congenital anomalies
6 320-389	Diseases of the nervous system and sense organs	**15** 760-779	Certain conditions originating in the perinatal period
7 390-459	Diseases of the circulatory system	**16** 780-799	Symptoms, signs, and ill-defined conditions
8 460-519	Diseases of the respiratory system	**17** 800-999	Injury and poisoning
9 520-579	Diseases of the digestive system	**18** E000-E999	Supplementary classification of external causes of injury and poisoning

3.2 Model Analyses

Dust storm periods will be identified by matching the hourly average PM_{10} exceeding 150 $\mu g/m^3$ and high winds above 10 m/s (Rivera et al., 2009). In order to estimate the influence of dust storm's particulate matter from hospitalizations, a regression model will be generated to determine the correlations between the identified dust storms and hospitalizations during one-week period (the day of the storm and week after the dust storm) identified in El Paso county.

SES data from the U.S. Census Bureau's American Community Survey for the 2010-2014 period will be connected with the RDF's Address Census Block Group code of each hospitalized patient identified. An association between diseases outcomes and SES, (including income, poverty level, occupation and education level at county level in El Paso, TX) will be looked upon. Also, it will be searched if there is a remarkable reduction/increase in the incidence of hospitalized residents with any disease that is affected by dust events from 2010-2014. A search will be conducted for an association between diseases and SES, including age, sex, and race at county level county level in El Paso, TX.

Model analysis will be applied using data mining. A conceptual model will be applied to establish a basic model to explore the associations between the predictor variable (Dust events) and response variables (admissions, age, sex and SES). I will remove long-term trends and seasonal patterns from the data to protect against confounding by omitted variables. I will control for season and long-term trend with a natural cubic regression spline with 1.5 degrees of freedom (df) for each season and year (corresponding to 6 df per year). In addition, I will include natural splines with three df for temperature on the day of the admission and with 2 df for the six following days and a linear term for daily average humidity and dummy variables for the day of the week effect and public holidays. Once the data is normalized, each diagnosis code will be categorized into; acute, chronic and mental, in order to have a better understanding of the associations between DEs and diagnosis. Separated models will be run for each outcome of significant primary diagnosis. Models for present (2010-2014) and future projections (2020 and 2050) will be modeled separately.

In addition, geographic maps will be created in each municipality indicating their PM_{10} levels and hospital admissions percentage during a DE and their association between each socio-economic factor per 1000 population in El Paso, TX between 2010 and 2014 and projected outcomes (2020 and 2050). This will be done by using the Empirical Bayesian Kriging (EBK) Regression Prediction Method by ArcGIS (ESRI, Redlands, CA, USA).

4 Preliminary Results

We propose that projected health outcomes due to DEs are manifested by patient hospitalization which is associated with environment, demographic and socio-economic factors as the following model and formulas indicate (Figure 4).

$$\alpha = \mathrm{i}\ (\gamma+\delta+\varepsilon+\iota) \pm id, \qquad (1)$$

where Educational attainment (α) is defined by: γ= Neighborhood, δ= Access to education, ε= Parent expectations about children, ι= Local inequities/disparities (these factors are rated from 1 to 10, being 1 the lowest given value and 10 the highest according the present and projected ratings for 2020 and 2050 in each Census Block Group code at the El Paso County) and ι= Income (value given from the Census Block Group code at the El Paso County and projected values for 2020 and 2050).

$$Z=(\zeta)^{gi}, \qquad (2)$$

where Occupation (Z) is defined by: ζ=Occupation (value given from the Census Block Group code at the El Paso County) and gi= Inequities/disparities based on globalization (rated from 0 to 10, being 1 the lowest given value and 10 the highest according the present and projected ratings in each Census Block Group code at the El Paso County).

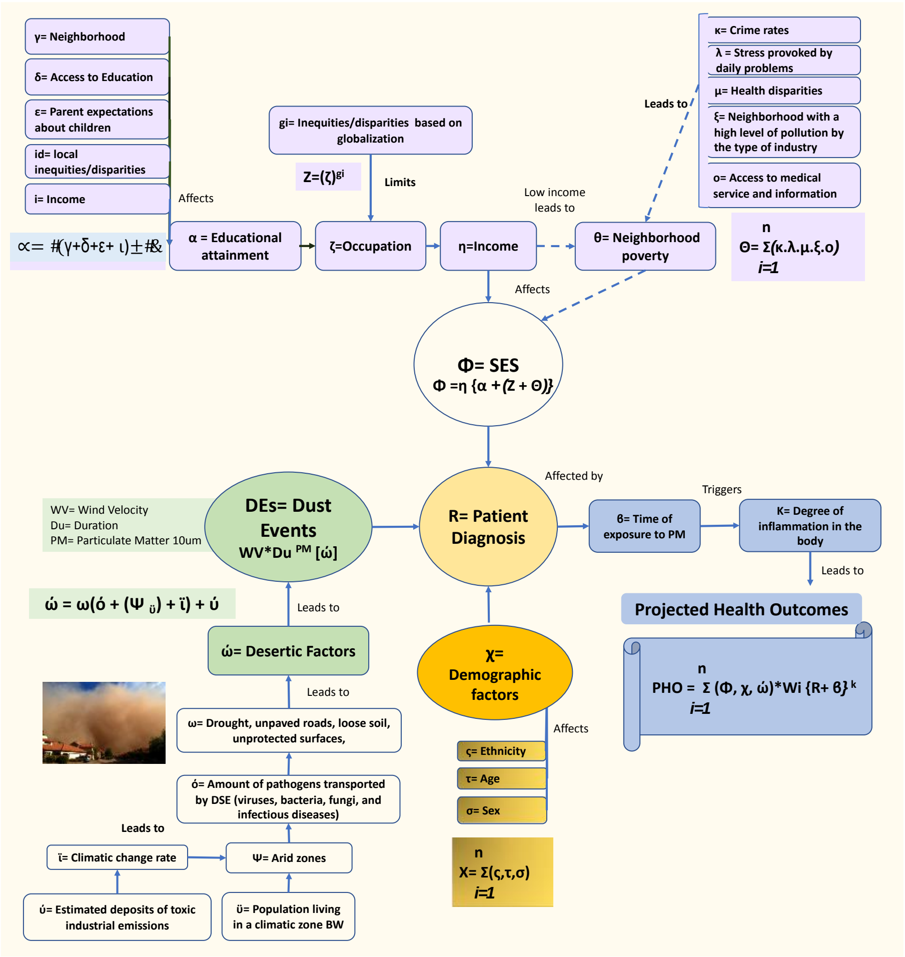

γ= Neighborhood
δ= Access to Education
ε= Parent expectations about children
id= local inequities/disparities
i= Income
Affects
∝= #(γ+δ+ε+ ι)±#&
α = Educational attainment
gi= Inequities/disparities based on globalization
Z=(ζ)gi
Limits
ζ=Occupation
η=Income
Low income leads to
θ= Neighborhood poverty
Leads to
κ= Crime rates
λ = Stress provoked by daily problems
μ= Health disparities
ξ= Neighborhood with a high level of pollution by the type of industry
o= Access to medical service and information
n
Θ= Σ(κ.λ.μ.ξ.o)
i=1
Affects
Φ= SES
Φ =η {α +(Z + Θ)}
WV= Wind Velocity
Du= Duration
PM= Particulate Matter 10um
DEs= Dust Events
WV*Du PM [ώ]
ώ = ω(ó + (Ψ ü) + ï) + ú
Leads to
ώ= Desertic Factors
Leads to
ω= Drought, unpaved roads, loose soil, unprotected surfaces,
ó= Amount of pathogens transported by DSE (viruses, bacteria, fungi, and infectious diseases)
Leads to
ï= Climatic change rate
Ψ= Arid zones
ú= Estimated deposits of toxic industrial emissions
ü= Population living in a climatic zone BW
Affected by
R= Patient Diagnosis
β= Time of exposure to PM
Triggers
K= Degree of inflammation in the body
Leads to
Projected Health Outcomes
n
PHO = Σ (Φ, χ, ώ)*Wi {R+ β} k
i=1
χ= Demographic factors
Affects
ς= Ethnicity
τ= Age
σ= Sex
n
X= Σ(ς,τ,σ)
i=1

Fig. 1. Conceptual predictive model of DEs and its health outcomes in a society of BW climate.

$$\Theta = \sum_{i=1}^{n} (\kappa.\lambda.\mu.\xi.o.), \tag{3}$$

where Neighborhood poverty (θ) is defined by: κ= crime rates, λ = stress provoked by discrimination, μ= health disparities, ξ= Neighborhood with a certain level of pollution according to the type of industry and o= access to medical service and information.

$$\Phi = \eta \; \{\alpha + (Z + \Theta)\}, \tag{4}$$

where SES (Φ) is defined by: η= income, α = Educational attainment, Z= Occupation, θ= Neighborhood poverty.

$$X=\sum_{i=1}^{n}(\varsigma,\tau,\sigma), \tag{5}$$

where Demographic Factors (χ) is defined by: ς= Ethnicity, τ= Age and σ= Sex.

$$\acute{\omega} = \omega(\acute{o} + (\Psi_{\ddot{\upsilon}}) + \ddot{\imath}\,) + \acute{\upsilon}, \tag{6}$$

where Desertic factors (ώ) is defined by: ό= Amount of pathogens transported by DSE (viruses, bacteria, fungi, and infectious diseases), Ψ= Arid Zones, ϋ= Population living in a climatic zone BW, ï= Climatic Change, ω= Drought, unpaved roads, loose soil, unprotected surfaces and ύ= Estimated deposits of toxic industrial emissions.

$$DEs=WV{*}Du^{PM}\,[\acute{\omega}], \tag{7}$$

where Dust events (DEs) is defined by: WV= Wind Velocity, Du= Duration, PM_{10}= Particulate Matter and ώ= desertic factors.

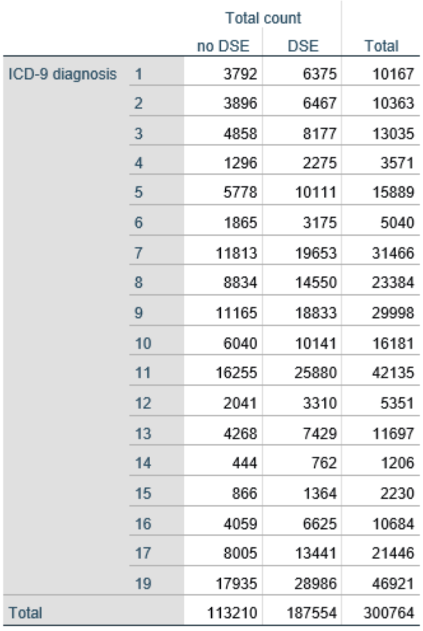
$$PHO = \sum_{i=1,}^{n} (\Phi, \chi, \acute{\omega}){*}Wi\,\{R+ \beta\}^{k} \tag{8}$$

where Projected Health Outcomes for 2030 – 2050 (PHO) is defined by: Φ= SES, χ= demographic factors, ώ= DEs, R= Patient hospitalization, ϐ= Exposure time to PM10 and K= Degree of inflammation in the body. In this equation, PHO refers to the Projected Health Outcomes ether for the present (2010-2014), or projections for 2030 and 2050 due to DEs: denotes the sum of Φ= SES, χ= Congenital factors and ώ= DSE; which affect the R= Patient hospitalization due to; ϐ= Exposure time to PM; and may be exacerbated by ϑ= risk of detrimental structural and cognitive effects, neurodegenerative as well as mental illnesses. Now we have a predictive model to analyze the health outcomes of dust events in a society with Köppen climate classification type BW.

Preliminary results of the investigation have find a preponderant value between the relationship between the location of patients in the metropolitan area of El Paso, TX and the correlation present between the age of patients and their income, which will allow explain how susceptible people are poorer and affected due to possible malfunction of their own houses and susceptibility, which may be not prepared for continuous events associated with DSE, which continuously affect patients (see Figure 11). A primary aspect of our developed model, is that it can adequately estimate the prevalence of a disease or group of diseases associated with a DSE considering its duration and frequency.

Spearman’s correlations indicate that dust events (events with high PM10 and wind speed values) are significant associated to diagnosis with a p value of 0.008. Figure 5 shows that from 2010-2014 there were more hospitalizations in a DE (62%) than in a regular day (RD) (38%). Figure 6 shows that during DE there were 0.4 more hospitalizations due to acute conditions; 0.4 more from chronic conditions and 0.5 more from mental health than in a regular day from 2010-2014. Figure 7 shows the increase in hospitalizations during 8 days after a DE and emphasized the possible effect of PM exposure during these events and hospitalizations; the effect of a DE on hospitalizations might be highest during the actual day of the DE and such effect decreases after that.

Table 1. Comparison of ICD-9 diagnosis in a regular day and in a DE from 2010-2014.



		Total count		
		no DSE	DSE	Total
ICD-9 diagnosis	1	3792	6375	10167
	2	3896	6467	10363
	3	4858	8177	13035
	4	1296	2275	3571
	5	5778	10111	15889
	6	1865	3175	5040
	7	11813	19653	31466
	8	8834	14550	23384
	9	11165	18833	29998
	10	6040	10141	16181
	11	16255	25880	42135
	12	2041	3310	5351
	13	4268	7429	11697
	14	444	762	1206
	15	866	1364	2230
	16	4059	6625	10684
	17	8005	13441	21446
	19	17935	28986	46921
Total		113210	187554	300764

The top 7 causes of admission during a DE from 2010-2014 are: causes of injury & poisoning (15.6%), complications of pregnancy, childbirth, & the puerperium (14%), diseases of the circulatory system (10.5%), diseases of the digestive system (10%), and diseases of the respiratory system (7.8%), diseases of the genitourinary system (5.4%) and mental disorders (5.3%) (Table 3 and Figure 8).

Table 4 shows the top high-risk reasons for hospitalizations, aside from deliveries, respiratory (pneumonia, obstructive chronic bronchitis, asthma) mental (unspecified episodic mood disorder, cerebral artery occlusion, unspecified with cerebral infarction, schizo-affective type schizophrenia unspecified state); cardiovascular (other chest pain, coronary atherosclerosis of native coronary artery, atrial fibrillation); and infectious (urinary tract infection, acute pancreatitis, acute appendicitis without mention of peritonitis) which are affected by bacteria, virus, or due to inflammation.

More patients live in areas with more roads and DE shows to affect the population with all incomes but more frequent patients with a family income of <40,000 dollars; and there are more cases of single born in areas with low income at El Paso, TX from 2010-2014 (Figure 10). There are 59.5 more females hospitalized than males (40.5%) during 2010-2014 at El Paso County (Figure 9).

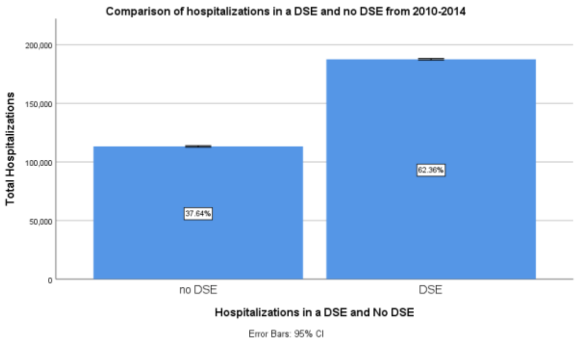


Fig. 2. Comparison of total hospitalizations in regular Days and DE from 2010-2014. There were more hospitalizations in a DE (62%) than in a regular day (RD) (38%).

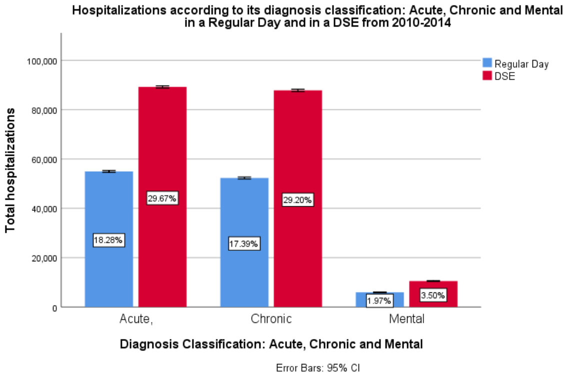


Fig. 3. Comparison of Acute, Chronic and Mental total Admissions code in a DE vs a regular day. During DEs there were 0.4 more hospitalizations due to acute conditions; 0.4 more from chronic conditions and 0.5 more from mental health than in a regular day from 2010-2014.

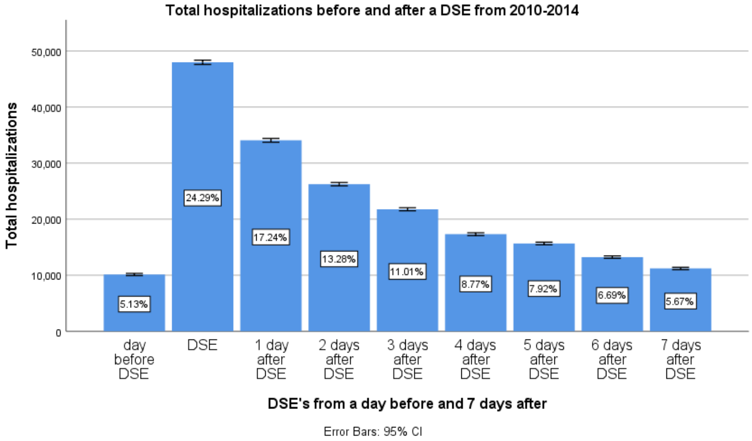


Fig. 4. Total Hospitalization percentage per day before and after DE (each day has many hospitalizations) from 2010-2014. There is an increase in hospitalizations after a DE and emphasized the possible effect of PM_{10} exposure during these events and hospitalizations; the effect of a DE on hospitalizations might be highest during the actual day of the DE and such effect decreases after that.

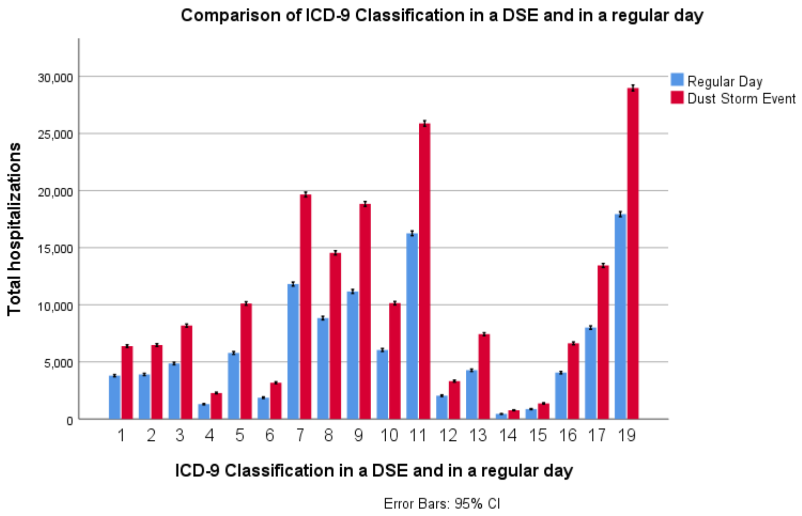


Fig. 5. Comparison of total IDC-9 Admissions code in a DE vs a regular day from 2010-2014. The top 7 causes of admission during a DE from 2010-2014 are: causes of injury & poisoning (15.6%), complications of pregnancy, childbirth, & the puerperium (14%), diseases of the circulatory system (10.5%), diseases of the digestive system (10%), and diseases of the respiratory system (7.8%), diseases of the genitourinary system (5.4%) and mental disorders (5.3%).

Table 2. List of most frequent ICD-9 diagnosis during DEs due to high-risk Respiratory, Mental, Cardiovascular and Infection causes from 2010-2014.

	Top reasons for hospitalizations	**ICD-9**
Respiratory	Pneumonia, organism unspecified	486
	Obstructive chronic bronchitis with (acute) exacerbation	491.21
	Asthma, unspecified type, with (acute) exacerbation	493.92
Mental	Unspecified episodic mood disorder	296.90
	Cerebral artery occlusion, unspecified with cerebral infarction	434.91
	Schizo-affective type schizophrenia unspecified state	295.90
Cardiovascular	Other chest pain	786.59
	Coronary atherosclerosis of native coronary artery	414.01
	Atrial fibrillation	427.31
Infection	Urinary tract infection, site not specified	599.0
	Acute pancreatitis	577.0
	Acute appendicitis without mention of peritonitis	540.9

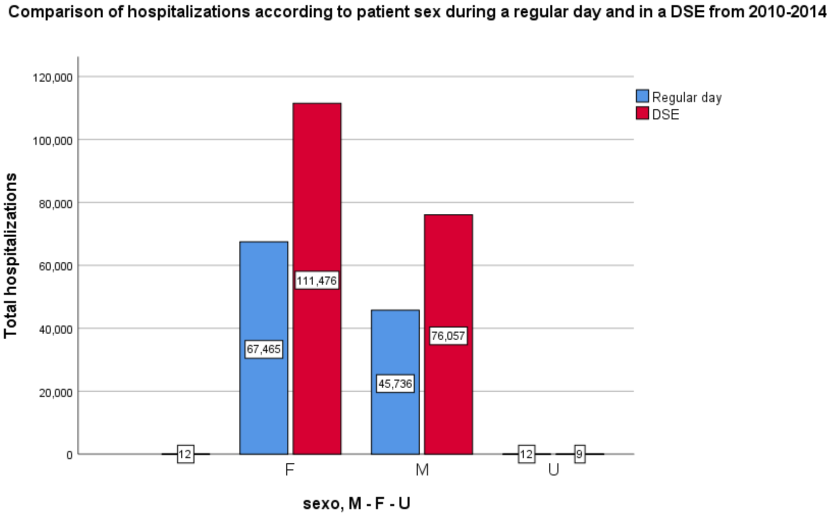
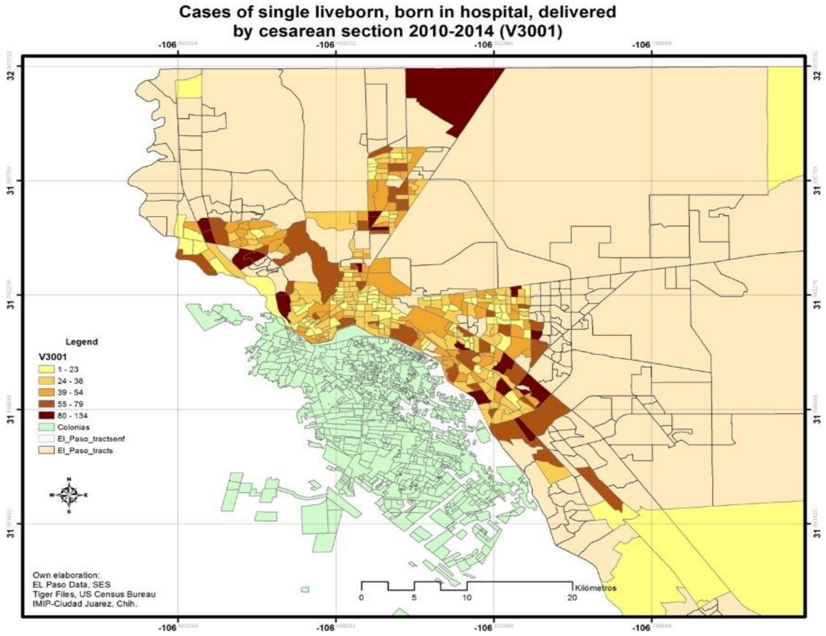

Comparison of hospitalizations according to patient sex during a regular day and in a DSE from 2010-2014

Fig. 9. Comparison of total hospitalizations by Female, Male and Unknown in a RD and in a DE from 2010-2014. There are 59.5 more females hospitalized than males (40.5%) during 2010-2014 at El Paso County.


Cases of single liveborn, born in hospital, delivered by cesarean section 2010-2014 (V3001)

Own elaboration:
EL Paso Data, SES
Tiger Files, US Census Bureau
IMIP-Ciudad Juarez, Chih.

Fig. 10. Map of cases of single born by cesarean section during DEs at El Paso, TX from 2010-2014. More patients live in areas with heavily trafficked roads and DE shows to affect the population with all incomes but more frequent patients with a family income of <40,000 dollars; and there are more cases of single born in areas with low income at El Paso, TX from 2010-2014.

5 Conclusions and Future Work

Several studies have tried to explain the most relevant aspects of the adverse outcomes of DEs in climates type BW, however none had proposed a reliable model associated with the numerical prediction of the present and projected impacts for 2020 and 2050. This research discusses a multifactorial problem, which requires a multivariate analysis which will be elaborated in the following research phase. In addition, we will Investigate whether dust exposure to PM_{10} during a DE (day of and 7 days after) between 2010 and 2014 is associated with hospital admissions due to acute or accelerated disease progression of neurodegenerative diseases (Parkinson's, Alzheimer's, and Huntington's), mental illness (depression and anxiety) and OD (e.g. respiratory, cardiovascular, infectious diseases and top diagnosis significantly associated by DE -diagnosis listed in the ICD9) in El Paso, TX. In addition, we will look into the biological plausibility of these diseases in order to establish a cause-and-effect relationship between PM_{10} during a DE and each significantly associated disease.

Note *: The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

References

1. Adler, N., Pantell, M.S., O'Donovan, A.: Educational attainment and late life telomere length in the Health, Aging and Body Composition Study. Brain Behav. Immun. 27, 15–21 (2013)
2. Agier, L., Deroubaix, A., Martiny, N., Yaka, P., Djibo, A., Broutin, H.: Seasonality of meningitis in Africa and climate forcing: Aerosols stand out. Journal of the Royal Society Interface, 10(79) (2013)
3. Aguilar, M., Bhuket, T., Torres, S., Liu, B., Wong, R.J.: Prevalence of the metabolic syndrome in the United States, 2003–2012. JAMA: J. Am. Med. Assoc. 313, 1973–1974 (2015)
4. Al, B., Bogan, M., Zengin, S., Sabak, M., Kul, S., Oktay, M. M., … Vuruskan, E.: Effects of Dust Storms and Climatological Factors on Mortality and Morbidity of Cardiovascular Diseases Admitted to ED. Emergency Medicine International, 2018(Dm), 1–7 (2018).http://doi.org/10.1155/2018/3758506
5. Alexander, R., Nugent, C., Nugent, K.: The Dust Bowl in the US: An Analysis Based on Current Environmental and Clinical Studies. The American Journal of the Medical Sciences, 356(2), 90–96 (2018). http://doi.org/10.1016/J.AMJMS.2018.03.015
6. Anderson, HR, Favarato, G, Atkinson, RW.: Long-term exposure to air pollution and the incidence of asthma: meta-analysis of cohort studies. Air Qual Atmos Health 6:47–56 (2013)
7. Badii, M.H., Guillen, A., Abreu, J. L.: Tamaño Óptimo de Muestra en Ciencias Sociales y Naturales Optimal Simple Size (OSS) in Social and Natural Sciences, 9(2), 41–51 (2014).
8. Bai, Y., Sun, Q.: Macrophage recruitment in obese adipose tissue. Obes. Rev. 16, 127–136 (2015). http://dx.doi.org/10.1111/obr.12242
9. Barrett, B., Charles, J. W., Temte, J. L.: Climate change, human health, and epidemiological transition. Preventive Medicine, 70, 69–75 (2015). http://doi.org/10.1016/j.ypmed.2014.11.013
10. Behzad, H., Mineta, K., Gojobori, T., Arabia, S., Arabia, S., Sciences, M., Gojobori, T.: Global Ramifications of Dust and Sandstorm Microbiota. Genome Biology and Evolution, 10(July 2018), 1970–1987 (2018). http://doi.org/10.1093/gbe/evy134/5046809

11. Bernier, S.A., Gill, T.E., Peterson, R.E.: Climatology of dust in the Southern High Plains of Texas. In: Busacca, A. (Ed.), Proceedings of the International Conference on Dust Aerosols, Loess Soils and Global Change, Seattle, October 1998. Washington State University College of Agriculture and Home Economics Miscellaneous Publication, vol. 190, pp. 4–7 (1998)
12. Bhagia, L. J.: Non-occupational exposure to silica dust. Indian journal of occupational and environmental medicine, 16(3), 95-100 (2012)
13. Bozlaker, A., Peccia, J., Chellam, S.: Indoor/Outdoor Relationships and Anthropogenic Elemental Signatures in Airborne PM2.5 at a High School: Impacts of Petroleum Refining Emissions on Lanthanoid Enrichment. Environmental Science & Technology, 51(9), 4851–4859 (2017). https://doi.org/10.1021/acs.est.6b06252
14. Braveman, P. A., Kumanyika, S., Fielding, J., LaVeist, T., Borrell, L. N., Manderscheid, R., Troutman, A.: Health disparities and health equity: The issue is justice. American Journal of Public Health, 101(SUPPL. 1), 149–155 (2011). http://doi.org/10.2105/AJPH.2010.300062
15. Brodie, EL, et al.: Urban aerosols harbor diverse and dynamic bacterial populations. Proc Natl Acad Sci USA. 104(1):299–304 (2007)
16. Brook, R.D., Rajagopalan, S., Pope 3rd, C.A.: Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. Circulation 121 (21), 2331–2378 (2010)
17. Brown, E. G., Selma,G., Laybourn, R. L.: Dust Storms and Their Possible Effect on Health. Public Health Reports, 50(40), 20 (1935)
18. Calderón-Garcidueñas, L., Kulesza, R. J., Doty, R. L., D'Angiulli, A., Torres-Jardón, R.: Megacities air pollution problems: Mexico City Metropolitan Area critical issues on the central nervous system pediatric impact. Environmental Research, 137, 157–169 (2015). https://doi.org/10.1016/j.envres.2014.12.012
19. Carvalho-Oliveira, R., Pires-Neto, R. C., Bustillos, J. O., Macchione, M., Dolhnikoff, M., Saldiva, P. H., Garcia, M. L.: Chemical composition modulates the adverse effects of particles on the mucociliary epithelium. Clinics (Sao Paulo, Brazil), 70(10), 706-13 (2015)
20. Chan, Y., Van Nostrand, JD, Zhou, J., Pointing, SB, Farrell, RL.: Functional ecology of an antarctic dry valley. Proc Natl Acad Sci USA. 110(22):8990–8995 (2013)
21. Chan, Y., Teng, J. C., Liu, T., Peng, Y., Chan, Y.: Asian dust storms and diabetes hospitalization: a nationwide population-based study, 1 (2018)
22. Chen, P.S., Tsai, F.T., Lin, C.K., Yang, C.Y., Chan, C.C., Young, C.Y., Lee, C.H.: Ambient influenza and avian influenza virus during dust storm days and background days. Environ. Health Perspective. 118, 1211–1216 (2010)
23. Chiarelli, S. P., Amador Pereira, L. A., Nascimento Saldiva, P. H. do, Ferreira Filho, C., Bueno Garcia, M. L., Ferreira Braga, A. L., Conceição Martins, L.: The association between air pollution and blood pressure in traffic controllers in Santo André, São Paulo, Brazil. Environmental Research, 111(5), 650–5 (2011). https://doi.org/10.1016/j.envres.2011.04.007
24. Chien, L.C., Yang, C.H., Yu, H.L.: Estimated effects of Asian dust storms on spatiotemporal distributions of clinic visits for respiratory diseases in Taipei children (Taiwan). Environ. Health Perspect. 120, 1215e1220 (2012)
25. Chung, H., Sobsey, M.D.: Comparative survival of indicator viruses and enteric viruses in seawater and sediment. Water Sci. Technol. 27, 425–428 (1993)
26. Colls, J.: Air Pollution: Measurement, Modelling and Mitigation, Second Edition. London: CRC Press (2002)
27. Cooper, L. A., Purnell, T. S., Ibe, C. A., Halbert, J. P., Bone, L. R., Carson, K. A., Levine, D. M.: Reaching for Health Equity and Social Justice in Baltimore: The Evolution of an Academic-Community Partnership and Conceptual Framework to Address Hypertension Disparities. Ethnicity & Disease, 26(3), 369–78 (2016). http://doi.org/10.18865/ed.26.3.369

28. Cox, C.S.: Stability of Airborne Microbes and Allergens. In: Cox, C.S., Wathes, C.M. (Eds.), Bioaerosols Handbook. Lewis Publishers, London, UK (1995)
29. Despres, V. R., Huffman, J. A., Burrows, S. M., Hoose, C., Safatov, A. S., Buryak, G., Frohlich-Nowoisky, J., Elbert, W., Andreae, M. O., Pöschl, U., Jaenicke, R.: Primary biological aerosol particles in the atmosphere: a review, Tellus Ser. B, 64, 15598, doi:10.3402/tellusb.v64i0.15598 (2012)
30. Dostert, C., Pétrilli, V., Van Bruggen, R., Steele, C., Mossman, B. T., Tschopp, J.: Innate immune activation through Nalp3 inflammasome sensing of asbestos and silica. Science, 320(5876), 674–677 (2008). http://doi.org/10.1126/science.1156995
31. Ebrahimi, S. J. A., Ebrahimzadeh, L., Eslami, A., Bidarpoor, F.: Effects of dust storm events on emergency admissions for cardiovascular and respiratory diseases in Sanandaj, Iran. Journal of Environmental Health Science and Engineering, 12(1), 1–5 (2014). http://doi.org/10.1186/s40201-014-0110-x
32. EPA Homepage. Criteria Air Pollutants. Retrieved from https://www.epa.gov/criteria-air-pollutants/naaqs-table, last accessed 2018/1/9.
33. EPA Homepage, Reference News Release: Oil Refiners to Reduce Air Pollution at Six Refineries Under Settlement with EPA and Department of Justice. Retrieved from https://www.epa.gov/enforcement/reference-news-release-oil-refiners-reduce-air-pollution-six-refineries-under-settlement, last accessed 2018/7/20.
34. EPA b Homepage, Criteria Air Pollutants. Retrieved from https://www.epa.gov/criteria-air-pollutants/naaqs-table, last accessed 2018/1/9)
35. Etemadifar Z, Gholami M, Derikvand P.: UV-resistant bacteria with multiple-stress tolerance isolated from desert areas in Iran. Geomicrobiol J. 33(7):1–598 (2016)
36. Finkel, M. L.: The impact of oil sands on the environment and health. Environmental Science & Health, 3, 52–55 (2018). https://doi.org/10.1016/J.COESH.2018.05.002
37. Franchini, M., Mannucci, P. M.: Inhibitors of propagation of coagulation (factors VIII, IX and XI): A review of current therapeutic practice. British Journal of Clinical Pharmacology, 72(4), 553–562 (2011). http://doi.org/10.1111/j.1365-2125.2010.03899.x
38. Franzetti, A, Gandolfi, I, Gaspari, E, Ambrosini, R, Bestetti, G.: Seasonal variability of bacteria in fine and coarse urban air particulate matter. Appl Microbiol Biotechnol. 90(2):745–753 (2011)
39. Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., Denier Van Der Gon, H., Facchini, M. C., Gilardoni, S.: Particulate matter, air quality and climate: Lessons learned and future needs. Atmospheric Chemistry and Physics, 15(14), 8217–8299 (2015). http://doi.org/10.5194/acp-15-8217-2015
40. Garcia, G. J. M., Schroeter, J. D., Kimbell, J. S.: Olfactory deposition of inhaled nanoparticles in humans. Inhalation Toxicology, 27(8), 394–403 (2015). http://doi.org/10.3109/08958378.2015.1066904
41. García, J. H., Li, W.-W., Arimoto, R., Okrasinski, R., Greenlee, J., Walton, J., Sage, S.: Characterization and implication of potential fugitive dust sources in the Paso del Norte region. Science of the Total Environment, 325(1), 95–112 (2004). http://doi.org/10.1016/j.scitotenv.2003.11.011
42. Gee, G. C., Walsemann, K. M., Brondolo, E.: A life course perspective on how racism may be related to health inequities. American Journal of Public Health, 102(5), 967–974 (2012). http://doi.org/10.2105/AJPH.2012.300666
43. Ghio, A. J., Huang, Y.-C. T.: Exposure to Concentrated Ambient Particles (CAPs): A Review. Inhalation Toxicology, 16(1), 53–59 (2004). http://doi.org/10.1080/08958370490258390
44. Gill, T.E., Stout, J.E., Peinado, P.: Composition and characteristics of aerosols in the Southern High Plains of Texas, USA. AIP Conference Proceedings 1099: 255- 258 (2009)

45. Ginoux, P., Prospero, J., Gill, T., Hsu, N., Zhao, M.: Globalscale attribution of anthropogenic and natural dust sources and their emission rates based on modis deep blue aerosol products, Rev. Geophys., 50, RG3005 (2012). http://doi:10.1029/2012RG000388.
46. Global Change Homepage, https://health2016.globalchange.gov/report-guide#figure-90, last accessed 2017/1/6
47. Griffin, D.W.: Atmospheric movement of microorganisms in clouds of desert dust and implications for human health. Clin. Microbiol. Rev. 20, 459–477 (2007)
48. Grineski, S. E., Herrera, J. M., Bulathsinhala, P., Staniswalis, J. G.: Is there a Hispanic Health Paradox in sensitivity to air pollution? Hospital admissions for asthma, chronic obstructive pulmonary disease and congestive heart failure associated with NO2 and PM2.5 in El Paso, TX, 2005–2010. Atmospheric Environment, 119, 314–321 (2015). http://doi.org/10.1016/j.atmosenv.2015.08.027
49. Grineski, S. E., Staniswalis, J. G., Bulathsinhala, P., Peng, Y., Gill, T. E.: Hospital admissions for asthma and acute bronchitis in El Paso, Texas: Do age, sex, and insurance status modify the effects of dust and low wind events? Environmental Research, 111(8), 1148–1155 (2011). http://doi.org/10.1016/j.envres.2011.06.007
50. Gurgueira, S.A., Lawrence, J., Coull, B., Murthy, G., Gonzalez-Flecha, B.: Rapid increases in the steady-state concentration of reactive oxygen species in the lungs and heart after particulate air pollution inhalation. Environ. Health Perspect. 110, 749–755 (2002). http://dx.doi.org/10.1289/ehp.02110749
51. Halonen, J. I., Blangiardo, M., Toledano, M. B., Fecht, D., Gulliver, J., Anderson, H. R., Tonne, C.: Long-term exposure to traffic pollution and hospital admissions in London. Environmental Pollution (Barking, Essex: 1987), 208(Pt A), 48–57 (2016). https://doi.org/10.1016/j.envpol.2015.09.051
52. Hansen, J, Sato, M, Ruedy, R, Lo, K, Lea, DW, Medina-Elizade, M.: Global temperature change. Proc Natl Acad Sci USA;103:14288–93 (2006)
53. Heusinkveld, H. J., Wahle, T., Campbell, A., Westerink, R. H. S., Tran, L., Johnston, H., … Schins, R. P. F.: Neurodegenerative and neurological disorders by small inhaled particles. NeuroToxicology, 56, 94–106 (2016). http://doi.org/10.1016/j.neuro.2016.07.007
54. Holliday, V.T.: The Blackwater Draw Formation (Quaternary): a 1.4-plus-m.y. record of aeolian sedimentation and soil formation on the Southern High Plains. Geol. Soc. Am. Bull. 101, 1598–1607 (1989)
55. Hosiokangas, J., Vallius, M., Ruuskanen, J., Mirme, A., Pekkanen, J.: Resus- pended dust episodes as an urban air-quality problem in subarctic regions. Scandinavian Journal of Work and Environmental Health 30, 28–35 (2004)
56. Howard, G., Peace, F., Howard, V. J.: The contributions of selected diseases to disparities in death rates and years of life lost for racial/ethnic minorities in the United States, 1999-2010. Preventing Chronic Disease, 11, E129 (2014). http://doi.org/10.5888/pcd11.140138
57. Huang, M, Wang, W, Chan, CY, Cheung, KC, Man, YB, Wang, X, et al.: Contamination and risk assessment (based on bioaccessibility via ingestion and inhalation) of metal(loid)s in outdoor and in- door particles from urban centers of Guangzhou, China. Sci Total Environ;479-480:117-124 (2014)
58. Husar, RB, et al.: Asian dust events of April 1998. J Geophys Res. 106(D16):18317–18330 (2001)
59. Intergovernmental Panel on Climate Change (IPCC) a: Climate Change 2007: Impacts, Adaptation and Vulnerability. In: Parry, M.L., Canziani, O.F., Palutikof, J.P., van der Linden, P.J., Hanson, C.E. (Eds.), Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK (2007a)
60. Intergovernmental Panel on Climate Change (IPCC) b: Climate change 2007: impacts, adaptation and vulnerability contribution of Working Group II to the Fourth Assess- ment

Report of the Intergovernmental Panel on Climate Change. In: Parry ML, Canzani OP, Palutikof JP, et al, editors. Cambridge, New York: Cambridge University Press (2007)

61. Jacob, D. J., Darrel, A.: (2009). Effect of climate change on air quality. Atmospheric Environment 43(1): 51-63 (2007b)
62. Kanatani, K. T., Ito, I., Al-Delaimy, W. K., Adachi, Y., Mathews, W. C., Ramsdell, J. W., Yamamoto, J.: Desert dust exposure is associated with increased risk of asthma hospitalization in children. American Journal of Respiratory and Critical Care Medicine, 182(12), 1475–1481 (2010). http://doi.org/10.1164/rccm.201002-0296OC
63. Kaplan, R. M.: Behavior change and reducing health disparities. Preventive Medicine, 68, 5–10 (2014). http://doi.org/10.1016/j.ypmed.2014.04.014
64. Karagulian, F., Belis, C. A., Dora, C. F. C., Prüss-Ustün, A. M., Bonjour, S., Adair-Rohani, H., Amann, M.: Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. Atmospheric Environment, 120, 475–483 (2015). http://doi.org/10.1016/j.atmosenv.2015.08.087
65. Karanasiou, A., Moreno, N., Moreno, T., Viana, M., de Leeuw, F., Querol, X.: Health effects from Sahara dust episodes in Europe: Literature review and research gaps, Environment Int., 47, 107–114 (2012). doi:10.1016/j.envint.2012.06.012
66. Keil, D. E., Buck, B., Goossens, D., McLaurin, B., Murphy, L., Leetham-Spencer, M., DeWitt, J. C.: Nevada desert dust with heavy metals suppresses IgM antibody production. Toxicology Reports, 5(June 2017), 258–269 (2018). http://doi.org/10.1016/j.toxrep.2018.01.006
67. Keil, D. E., Buck, B., Goossens, D., Teng, Y., Pollard, J., McLaurin, B., DeWitt, J.: Health effects from exposure to atmospheric mineral dust near Las Vegas, NV, USA. Toxicology Reports, 3, 785–795 (2016). http://doi.org/10.1016/j.toxrep.2016.09.009
68. Keil, D.E., Buck, B., Goossens, D., Teng, Y., Spencer, M., Murphy, L., Pollard, J., Eggers, M., McLaurin, B., Gerads, R., DeWitt, J.: Immunotoxicological and neurotoxicological profile of health effects following subacute exposure to geogenic dust from sand dunes at the Nellis Dunes Recreation Area Las Vegas, NV. Toxicol. Appl. Pharmacol. 2016;291:1–12 (2016)
69. Khan, R. K., Strand, M. A.: Road dust and its effect on human health: a literature review. Epidemiology and health, 40, e2018013(2018). Doi:10.4178/epih.e2018013
70. Khaniabadi, Y. O., Fanelli, R., De Marco, A., Daryanoosh, S. M., Kloog, I., Hopke, P. K., Goudarzi, G. (2017). Hospital admissions in Iran for cardiovascular and respiratory diseases attributed to the Middle Eastern Dust storms. Environmental Science and Pollution Research, 24(20), 16860–16868. http://doi.org/10.1007/s11356-017-9298-5
71. Kinge, J. M., Steingrímsdóttir, Ó. A., Strand, B. H., Kravdal, Ø.: Can socioeconomic factors explain geographic variation in overweight in Norway? SSM - Population Health, 2, 333–340 (2016). http://doi.org/https://doi.org/10.1016/j.ssmph.2016.04.010
72. Kioumourtzoglou, M. A., Schwartz, J. D., Weisskopf, M. G., Melly, S. J., Wang, Y., Dominici, F., Zanobetti, A.: Long-term PM Exposure and Neurological Hospital Admissions in the Northeastern United States. Environmental Health Perspectives, 124(1), 23–29 (2015). http://doi.org/10.1289/ehp.1408973
73. Kirkland, TN, Fierer, J.: Coccidioidomycosis: a reemerging infectious disease. Emerg Infect Dis. 2(3):192–199 (1996)
74. Krieger, N.: Living and Dying at the Crossroads: Racism, Embodiment, and Why Theory Is Essential for a Public Health of Consequence. American Journal of Public Health, 106(5), 832–833 (2016). http://doi.org/10.2105/AJPH.2016.303100
75. Lee, J. A., Gill, T. E.: Multiple causes of wind erosion in the Dust Bowl. Aeolian Research, 19(November 2006), 15–36 (2015). http://doi.org/10.1016/j.aeolia.2015.09.002
76. Lee, J. A., Baddock, M. C., Mbuh, M. J., Gill, T. E.: Geomorphic and land cover characteristics of aeolian dust sources in West Texas and eastern New Mexico, USA. Aeolian Research, 3(4), 459–466 (2012). http://doi.org/10.1016/j.aeolia.2011.08.001

77. Lee, J. A., Gill, T. E., Mulligan, K. R., Dominguez Acosta, M., Perez, A. E.: Land use/land cover and point sources of the 15 December 2003 dust storm in southwestern North America. Geomorphology, 105(1–2), 18–27 (2009). https://doi.org/10.1016/j.geomorph.2007.12.016
78. Lee, J., Baddock, M., Mbuh, M., Gill, T.: Geomorphic and land cover characteristics of aeolian dust sources in West Texas and eastern New Mexico, USA. Aeolian Research (Vol. 3) (2012). http://doi.org/10.1016/j.aeolia.2011.08.001
79. Lee, J.A., Tchakerian, V.P.: Magnitude and frequency of blowing dust on the Southern High Plains of the United States, 1947–1989. Ann. Assoc. Am. Geogr. 85, 684–693 (1995).
80. Li, J., Kandakji, T., Lee, J. A., Tatarko, J., Blackwell, J., Gill, T. E., Collins, J. D.: Blowing dust and highway safety in the southwestern United States: Characteristics of dust emission “hotspots” and management implications. Science of The Total Environment, 621, 1023–1032 (2018). http://doi.org/10.1016/J.SCITOTENV.2017.10.124
81. Li, R., Kou, X., Geng, H., Xie, J., Tian, J., Cai, Z., Dong, C.: Mitochondrial damage: an important mechanism of ambient PM2.5 exposure-induced acute heart injury in rats. Journal of Hazardous Materials, 287, 392–401 (2015). https://doi.org/10.1016/j.jhazmat.2015.02.006
82. Li, W.-W., Orquiz, R., Garcia, J. H., Espino, T., Pingitore, N. E., Gardea-Torresdey, J., Watson, J. G.: Analysis of temporal and spatial dichotomous PM air samples in the El Paso-Cd. Juarez air quality basin. Journal of the Air & Waste Management Association (1995), 51(11), 1551–60 (2001). http://doi.org/10.1080/10473289.2001.10464377
83. Lim, J.Y., Chun, Y.: The characteristics of Asian dust events in Northeast Asia during the springtime from 1993 to 2004. Global and Planetary Change, 52, 231–247 (2006)
84. Ma, Y., Zhang, H., Zhao, Y., Zhou, J., Yang, S., Zheng, X., Wang, S.: Short-term effects of air pollution on daily hospital admissions for cardiovascular diseases in western China. Environmental Science and Pollution Research, 24(16), 14071–14079 (2017). http://doi.org/10.1007/s11356-017-8971-z
85. Middleton, N. J: Desert dust hazards: A global review. Aeolian Research, 24, 53–63 (2017). https://doi.org/https://doi.org/10.1016/j.aeolia.2016.12.001
86. Morman, S. A., Plumlee, G. S.: The role of airborne mineral dusts in human disease. Aeolian Research, 9, 203–212 (2013). http://doi.org/10.1016/j.aeolia.2012.12.001
87. Morman, S.A., Plumlee, G.S.: Dust and human health. In Mineral Dust (pp. 385-409). Springer, Dordrecht (2014)
88. Nemmar, A., Al-Salam, S., Zia, S., Dhanasekaran, S., Shudadevi, M., Ali, B. H.: Time-course effects of systemically administered diesel exhaust particles in rats. Toxicology Letters, 194(3), 58–65 (2010). http://doi.org/10.1016/J.TOXLET.2010.02.001
89. Nemmar, A., Hoylaerts, M. F., Hoet, P. H.., Vermylen, J., Nemery, B.: Size effect of intratracheally instilled particles on pulmonary inflammation and vascular thrombosis. Toxicology and Applied Pharmacology, 186(1), 38–45 (2003). https://doi.org/10.1016/S0041-008X(02)00024-8
90. Nery, L. E., Florencio, R. T., Sandoval, P. R. M., Rodrigues, R. T., Alonso, G., Mason, G. R.: Additive effects of exposure to silica dust and smoking on pulmonary epithelial permeability: A radioaerosol study with technetium-99m labelled DTPA. Thorax, 48(3), 264–268 (1993). http://doi.org/10.1136/thx.48.3.264
91. Novlan, D. J., Hardiman, M., Gill, T. E.: A Synoptic Climatology of Blowing Dust Events in El Paso, Texas from 1932-2005. American Meteorological Society J., 3 (2007)
92. O’Neill, L. A. J., Dunne, A.: The Interleukin-1 Receptor / Toll-Like Receptor Superfamily, (February), 1–18 (2008)
93. Oberdörster, G., Sharp, Z., Atudorei, V., Elder, A., Gelein, R., Kreyling, W., Cox, C.: Translocation of inhaled ultrafine particles to the brain. Inhalation Toxicology, 16(6–7), 437–445 (2004). https://doi.org/10.1080/08958370490439597

94. Oil Change International Homepage, Refineries, http://refineryreport.org/refineries.php, last accessed 2018/7/5
95. Olvera, H. A., Kubzansky, L. D., Campen, M. J., Slavich, G. M.: Neuroscience and Biobehavioral Reviews Early life stress, air pollution, inflammation, and disease: An integrative review and immunologic model of social-environmental adversity and lifespan health. Neuroscience and Biobehavioral Reviews, 92(May), 226–242 (2018). https://doi.org/10.1016/j.neubiorev.2018.06.002
96. Orgill M.M., Sehmel G.A.: Frequency and diurnal variation of dust storms in the contiguous U.S.A. Atmospheric Environment, 10 pp. 813-825 (1976)
97. Prospero, J.M., Ginoux, P., Torres, O., Nicholson, S.E., Gill, T.E.: Environmental characterization of global sources of atmospheric soil dust identified with the nimbus 7 total ozone mapping spectrometer (toms) absorbing aerosol product. Reviews of Geophysics 40: doi: 10.1029/2000RG000095. issn: 8755-1209 (2002)
98. Prussin A. J., Garcia E. B., Marr L. C.: Total concentrations of virus and bacteria in indoor and outdoor air. Environ Sci Tech Lett. 2(4):84–88 (2015)
99. Reynolds, K., Go, A. S., Leong, T. K., Boudreau, D. M., Cassidy-bushrow, A. E., Fortmann, S. P., Sidney, S.: Trends in Incidence of Hospitalized Acute Myocardial Infarction in the Cardiovascular Research Network (CVRN). The American Journal of Medicine (2016). http://doi.org/10.1016/j.amjmed.2016.09.014
100. Rivas, Jose A. Jr., Thomas E. Gill, Elizabeth J. Walsh, Robert L. Wallace.: Characterization of Dust Transported to El Paso, Texas. Abstracts of the 18th Joint Conference on the Applications of Air Pollution Meteorology with the A&WMA, American Meteorological Society Annual Meeting, Atlanta, GA, February 2014, Abstract # 236614 (2014)
101. Rivera Rivera, N. I., Gill, T. E., Gebhart, K. A., Hand, J. L., Bleiweiss, M. P., Fitzgerald, R. M.: Wind modeling of Chihuahuan Desert dust outbreaks. Atmospheric Environment, 43(2), 347–354 (2009). http://doi.org/10.1016/j.atmosenv.2008.09.069
102. Rivera, N. I. R., Gill, T. E., Bleiweiss, M. P., Hand, J. L.: Source characteristics of hazardous Chihuahuan Desert dust outbreaks. Atmospheric Environment, 44(20), 2457–2468 (2010). https://doi.org/10.1016/J.ATMOSENV.2010.03.019
103. Rodo X, et al.: Tropospheric winds from northeastern china carry the etiologic agent of Kawasaki disease from its source to japan. Proc Natl Acad Sci USA. 111(22):7952–7957(2014).
104. Rodopoulou, S., Chalbot, M. C., Samoli, E., DuBois, D. W., San Filippo, B. D., Kavouras, I. G.: Air pollution and hospital emergency room and admissions for cardiovascular and respiratory diseases in Doña Ana County, New Mexico. Environmental Research, 129, 39–46 (2014). http://doi.org/10.1016/j.envres.2013.12.006
105. Schwarze, P.E., Totlandsdal, A.I., Låg, M., Refsnes, M., Holme, J.A., Øvrevik, J.: Inflammation-related effects of diesel engine exhaust particles: studies on lung cells in vitro. Biomed. Res. Int. 2013, 1–13 (2013). http://dx.doi.org/10.1155/2013/685142
106. Schweitzer, M. D., Calzadilla, A. S., Salamo, O., Sharifi, A., Kumar, N., Holt, G., Mirsaeidi, M.: Lung health in era of climate change and dust storms. Environmental Research, 163(February), 36–42 (2018). http://doi.org/10.1016/j.envres.2018.02.001
107. Shao, Y.: Physics and modelling of wind erosion (2nd edn.). Heidelberg: Springer (2008)
108. Shrey, K., Suchit, A., Deepika, D., Shruti, K., Vibha, R.: Air pollutants: the key stages in the pathway towards the development of cardiovascular disorders. Environmental Toxicology and Pharmacology, 31(1), 1–9 (2011). http://doi.org/10.1016/j.etap.2010.09.002
109. Song, Z., Wang, J., Wang, S.: Quantitative classification of northeast Asian dust events. Journal of Geophysical Research, 112(D4), D04211 (2007). http://doi.org/10.1029/2006JD007048
110. Sujaritpong, S., Dear, K., Cope, M., Walsh, S., Kjellstrom, T.: Quantifying the health impacts of air pollution under a changing climate-a review of approaches and methodology.

International Journal of Biometeorology, 58(2), 149–160 (2014). http://doi.org/10.1007/s00484-012-0625-8

111. Sun, J., Fu, J. S., Huang, K., Gao, Y.: Estimation of future PM2.5- and ozone-related mortality over the continental United States in a changing climate: An application of high-resolution dynamical downscaling technique. Journal of the Air & Waste Management Association, 65(5), 611–623 (2015). http://doi.org/10.1080/10962247.2015.1033068
112. Tang K, et al.: Characterization of atmospheric bioaerosols along the transport pathway of Asian dust during the dust-bioaerosol 2016 campaign. Atmos Chem Phys. 18:7131–7148 (2018). doi: 10.5194/acp-18-7131
113. Tarhan, C., Ozcan, N. S., Ozkan, S. P.: The Relationship between Respiratory Systems' Cases and Environmental Urban Factors. Procedia - Social and Behavioral Sciences, 216, 622–631 (2016). http://doi.org/10.1016/j.sbspro.2015.12.040
114. Texas Climate Descriptions Homepage, http://web2.airmail.net/danb1/texas_climate_descriptions.htm, last accessed 2016/2/6
115. Tili, E., Michaille, J.-J., Wernicke, D., Alder, H., Costinean, S., Volinia, S., Croce, C. M.: Mutator activity induced by microRNA-155 (miR-155) links inflammation and cancer. Proceedings of the National Academy of Sciences of the United States of America, 108(12), 4908–13 (2011). http://doi.org/10.1073/pnas.1101795108
116. Tong DQ, Wang JXL, Gill TE, Lei H, Wang BY.: Intensified dust storm activity and valley fever infection in the southwestern United States. Geophys Res Lett. 44(9):4304–4312 (2017).
117. Tonne, C., Halonen, J. I., Beevers, S. D., Dajnak, D., Gulliver, J., Kelly, F. J., Anderson, H. R.: Long-term traffic air and noise pollution in relation to mortality and hospital readmission among myocardial infarction survivors. International Journal of Hygiene and Environmental Health, 219(1), 72–78 (2016). https://doi.org/10.1016/j.ijheh.2015.09.003
118. Touloumi, G., Atkinson, R., Le Tertre, A., Samoli, E., Schwartz, J., Schindler, C., et al.: Analysis of health outcome time series data in epidemiological studies. Environmetrics 15, 101–117 (2004)
119. Van Pelt, R. S., Zobeck, T. M.: Chemical constituents of fugitive dust. Environmental Monitoring and Assessment, 130(1–3), 3–16 (2007). http://doi.org/10.1007/s10661-006-9446-8
120. Vineis, P., Foraster, F., Hoek, G., Lippsett, M.: Outdoor air pollution and lung cancer: recent epidemiological evidence. Int. J. Cancer 111, 647e652 (2004)
121. Wang Y, et al.: Effects of dust storm events on weekly clinic visits related to pulmonary tuberculosis disease in Minqin, China. Atmos Environ. 127:205–212 (2016)
122. Weil T, et al.: Legal immigrants: invasion of alien microbial communities during winter occurring desert dust storms. Microbiome 5(1):32 (2017)
123. Weintraub, D., Xie, S., Karlawish, J., Siderowf, A.: Differences in depression symptoms in patients with Alzheimer's and Parkinson's diseases: evidence from the 15-item Geriatric Depression Scale (GDS-15). International journal of geriatric psychiatry, 22(10), 1025-30 (2007)
124. Wolf, K., Schneider, A., Breitner, S., Meisinger, C., Heier, M., Cyrys, J., Peters, A.: Associations between short-term exposure to particulate matter and ultrafine particles and myocardial infarction in Augsburg, Germany. International Journal of Hygiene and Environmental Health, 218(6), 535–42 (2015). https://doi.org/10.1016/j.ijheh.2015.05.002
125. Wu, S., Ni, Y., Li, H., Pan, L., Yang, D., Baccarelli, A. A., Guo, X.: Short-term exposure to high ambient air pollution increases airway inflammation and respiratory symptoms in chronic obstructive pulmonary disease patients in Beijing, China. Environ Int, 94, 76–82 (2016). https://doi.org/10.1016/j.envint.2016.05.004
126. Yu, H. L., Chien, L. C., Yang, C. H.: Asian dust storm elevates children's respiratory health risks: A spatiotemporal analysis of children's clinic visits across Taipei (Taiwan). PLoS ONE, 7(7), 1–9 (2012). http://doi.org/10.1371/journal.pone.0041317

127. Zablocki O, Adriaenssens EM, Cowan D.: Diversity and ecology of viruses in hyperarid desert soils. Appl Environ Microbiol. 82(3):770–777 (2016)
128. Zhang, X., Zhao, L., Tong, Q. D., Wu, G., Dan, M., Teng, B.: A Systematic Review of Global Desert Dust and Associated Human Health Effects. Atmosphere (2016). http://doi.org/10.3390/atmos7120158

ISSN 1870-4069

Determining the Optimal and Ideal Helmet for an Italian Scooter Used in a Smart City Considering Cranial Anthropometry and Intelligent Data Analysis

Mariana Martinez-Valencia[1], José-Antonio Vazquez-Lopez[1],
Carolina Hernandez-Navarro[2], Alberto Ochoa[3], Juan-Luis Hernandez-Arellano[3]

[1] Instituto Tecnológico Nacional de México en Celaya, Departamento de Ingeniería Industrial,
Celaya, Guanajuato., Mexico

[2] Instituto Tecnológico Nacional de México en Celaya, Departamento de Ingeniería Mecánica,
Celaya, Guanajuato., Mexico

[3] Universidad Autónoma de Ciudad Juárez, Laboratorio de Diseño Ergonómico de Producto,
Ciudad Juárez, Chihuahua, Mexico

luis.hernandez@uacj.mx

Abstract. The main objective of the present research was to develop an intelligent data analysis from anthropometric data to find tendencies and patrons before stating the design of an optimal and ideal helmet to be used in an Italian scooter. Using a database of the anthropometric properties of the craniofacial structures of a sample of students in a border society. A set of 14 craniofacial dimensions were obtained using 13 reference anthropometric points (Glabella, Vertex, Opisthocranion, Eurion, Alare, Gnathion, Nasion, Nasoespinhale, Frontotemporale, Porion, Exocanthion, Endocantion, and Zygion). We used a ROSSCRAFT anthropometer model CAMPBELL 10 RC-10, a ROSSCRAFT metallic tape, and an ErgoTechMx brand ErgoMeasure model anthropometer. 130 students, 69 men, and 61 women enrolled in said University were measured. The values of mean, standard deviation, maximum and minimum were calculated. Finally, we analyze the data obtained to determine the ideal thickness of the helmet and how it can help reduce deaths in road accidents.

Keywords: anthropometry, intelligent data analysis, product design.

1 Introduction

1.1 Background

An Italian scooter is a type of motorized two-wheeled vehicle, with an open frame in which the driver sits without straddling any part of the engine. Most modern scooters have smaller wheels than motorcycles, between 12 and 15 inches (30-37.5 cm) in diameter and were ideally created for urban use and are currently being promoted for mass use in a Smart City, as is shown in Figure 1.

Fig. 1. An Italian Scooter used in diverse societies.

In terms of its design and mechanics, in contrast to most of the motorcycles, scooters tend to have a carriage, including a frontal protection for the legs and a body that hides all or most of the mechanics. The classic design of the scooter has a flat floor for the driver's feet and often includes some integrated storage space, either under the seat, in the front leg protection or both.

Most old scooters and some recent retro models have a manual transmission with the shift lever and the clutch on the left handlebar. The motor of the scooter is usually under the seat with a continuous variable transmission transferring the power to the rear wheel, often in a front axle arrangement that allows the rear of the engine to oscillate vertically in conjunction with the movement of the rear wheel.

Until relatively recently, most modern scooters had air- cooled two-stroke engines with fuel and air blends by carburizing, although some high-end ones are water-cooled, such as the Honda FC50 or the Yamaha YG50 of 2002. Most scooters have smaller engines than motorcycles (between 30ccs and 250ccs with a single cylinder). Those of 49ccs or less displacement are classified in most countries as a moped and are subject to safety restrictions and reduced rates. Since the 90s are increasingly common four-stroke engines that allow compliance with the strictest emission controls, the mixture of fuel and air by injection and models with larger displacement: 300cc, 400cc, 500cc and up to 800cc that are usually called maxiscooters. Significant examples of these models are the Kymco SuperDink 300, Yamaha T-Max, SYM Joymax 300i or the Piaggio X9. Recently, they are also appearing in the market and they begin to popularize the scooters of three wheels to which we should consider tricycles and not bicycles like the Piaggio MP3 or the Yamaha Tricity 125.

1.2 Anthropometry

Anthropometry is a simple and reliable method to quantify the size and proportions of the body by measuring the length, width, circumference, and thickness of the skin fold of the body [1], [2]. Updated anthropometric data are very important for any population/civilization since the determination of the correct dimensions of objects depends to a large extent on the improvement of existing anthropometric data [3]. Anthropometry practices aim, through the correct use of the conventional anthropometer, to identify, correctly measure the body segments of the human being, perform the appropriate statistical calculation to integrate the data acquired in an anthropometric chart and correctly apply the measurements in the design of products, objects, spaces and work stations, among other applications [4].

Craniofacial anthropometry has become an important tool used by both clinical geneticists and reconstructive surgeons [5]. The study of the craniofacial anthropometric characteristics is of fundamental importance to solve problems related to the identification and quantification of syndromic clinical characteristics, the planning of treatments and reconstructive surgeries, the control of the operative results and the evaluation of the longitudinal change of the skull.

The craniometric analysis is done by locating the anthropometric points and determining the anthropometric measurements, which have already been established in the literature, these can usually be obtained by simple rules, calibers and other specific tools [6].

1.3 Data Mining

Data Mining is the extraction of hiding and predictable information inside great data bases, also, is a powerful new technology with great potential to help to the companies or organizations to focus on the most important information in their Bases of Information (Data Warehouse). Data Mining tools predict future tendencies and behaviors, allowing businesses to make proactive decisions leaded by knowledge-driven information. The automated prospective analyses offered by a product thus go beyond past events provided by retrospective typical tools of decision support systems. Data Mining tools can respond to questions of businesses that traditionally consume too much time to be solved and to which the users of this information almost are not willing to accept. These tools explore the data bases searching for hidden patterns, finding predictable information that sometimes an expert cannot find because this is outside expectations as is shown in Figure 3.

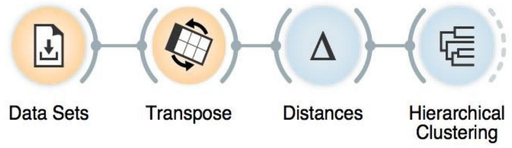


Fig. 3. Intelligent Data Analysis.

The motivation to make an approach by means of applications with Data Mining is based on previous works of Social Data Mining in this research area. This research area emphasizes the role of the collective analysis of conduct effort, rather that the individual one. A social tendency results from the decisions of many individuals joined only in the location in where they choose to coexist, yet this, still it reflects a rough notion of what the researchers of the area find of what could be a correct and valid social tendency. The social tendency reflects the history of the use of a collective behavior and serves as a base to characterize the behavior of future descendants. The Data Mining approaches for social aspects look for analogous situations in the behavior registers. The investigators look for situations where the groups of people are producing computer registers (such as documents, USENET messages, or Web sites and links to groups with a specific profile) like part of its normal activity. The potentially useful information implicit in these files is identified and the computer techniques to display the results are designed. Thus, the computer discovers and makes explicit the "social tendencies through the time" created by a particular type of community.

1.4 Objectives

Once the background has been stablished, the main objectives of this study are to determine anthropometrics reference points and to develop and intelligent data analysis, both will be used in the design of an ergonomic helmet for an Italian scooter.

2 Methodology

2.1 Study Design

A transversal, descriptive and correlational study is present. A team of 3 anthropometrists was trained to perform craniofacial anthropometric measurements. The measurements were made in the Product Ergonomic Design Laboratory of the UACJ. The ethics committee of the Autonomous University of Ciudad Juarez, Mexico, reviewed and approved the study. The participants signed a consent form accepting their participation in the study, as well as the absence of health risks when participating in the study. The information collected was treated confidentially and was used only for academic purposes.

2.2 Sample

A convenience sample consisted of 130 students of the UACJ participated in the study. The age was between 18 and 30 years. The inclusion criteria were that the participants were free of physical injuries, without craniofacial fractures, deformities, or having undergone surgeries in the skull. The exclusion criteria included facial fractures, defects or posttraumatic deformities, congenital malformations, development of asymmetries, or the presence of implants.

2.3 Materials

An anthropometer brand ROSSCRAFT model CAMPBELL 10 RC-10, range 18 cm, and a metal tape ROSSCRAFT mark, for anthropometric use range 200 cm. The equipment has an accuracy of 0.5 mm, is certified and calibrated by the manufacturer.

ErgoMeasure vertical anthropometer range to 500 cm was used. The precision of the instrument is ±1 mm.

2.4 Variables

Fourteen dimensions of the head corresponding to distances between thirteen craniofacial anthropometric reference points (Glabella, Vertex, Opisthocranion, Eurion, Alare, Gnathion, Nasion, Nasoespinhale, Frontotemporale, Porion, Exocanthion, Endocantion, and Zygion) were measured, which are described in Table 1. Table 2 shows the names of the craniofacial anthropometric dimensions. Figure 2 shows the cranial and head dimensions used in the study.

Table 1. Anthropometric anatomical points.

N°	Name	Abbreviation
1	Eurion	Eu
2	Vertex	V
3	Opisthocranion	Op Ft
4	Frontotemporale	
5	Porion	Po
6	Gnathion	Gn
7	Zygion	Zy
8	Glabela	G
9	Nasion	N
10	Nasoespinhale	Ns
11	Alare	Al
12	Exocanthion	Ex
13	Endocanthion	It

Table 2. Craniofacial anthropometric measurements.

N°	Anthropometric Dimension Reference	Anthropometric Points
1	Head Width	Eu-Eu
2	Skull length	G-Op
3	Nasal amplitude	Al-Al
4	Nasal height	N-Ns
5	Facial amplitude	N-Gn
6	Forehead width	Ft-Ft
7	External inter-chamber distance	Ex-Ex
8	Internal intercantal distance	En-En
9	Facial widht	Zy-Zy

N°	Anthropometric Dimension Reference	Anthropometric Points
10	Head circumference	G-Op
11	Length G-V-Op	G-V-Op
12	Length Eu-V-Eu	Eu-V-Eu
13	Length V-Gn	V-Gn
14	Length V-Po	V-Po
15	Head height	V-N

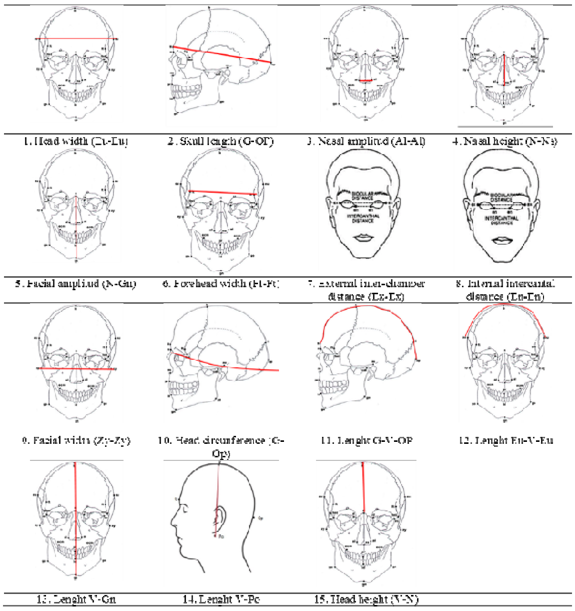
2.5 Data Analysis

Mean, maximum, minimum, and standard deviation were calculated for each item. The intelligent data analysis was developed using the Weka software, a program that contains multiple machine learning algorithms. Attributes that have less variation between instances to dismiss them from the analyzes were searched. A clustering was performed to group the data and find patterns in the data set.

3 Results Using Intelligent Data Analysis

3.1 Descriptive Analysis

Table 3 shows the mean, minimum, maximum and standard deviation of the collected data for men and women.



1. Head width (Eu-Eu)
2. Skull length (G-OP)
3. Nasal amplitud (Al-Al)
4. Nasal height (N-Ns)
5. Facial amplitud (N-Gn)
6. Forehead width (Ft-Ft)
7. External inter-chamber distance (Ex-Ex)
8. Internal intercantal distance (En-En)
9. Facial width (Zy-Zy)
10. Head circumference (G-Op)
11. Lenght G-V-OP
12. Lenght Eu-V-Eu
13. Lenght V-Gn
14. Lenght V-Po
15. Head height (V-N)

Fig. 2. Cranial and head dimensions used in the study.

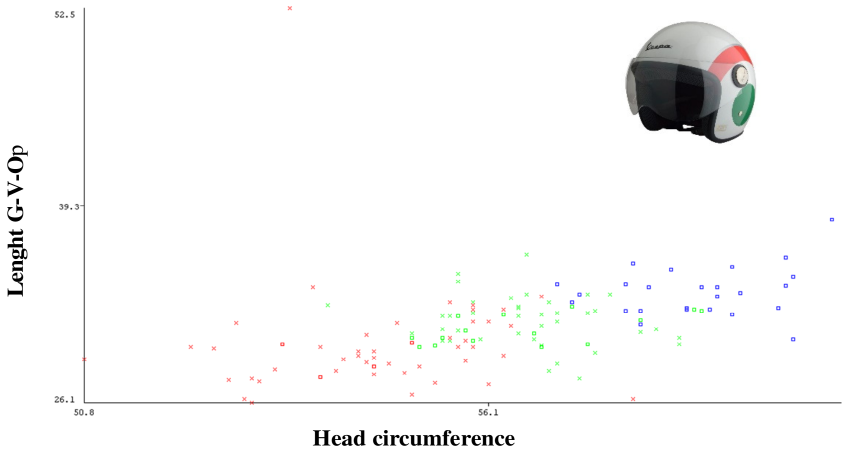


Fig. 3. Data Mining and Clustering Analysis.

Table 3. Descriptive statistics.

Dimension	Men (n=69)			Women (n=61)		
	MIN	MAX	Mean ± SD	MIN	MAX	Mean ± SD
Eu-Eu	142	170	156.62 ± 6.13	139	166	150 ± 5.65
G-Op	174	219	194.70 ± 9.05	171	201	185.66 ± 7.05
Al-Al	27	47	34.59 ± 3.02	26	36	31.97 ± 2.28
N-Ns	43	59	51.96 ± 3.34	38	57	46.92 ± 3.90
N-Gn	103	140	124.04 ± 6.93	99	128	112.54 ± 6.68
Ft-Ft	75	125	105.35 ± 9.46	77	126	99.2 ± 10.08
Ex-Ex	95	115	104.41 ± 4.34	87	112	99.67 ± 5.20
En-En	26	36	31.25 ± 2.55	24	35	29.34 ± 2.85
Zy-Zy	130	158	142.84 ± 6.85	105	154	133.26 ±9.57
Perímeter G-Op	534	614	573.52 ±17.07	508	601	552.57 ±17.21
G-V-Op	277	383	321.78 ±20.86	261	525	303.74 ±34.70
Eu-V-Eu	286	370	320.29 ±16.53	263	338	301.64 ±18.01
V-Gn	197	249	224.39 ±11.68	185	230	204.61 ±11.08
V-Po	119	158	140.39 ± 9.14	118	160	135.53 ±9.38
V-N	79	151	103.54 ±14.30	67	119	92.57 ±11.95

3.2 Data Mining and Clustering Analysis

A clustering was performed on the complete database to find distinctions between people. The algorithm, on a large scale, detected three groups. When doing the comparison with the class gender, there is an error of 35.38% in the classification of the instances, however, the clusters reveal different characteristics of people (see Figure 3).

3.3 Algorithms for Classification

The algorithms tested do not differ a change of significance of 5%. As a result, it is not possible to indicate which algorithm performs better. However, with the results provided, it is observed that the attributes are informative. This after observing that they are better than applying the baseline classifier Zero R, which, part of trivial assumptions.

Table 4. Algorithm classification.

Percent correct	Kappa statistic	Mean absolute error	RMS error	Relative absolute error	Root relative squared error	F measure	MCC	Area under ROC
0.83	0.66	0.17	0.38	0.34	0.76	0.84	0.67	0.92
0.06	0.13	0.06	0.08	0.12	0.02	0.05	0.13	0.05
J48								
0.76	0.51	0.25	0.47	0.50	0.95	0.75	0.52	0.78
0.07	0.14	0.07	0.08	0.14	0.15	0.08	0.14	0.09
Logistic								
0.81	0.63	0.21	0.40	0.42	0.79	0.83	0.63	0.87
0.09	0.17	0.06	0.09	0.13	0.17	0.08	0.17	0.07
OneR								
0.77	0.54	0.23	0.47	0.45	0.95	0.79	0.55	0.77
0.06	0.12	0.06	0.06	0.12	0.12	0.05	0.12	0.06
ZeroR								
0.53	0.00	0.50	0.50	1.00	1.00	0.69	NaN	0.50
0.01	0.00	0.00	0.00	0.00	0.00	0.01	NaN	0.00
IBk								
0.80	0.58	0.22	0.39	0.44	0.78	0.83	0.60	0.85
0.05	0.10	0.04	0.05	0.08	0.10	0.04	0.09	0.05
PART								
0.74	0.48	0.26	0.49	0.53	0.99	0.74	0.49	0.76
0.08	0.15	0.07	0.08	0.15	0.17	0.08	0.15	0.08

4 Discussion

When the Data mining tools are implemented in high performance parallel processing systems, can analyze massive data bases in just a few minutes. Faster processing means that users can automatically experiment with more models to understand complex data [7,8]. High speed is practical for the user and makes possible to analyze immense amounts of data. The great data bases, as well, can produce better predictions. The data bases can be huge as well on depth as well as on width, for example, more columns. As a result, many times, analysts must limit the number of variables to examine when the manual analysis is done due to limitations on time. However, variables that are suppressed because they seem without importance can provide information about unknown models. Bigger samples produce fewer estimation errors and deflections and allow users to make inferences about small but important population segments.

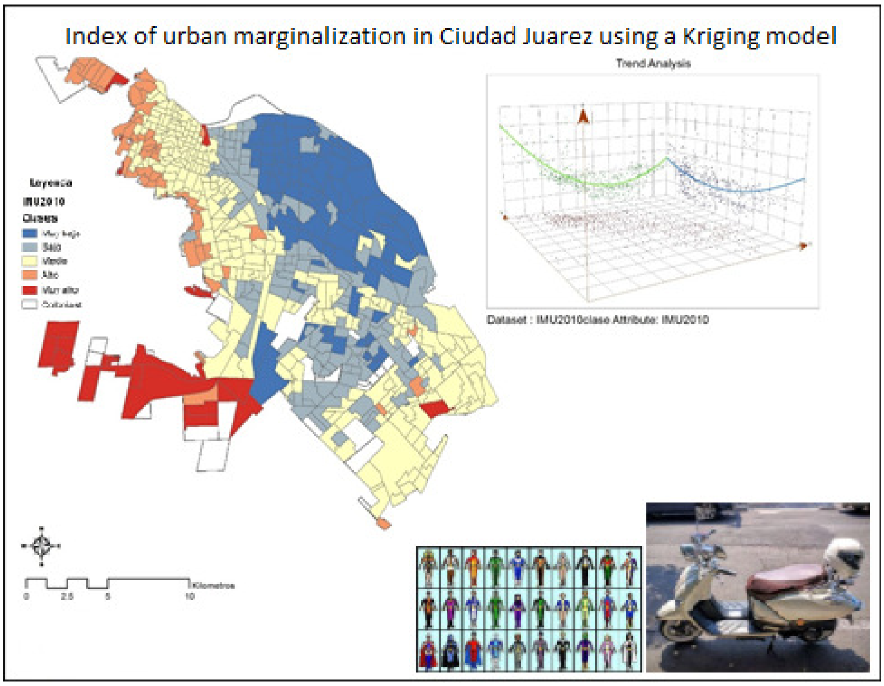

Index of urban marginalization in Ciudad Juarez using a Kriging model
Trend Analysis
Leyenda
IMU2010
Clases
Muy bajo
Bajo
Medio
Alto
Muy alto
Sin datos
0 2.5 5 10 Kilometros
Dataset : IMU2010clase Attribute: IMU2010

Fig. 4. Index of urban marginalization in Ciudad Juárez using Kriging model.

5. Conclusions and Future Research

The Data Mining tools to discover knowledge sweep the data bases and identify previously hidden models in only one step. Data mining techniques can generate benefits for the automatization of existing hardware and software platforms and can be implemented into new systems as the existing platforms get updated and new products are developed [7]. We use a Data mining tool called WEKA to analyze data.

As future research, we propose to develop a model that allows explaining the behavior showed by the environmental traffic and vial education, and how affects their activities like organize nocturnal travels together, activism, registration groups.

Another future research and application of the data gathered here is the design and the determination of a specific helmet to Go-karts.

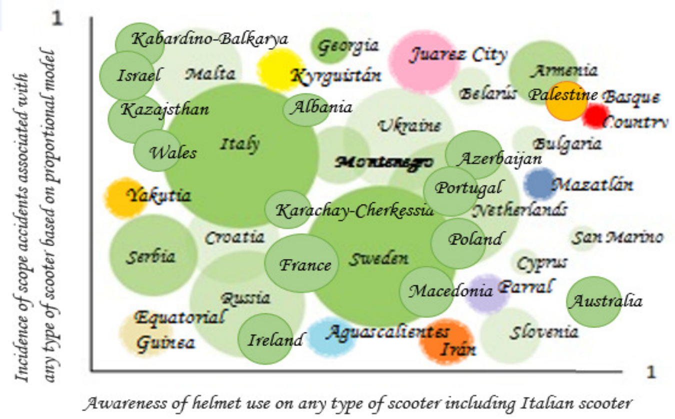


Fig. 6. Comparative of transit accidents in 39 different societies related with any kind of Motorbike including Italian Scooter: Green: European and Australian societies; Yellow and Orange: Asian societies; Brown: African Society; Blue and Purple: Mexican Societies and finally in Pink our border society of this study.

References

1. Wang, J., Thornton, J.C., Kolesnik, S., Pierson, R.N.: Anthropometry in body composition: an overview. Ann. N. Y. Acad. Sci. 904(1), 317–326 (2000)
2. Kroemer, K., Kroemer, H., Kroemer-Elbert, K.: Ergonomics: How to design for easy and efficiency. pp. 28–29 (2001)
3. Hernandez-Arellano, J.L., Talavera-Aguirre, G., Serratos-Perez, J.N., Maldonado-Macias, A.A., Garcia-Alcaraz, J.L.: Anthropometrics of University Students in Northern Mexico. Open J. Saf. Sci. Technol. 6(4), 143–155 (2016)
4. Hernandez-Arellano, J.L., Gómez-Bull, K.G.: Manual de prácticas de antropometría, biomecánica y fisiología. 1ª ed. Ciudad Juárez: Universidad Autónoma de Ciudad Juárez (2016)
5. Ward, R.E., Jamison, P.L.: Measurement precision and reliability in craniofacial anthropometry: implications and suggestions for clinical applications. J. Craniofac. Genet. Dev. Biol. 11(3), 156–164 (1991)
6. Droessler, J.: Craniometry and biological distance: biocultural continuity and change at the Late-Woodland-Mississippian interface, vol. 1. Center for Amer Archeology Press (1981)
7. Lacko, D., Huysmans, T., Vleugels, J., De Bruyne, G., Van Hulle, M.M., Sijbers, J., Verwulgen, S.: Product sizing with 3D anthropometry and k-medoids clustering. Comput. Des. 91, 60–74, (2017)
8. Probst, T., Fossati, A., Salzmann, M., Van Gool, L.: Efficient Model-free Anthropometry from Depth Data. In: 3D Vision (3DV), 2017 International Conference on, pp. 486–495 (2017)
9. Häuslschmid, R., Fritzsche, B., Butz, A.: Can a Helmet-mounted Display Make Motorcycling Safer? In: IUI, pp. 467–476 (2018)

10. Matviienko, A., Ananthanarayan, S., Borojeni, S.S., Feld, Y., Heuten, W., Boll, S.: Augmenting bicycles and helmets with multimodal warnings for children. In: MobileHCI 15:1-15:13 (2018)

ISSN 1870-4069

Development of a Graphic User Interface Focused on Multicriteria Analysis among a Plethora of Passive Exoskeletons to Improve the Social Inclusion of Infants in a Smart City

Jorge Restrepo[1,2], Juan Hernández[1] y Alberto Ochoa[1]

[1] Universidad Autónoma de Ciudad Juárez, Cd. Juárez, Chihuahua, Mexico
{al175623, luis.hernandez, alberto.ochoa}@uacj.mx
[2] Universidad Tecnológica de Pereira, Pereira, Ris, Colombia
jhrestrepoco@utp.edu.co

Abstract. Passive exoskeletons are portable devices that have beneficial functions that can address ergonomic needs taking advantage over other technological supports. They were developed to multiply human power. Passive exoskeletons have been introduced in some industrial environments, but there has been relatively little research that has examined the possible benefits, drawbacks, and tradeoffs of using the exoskeleton in a workplace. These exoskeletons require different approaches towards compliance with requirements such as use, acceptability in the workplace and possible security problems among others. Consequently, the present investigation has planned to evaluate different passive exoskeletons based on criteria such as: cost, multitasking, performance, physical demand, and influence of the mass of the tool. The criteria will be modeled and evaluated with the AHP methodology (Analytical Hierarchy Process), in order to support the acquisition decision, in the present investigation a multicriteria analysis is carried out that will allow identifying which would be the best exoskeleton of its type and how it would help children with some type of motor defect in a Smart City with regard to their social inclusion.

Keywords: passive exoskeletons, usability, multicriteria analysis, AHP.

1 Introduction

In recent years, a series of technological measures have emerged that offer help to the operator to perform certain industrial tasks of lifting, moving and unloading, such as: cranes, forklifts, and robots, but with limitations to access small places. Then, the exoskeletons appear as an alternative to counteract some of these limitations. Exoskeletons are collaborative robots that have beneficial functions that can address industrial ergonomic needs such as postural load compensation, the requirement of upper limbs or adaptability in the choice of tasks [1]. Exoskeletons are defined as portable mechanical structures that improve a person's strength, and reduce their exposure to the associated physical demand [2]. Thus, the advantages of exoskeletons over other robotic solutions are that they are intuitive to use, and are operated without requiring a robust and expensive infrastructure [3]. Then, the exoskeletons are a technological strategy that compensates the physical load without taking up space and allowing maneuverability in places where other technologies cannot operate.

pp. 103–110; rec. 2018-09-07; acc. 2018-10-07

Despite the fact that numerous work-related exoskeletons are commercially available and have been introduced into some occupational environments, there has been relatively little research that has examined the possible benefits, drawbacks, and compensations of exoskeleton use in a workplace [4]. The main benefit of the exoskeleton is a good fusion between the human flexibility and the robot improving the power, without the need of teaching or programming of robots; in addition, they can be used when others traditional solutions are not usable or effective [5]. Then, as a result of the growing tendency to use exoskeletons in the industry, where economic benefits are frequently a key to decision-making, and because they have been evaluated in terms of the reduction of physical demand, what was considered useful to evaluate its effectiveness in terms of performance and reduction of the risk of injuries [6].

An exoskeleton can be defined as an external and portable mechanical structure that improves the power of a person, and which can be classified as "active" or "passive" [7]. An active exoskeleton is composed of one or more actuators (for example, electric motors) that actively increases power to the human body, while a passive system does not use an external power source, but uses materials, springs or dampers with the ability to store the energy of human movements and release it when necessary [8]. Therefore, commercially developed exoskeletons are mainly passive in nature, with the aim of reducing the physical load during dynamic lifting and static flexion [9]. These exoskeletons, passive or active, require different approaches towards the fulfillment of requirements such as use, acceptability in the workplace and possible security problems [10]. Then, in the mechanical design of these systems, mobile ranges, safety, comfort, low inertia and adaptability should be especially considered [11]. So, usability is a key factor to take into account in the selection of technology. Usability is defined by [12], as the extent to which a system, product or service can be used by specific users to achieve specific objectives with effectiveness, efficiency and satisfaction in a specific context of use. Then, usability is as important as the design and technical characteristics with respect to high customer satisfaction and future sales [13]. So, when purchasing a team of these, it is necessary to systematically evaluate different criteria based on a methodology to support the decision. For example, [14] they applied the AHP in their study, mainly due to their inherent capacity to handle the qualitative and quantitative criteria used in equipment selection problems. Consequently, the present investigation has planned to evaluate different passive exoskeletons based on criteria such as: cost, multitasking, performance, weight, physical demand, and influence of the mass of the tool. The criteria will be modeled and evaluated with the AHP methodology (Analytical Hierarchy Process), to support the acquisition decision.

Table 1. Selected criteria.

Criterion	Sub-criterion	Unit of measurement	Specifications
Cost	Does not apply	Dollars	Does not apply
Multitasking	-With tool support -Without tool holder	- # of supports - # of supports	-Type of tool support -Type of arm support
Performance	-Productivity	-tasks / minute	-Standard time

Criterion	Sub-criterion	Unit of measurement	Specifications
	-Quality	-Milimeters	-Adjustment, tolerance allowed for the task
Weight	Does not apply	kilograms	Limit allowed
Physical Demand	-Commodity - Muscular load	-Scale Likert -Newtons	-Scale range with extremes: Discomfort and comfort -Force applied with and without tool
Mass influence tool	-With it -Without it	-Milivoltics -Milivoltics	-Maximum set for muscle -Maximum established for muscle
Social inclusion in children with motor disabilities	Does not apply	cost-benefit	Does not apply

Figure 1 depicts the parts of a passive exoskeleton with tool support [15], and Figure 2 presents the parts of a passive exoskeleton to support the shoulder [16].

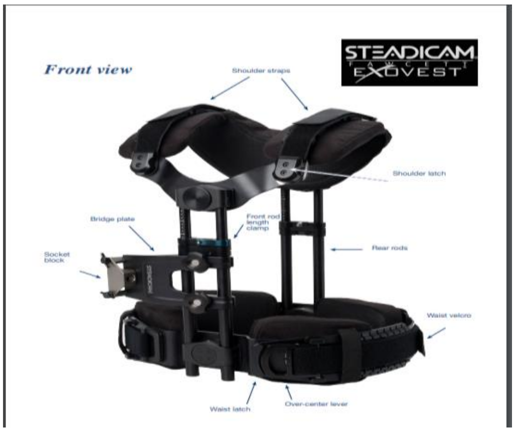


Fig. 1. Passive exoskeleton with tool support.

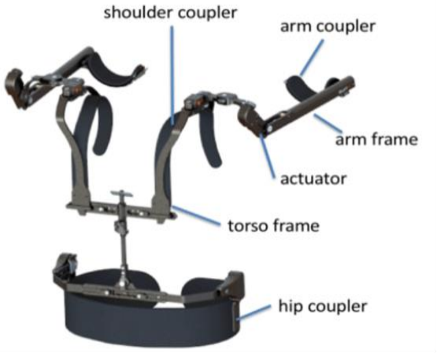


Fig. 2. Passive exoskeleton to support the shoulder.

2 Proposed Methodology

The AHP was proposed by Saaty in 1991, and it uses an objective function to add the different aspects of the problem, and its goal is to select the alternative that has the highest values of the objective function [17]. The application of the AHP procedure involves three basic steps: (1) decomposition, or hierarchical construction; (2) comparative judgments, or definition and execution of data collection to obtain pair comparison data on the elements of the hierarchical structure; synthesis of priorities, or construction of a total priority rate [18]. The weights of the criteria and sub-criteria are calculated by the AHP, where it calculates the relative importance of each criterion by means of an exploration of a multilevel hierarchy of a structure of decision making [19].

[20] presents the following decision sequence with AHP in his research:

a. Confirm the evaluation problem.
b. List the evaluation elements.
c. Establish hierarchies.
d. Establish the pair comparison matrix.
e. Obtain the eigenvector and the maximum eigenvalue of the matrix.
f. Obtain the index and the proportion of consistency.
g. Determine if the matrix of comparisons is consistent. If it is consistent, refer for evaluation. Otherwise, go back to point d.

[21] in its application of AHP for the selection of equipment suggests to keep present:

a. There is no correct hierarchy.
b. The context plays an important role for the lower hierarchical levels.
c. The factors of each level must be related to the previous level.
d. An attribute must have a name associated with its meaning and be perceived by the user as is.
e. Avoid duplication.
f. Avoid factors that generate identical alternatives.
g. A decision factor is not limited to belonging to a set.
h. The fact of moving to the comparison of pairs does not mean that the hierarchy has ended.

We propose a decision support system based on AHP, which allows analyzing diverse criteria associated with the use of passive exoskeletons, which include:

a. Cost,
b. Multitasking,
c. Performance,
d. Weight,
e. Physical demand,
f. Influence mass tool,
g. Social inclusion in children with motor disabilities.

People with disabilities also deserve to be recognized, that is why December 3 is the International Day of Persons with Disabilities. In our society we still have a hard time understanding that they are citizens with full rights and obligations, that they can perform the same jobs as the rest and that they should have the same opportunities.

The current reality is that, in general, people with disabilities have a poorer quality of life than the rest of the citizens, since their access to education is lower and, as a consequence, their labor insertion is also lower, which at the same time, it condemns them to having higher poverty rates than people without disabilities. The simple fact of being disabled carries a series of consequences that affect for life. We are in a moment in which we celebrate that a person with disabilities goes to the University and finish the race, a process that thousands of young people spend each academic year, but that seems an extraordinary fact if performed by a "disabled". All this is triggered by the fact that we still do not see it as something normal and generalized, and there is still a long way to go.

There are many who can lead a normal life, but for many people with disabilities and those who care for them, life may not be easy. Disabilities affect the whole family. Satisfying the complex needs of a person with a disability can cause a tremendous level of stress for the family, emotionally as well as economically, and often physically. It will depend on the type and degree of disability that you have. But with the policies of cuts that have been suffered in recent years it is difficult to find financial support to face the day to day of these people.

3 Design of a Graphical User Interface Associated with the Correct Selection of a Passive Exoskeleton to Improve Social Inclusion

An aspect of great relevance in our research, was not only to analyze the different types of existing exoskeletons, but to determine the functionality that each one has, as can be seen in Figure 3, and specify which could be useful to help social inclusion of childhood with motor problems.

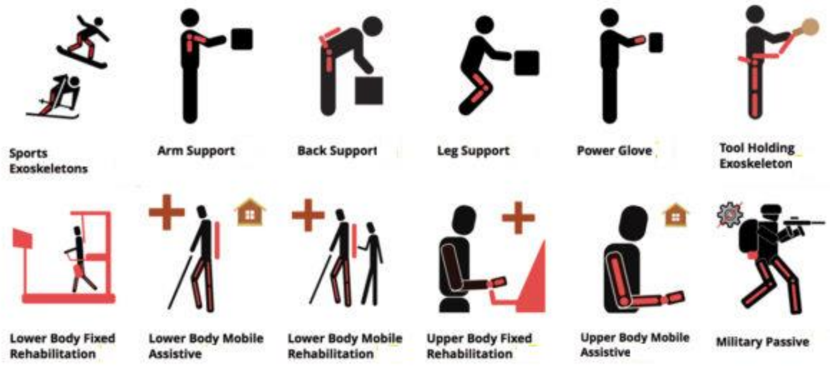

Sports Exoskeletons
Arm Support
Back Support
Leg Support
Power Glove
Tool Holding Exoskeleton
Lower Body Fixed Rehabilitation
Lower Body Mobile Assistive
Lower Body Mobile Rehabilitation
Upper Body Fixed Rehabilitation
Upper Body Mobile Assistive
Military Passive

Fig. 3. Specifications and use of different types of exoskeletons existing in the Market.

Through the implementation of an adequate and optimal graphical user interface associated with the system and composed of seven different modules can be obtained in the case of the social inclusion module, an analysis of each type of child and his motor problem, then specify the matching characteristics, then through a multicriteria analysis visualize the performance of each one of the exoskeletons that can cover the range of needs of that type of child and a visualization of how this functionary. Figure 4 shows a conceptual representation of decision making. Finally, a representation of the associated social inclusion can be observed with the performance given by exoskeleton for that particular type of motor problem, as can be seen in the ranking and matching module of our Intelligent System., as is shown in Figure 5.

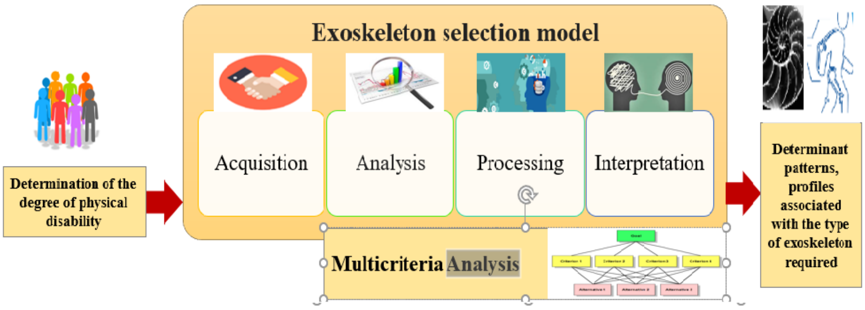
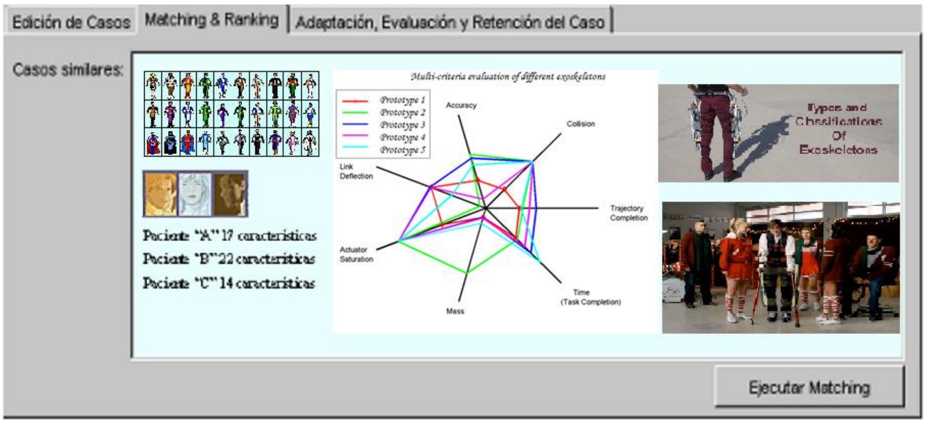

Exoskeleton selection model
Acquisition
Analysis
Processing
Interpretation
Determination of the degree of physical disability
Multicriteria Analysis
Determinant patterns, profiles associated with the type of exoskeleton required

Fig. 4. Conceptual representation of decision making.



Paciente "A" 17 características
Paciente "B"22 caracteristicas
Paciente "C" 14 caracteristicas

Fig. 5. Simulation module of our intelligent tool for the proper selection of an exoskeleton using multicriteria analysis.

Intelligent decision-making tools are based on multi-criteria analysis to be able to make an adequate selection of the best characteristics considering factors such as price, performance, usability, and above all the aspect of the human factor of their daily use.

4 Conclusions and Future Research

We have much to advance and overcome in this aspect, starting with the educational, social and labor inclusion of people with disabilities: recognizing that the difference in abilities among all people is a fact of social plurality, and not a handicap, as we have observed it until very recently. This change of mentality is what needs to be generalized in the whole of Spanish society and the world.

We cannot wait for the Administrations governments to change their procedures without first changing our way of thinking, because in the meantime we are missing the opportunity to take advantage of the enormous potential of people with disabilities.

If you want to contribute to this society with real changes in these issues, and educate from the base to eliminate or at least reduce discrimination or the non-inclusion of people who struggle every day with physical disability. The proposal of an intelligent tool that can help to select the most suitable exoskeleton depending on the type of motor disability.

References

1. Sylla, N., Bonnet, V., Colledani, F., Fraisse, P.: Ergonomic contribution of ABLE exoskeleton in automotive industry. Int. J. Ind. Ergon. 44(4), 475–481 (2014)
2. Theurel, J., Desbrosses, K., Roux, T., Savescu, A.: Physiological consequences of using an upper limb exoskeleton during manual handling tasks. Appl. Ergon. 67, pp. 211–217 (2018)
3. Carmichael, M.G., Liu, D., Waldron, K.J.: Investigation of reducing fatigue and musculoskeletal disorder with passive actuators. In: IEEE/RSJ 2010 International Conference on Intelligent Robots and Systems, IROS 2010, pp. 2481–2486 (2010)
4. Weston, E.B., Alizadeh, M., Knapik, G.G., Wang, X., Marras, W.S.: Biomechanical evaluation of exoskeleton use on loading of the lumbar spine. Appl. Ergon. 68, pp. 101–108 (2018)
5. Spada, S., Ghibaudo, L., Gilotta, S., Gastaldi, L., Cavatorta, M.P.: Investigation into the Applicability of a Passive Upper-limb Exoskeleton in Automotive Industry. Procedia Manuf. (2017)
6. Alabdulkarim S., Nussbaum, M.A.: Influences of different exoskeleton designs and tool mass on physical demands and performance in a simulated overhead drilling task. Appl. Ergon. 74, No. August 2018, pp. 55–66 (2019)
7. de Looze, M.P., Bosch, T., Krause, F., Stadler, K.S., O'Sullivan, L.W.: Exoskeletons for industrial application and their potential effects on physical work load. Ergonomics 0139, no. December, pp. 1–11 (2015)
8. Bosch, T., van Eck, J., Knitel, K., de Looze, M.: "The effects of a passive exoskeleton on muscle activity, discomfort and endurance time in forward bending work. Appl. Ergon. vol. 54, pp. 212–217 (2016)
9. Huysamen, K., Bosch, T., de Looze, M., Stadler, K.S., Graf, E., O'Sullivan, L.W.: Evaluation of a passive exoskeleton for static upper limb activities. Appl. Ergon., vol. 70, no. August 2017, pp. 148–155 (2018)
10. Spada, S., Ghibaudo, L., Gilotta, S., Gastaldi, L., Cavatorta, M.P.: Analysis of exoskeleton introduction in industrial reality: Main issues and EAWS risk assessment. Advances in Intelligent Systems and Computing 602, pp. 236–244 (2018)
11. Gopura, R.A.R.C., Kiguchi, K., Bandara, D.S.V.: A brief review on upper extremity robotic exoskeleton systems. In: 2011 6th International Conference on Industrial and Information Systems (2011)

12. Bevan, N., Carter, J., Harker, S.: ISO 9241-11 revised: What have we learnt about usability since 1998? In: Lecture Notes in Computer Science (2015)
13. Eraslan, E.: A multi-criteria usability assessment of similar types of touch screen mobile phones. J. Multi-Criteria Decis. Anal. (2013)
14. Bascetin, A.: Technical note: An application of the analytic hierarchy process in equipment selection at Orhaneli open pit coal mine. Min. Technol. (2004)
15. Tiffen: Vest: Steadicam Fawcett Exovest.
16. Van Engelhoven, L., Poon, N., Kazerooni, H., Barr, A., Rempel, D.: Evaluation of an adjustable support shoulder exoskeleton on static and dynamic overhead tasks. pp. 1–5 (2018)
17. Rashidi, M., Ghodrat, M., Samali, B., Kendall, B., Zhang, C.: Remedial Modelling of Steel Bridges through Application of Analytical Hierarchy Process (AHP). Appl. Sci. (2017)
18. Chen, C.-F.: Applying the Analytical Hierarchy Process (AHP) Approach to Convention Site Selection. J. Travel Res. (2006)
19. Ortiz-Barrios, M.A. et al.: The analytic decision-making preference model to evaluate the disaster readiness in emergency departments: The A.D.T. model. J. Multi-Criteria Decis. Anal. (2017)
20. Lin, Z.C., Yang, C.B.: Evaluation of machine selection by the AHP method. J. Mater. Process. Technol. (1996)
21. Shapira, A., Goldenberg, M.: AHP-Based Equipment Selection Model for Construction Projects. J. Constr. Eng. Manag. (2005)

ISSN 1870-4069

Development of a Serious Games for Asperger Syndrome Based on a Bio-inspired Algorithm to Measure Empathy Performance

Alejandro Lara, Jorge Rodas-Osollo, Alberto Ochoa, Gilberto Rivera

Universidad Autónoma de Ciudad Juárez, Ciudad Juárez, Chihuahua, Mexico

Abstract. Currently, the development of a Serious Game, whose purpose is to assist cognitive behavioral therapy for patients with Asperger Syndrome, should provide valuable information always to the specialist providing the psychological therapy, to determine the degree of progress that the patient is achieving, and related to facial recognition and social disabilities exercises. Thus, we can improve the symptoms and achieve a better social interaction with other people. The use of specific knowledge from specialists about empathy can make simple videogame in a serious game, in which not only the fun factor is involved, this knowledge gives a support for a satisfactory progress achieved by the individual while playing. This paper, presents the progress in the development of a serious game that helps the socialization of children with Asperger syndrome, general aspects, and methodology that is being used.

Keywords: serious game, psychological therapy, algorithms, artificial intelligence, Asperger syndrome.

1 Introduction

The game is a fundamental activity for human development. Usually played to have fun, to entertain. However, there are those who claim that it is mainly played to learn [4]. Health video games are designed to entertain while attempting to modify some aspect of behavior [1]. With this premise is how you can summarize the main function of any therapeutic video game that is focused on health, and more in this case, in which we focus on people suffering from Asperger Syndrome.

In the field of research on autism spectrum disorders, computer-based interventions are being used to improve emotion and face recognition skills, as well as language and social skills [2], which have achieved be successful as the patient continues treatment. This approach has been inspired, in part, by the findings that children with this type of disorder often enjoy playing computer games in their free time [2].

Knowledge elicitation tactics has played a key role in the construction of serious games. The idea of constructing this Serious Game is to use specialized knowledge to support the progress that the patients with Asperger has, and that information obtained from mental health specialist can use the treatment of the patients. The game is a fundamental means for the structuring of language and thought, it acts systematically on the psychosomatic balance; enables learning of strong significance; reduces the feeling of gravity against errors and failures; invites participation by the player; develops creativity, intellectual competence, emotional strength and personal stability. Finally, it can be affirmed that play constitutes a fundamental strategy to stimulate the integral development of the people in general [4]. The aim of this article is to show the progress of the project of developing a serious game to help the socialization of children with Asperger syndrome. In section 2, a brief explanation of flow theory, since has a significant impact when designing a video game. In section 3, the project structure is presented three mains items (psychological, technical and narrative of the video game). In section 4, is the proposed game with the narrative and how the therapy is including in the videogame. Section 5 presents the procedure of the project with the experimental activities, the location and time to test the serious game. In section 6, the statistical analysis shows the analysis of the existence of statistically significant differences in the pre-test and post-tests measurements of two groups.

2 Related Research

A theme related with videogame design is flow theory, because establish the fun of videogame.

2.1 Flow Theory

The flow theory also known as the zone is the operative state of mind in which a person is completely immersed in the activity executes. It is characterized by a feeling of focusing energy, of total involvement with the task, and of success in carrying out the activity. This feeling is experienced while the activity is in progress. The concept of flow was proposed by the psychologist Mihaly Csikszentmihalyi in 1975 [10]. This theory has been put into practice in different areas always with satisfactory results. In games, this practice results in having the player perform the activities in a way that is fun, and this is achieved by having a balance between the skills and the challenges that we face the player, we will achieve that the game is entertaining and motivating.

According to Csikszentmihalyi, the components of a flow experience are as follows [10]:

Table 1. The Components of a Flow Experience.

1	Clear objectives (Expectations and norms can be perceived, and goals are attainable appropriately with the set of skills and abilities).
2	Concentration and focus, a high degree of concentration in a limited field of attention (a person related to a single activity will have the opportunity to focus and delve into the matter).
3	Direct and immediate feedback (successes and failures in the course of the activity are obvious, so the behavior can be adjusted as needed).
4	Balance between skill level and challenge (The activity is neither too easy nor too complicated).
5	The activity is intrinsically rewarding, so no effort is noticeable when running.

3 Theoretical Framework

In every project work the area about theoretical framework is very essential. In this section we going to talk the structure of the project.

3.1 Structure

In the structure section we divide the project in two areas (psychological category and technical category).

Psychological Category

According to the DSM 5 Manual of the American Psychiatric Association (a reference manual diagnosing neurological problems) [15], Asperger syndrome is already within autism spectrum disorder, with Asperger being the lowest level within of said disorder. As discussed earlier, this article is about the progress that is developing a serious game called also serious video game, which will help the socialization of children who have the diagnosis of Asperger Syndrome. To achieve such socialization, it is important to see empathy as one of the aspects to be addressed to achieve the objective. For the video game to have the therapeutic character and can be used in the therapies that the mental health specialists carry out, it is necessary that it fulfills a series of stages that in a conventional therapy would be carried out, the stages in question They are:

Facial Recognition. The deficit would be manifested through a marked difficulty in understanding the emotional load that faces present, being the inability to interpret emotions through the face. Abnormal processing of faces in people with autistic disorder may be because faces are not a stimulus important to them or in any case social content, not processed in a usual manner [12].

Mind Theory. Deficit Mind Theory mentions that people with Asperger show serious difficulties to take the place of the other and intuit their mental world. People with Asperger 's are unable to intuit the mental world of others, so that the consequences of this inability or difficulty very serious [12].

Some of the consequences of this limitation for understanding the mental world of others would be the following (Table 2) [12]:

Table 2: The limitations for understanding the mental world.

Difficulty predicting the behavior of others.
Difficulty in realizing the intentions of others and knowing the true reasons that guide their behavior
Difficulty understanding emotions, both own and others, which leads them to show few empathic reactions
Difficulty to consider the degree of interest of the interlocutor on the topic of conversation
Difficulty anticipating what others may think about their behavior
Difficulty to lie and to understand deceit

Subtleties, Skills and Social Interaction. In basic interaction skills, children with autistic disorder present difficulties since basic social conventions such as greeting, if they are not known, do not do it, as well as thank or ask permission, do not decode them because they are not explicit, many times they do not greet by the amount and forms of greetings that appear, that fail to capture the key and to be so changing it generates in the boys' and girls' confusion and finally they do not integrate it in their repertoire. By performing these stages within the video game, we can say that the patient is conducting a therapy as it would in the specialist's office, of course, adding the playful part that any video game has.

3.2 Knowledge Elicitation Tactic

The methodology that we use in this project is KMoS-RE. The KMoS-RE system is a tool that generate explicit knowledge from tacit knowledge and is based on the premise that, to develop a software, you must understand the requirements, and for that it is necessary to understand the domain [16]. The strategy consists of 3 phases:

- Domain Modeling: Formalizes the properties of the domain concepts, making explicit the concepts, attributes, relationships between concepts and basic integrity constraints. A lexicon is used to identify, classify and define domain terms [16].
- System Modeling: The information used to develop this model is derived from the lexicon and the conceptual model, and in this way the requirements are formalized [16].
- Development Specification: In this stage the requirements obtained are used to generate a formal document with the necessary information about the project in question [16].

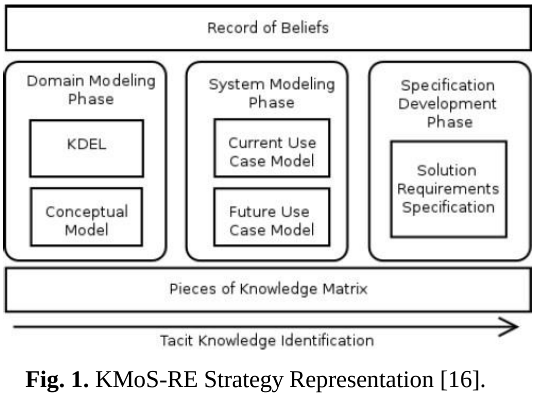

Record of Beliefs
Domain Modeling Phase
KDEL
Conceptual Model
System Modeling Phase
Current Use Case Model
Future Use Case Model
Specification Development Phase
Solution Requirements Specification
Pieces of Knowledge Matrix
Tacit Knowledge Identification

Fig. 1. KMoS-RE Strategy Representation [16].

Game Engine

The Game Engine refers to a series of routines that allow the execution of all elements of the game. It is where the control of each element of the game is represented and how it interacts with them. It is here that the AI (Artificial Intelligence), behavior, personality and ability of the elements of the game are combined, the sounds associated with each element of the game in each moment and all the graphic aspects associated with them, including the kinematics of the game [11].

The game engine selected for the creation of the video game is Unity which is the tool of creation of videogames more used today. It has an excellent editor, with a multitude of options, is very friendly, and is completely multiplatform (PC, MAC, Linux, Web browsers, iOS, Android, Blackberry, Windows Phone, Xbox 360, PlayStation 3, Wii, Xbox ONE, PlayStation 4, PlayStation Vita and WiiU), allowing to export a project for any of these platforms in a straightforward way [9]. This video game aims to follow the theory of flow, but applied to people suffering from Asperger syndrome, therefore, is adjusted for these conditions of use.

4 Proposed Game

In this section the narrative describes the therapy work together with the videogame to make the serious game.

4.1 Narrative Category

As mentioned in the section of psychology, video game should contain the issues in question to be considered therapeutic, but without removing the playful game contains everything, and is represented as follows. The reward is a key factor in the game, because we keep the attention and motivation of the child and on the other hand will tell the progress that has been obtained. The reward system consists of stars that the child is gaining as he achieves the objectives and is advancing.

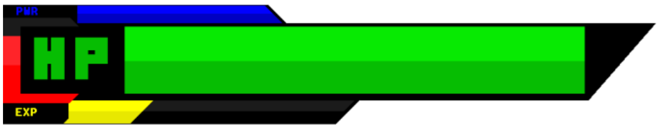


Fig.2 Lives representation.

First Part

In this section, and a scenario in 2D in which the game begins where a child with Asperger goes to sleep and dreams of an adventure that happens, which is immersed in a series of situations that must be resolved is contemplated being these situations the issues raised in the psychology category. First, there is the stage of face recognition, at this stage the child is in a village which is walled and has to leave but to get must be able to open the door of the wall with two keys which have them two people (Avatars) that have the face with the faction that we want the child to recognize in that case (happy, angry, sad, indifferent, etc.) and in case you ask the key to a different avatar will respond that he does not have The key that continues searching, it is necessary to remember that the main idea is that the boy recognizes the facial emotions. People with Asperger's Syndrome do not recognize facial expressions as other people would, as a defect they must recognize emotions, and this is reflected in the social behavior of the person.

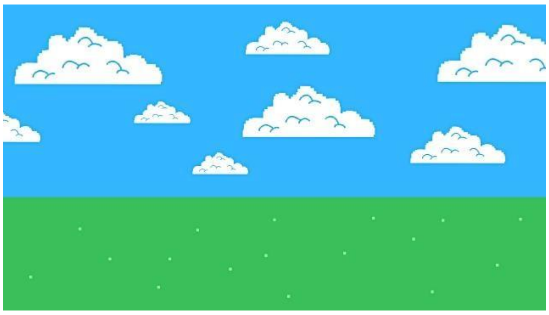


Fig. 3 Dream and adventure begins.

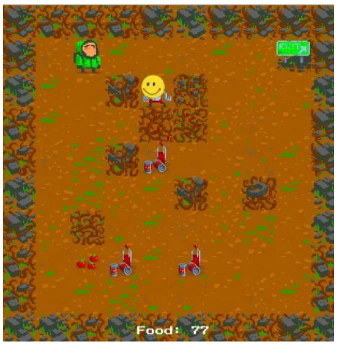


Fig.4 First level facial recognition

Second Part

Once the patient manages to open the gate of the wall he will have passed the first level of the game, later he will go through some paths and some bridges as part of his adventure that will take him to the next stage. In this second part, the following item in question is theory of mind, which is taken as reference one of the most common tests in therapy, which is the Test of Anne and Sally, will become clear is set for release of video game.

Upon reaching the next stage, the child will find two characters in a forest where there is a hut where they live and observes how one of them leaves an object in a place near the hut and gets in, and the other Character takes the object and hides it in the forest. On leaving the first character of the hut there that help you find the object and should start where I left last time.

For a person who has Asperger disorder will result by other obviously should look at where you left off last time, but if you have Asperger someone who does not understand that there should look first. To find the object, began dating all the lost items that had hidden him the second character from long ago, and as a reward give you a canoe that had to cross a river to continue looking for the lost treasure. The concept of mind theory refers to the ability to understand and predict the behavior of other people, their knowledge, their intentions and their beliefs [6].

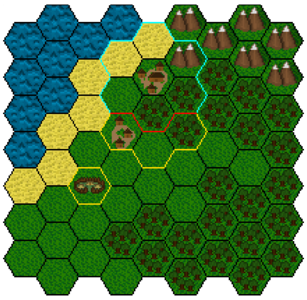


Fig.5 Second level theory of mind.

Third Part

It is clear to determine that to establish an improvement in therapy it is convenient to see the patient have a better social interaction with the people in their environment, so we can say that the therapy is effectively giving the expected result. As part of the last level of the game, there are a series of situations in which the patient is interacting with other avatars simulating everyday situations of life. In this part, social subtleties, social skills and social interaction will be seen together as these 3 areas have much in common.

Fig. 6 Village Representation.

Once the child has sailed the river reaches a village where is the lost treasure, but realizes that to get to the treasure, he must be kind, respectful and polite to the other characters there they are, for example, he should arrive greeting the characters to be told where the treasure is, and tell him across the bridge can get but to cross it must ask the guard to allow cross asking for a polite way. Reaches across a pyramid and tell him that there It is the treasure, but to open the pyramid deb and to have friends to help you and begins to befriend the characters of the village and together will open the pyramid and find the treasure.

At this point, the child wakes up at home and tells his mother that he dreamed of adventure.

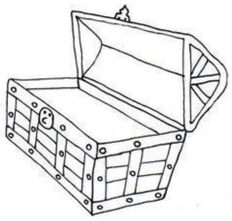


Fig 7. Treasure representation.

5 Procedure

The experimental activities were carried out at the Medical Center of Americas at El Paso, Texas. The study was carried out over the course of 3 months in sessions held once a week in which two groups of patients participated every hour. This study was carried out in accordance with the recommendations of American Psychology Association. The entire experiment was conducted in accordance with the Declaration of Helsinki. All participants gave also oral informed consent. Ethics approval was obtained from the Research Ethics Committee of the Health Research Unit. Tests were conducted in a sound-proofed room equipped with tables, chairs, and a computer connected to a projector. The groups consisted of between 10 and 15 patients, and the experiment lasted approximately 1 h. Participation was voluntarily. If any patient did not wish to take part, he or she could stop the activity at any time. The only criterion for exclusion was if the patient did not understand Spanish. The patients and were in- formed of these details prior to starting the experiment. None of the patients were excluded because of the language criterion, nor did anyone choose to leave during the sessions. The patients answered the Asperger's Standard Questionnaire [20] on patient's attitudes toward syndrome both before contact with serious game and after. In addition, once the test had ended, they were also asked to evaluate the usefulness and visual appeal of the program (on a scale from 0 to 10), whether they would recommend playing the game to a relative or friend or not, what they thought the game had taught them, and what improvements they would make on the game. In the control group, the patients participated in another video game not related to Asperger's with the same duration as Serious game. This group also answered the abovementioned questionnaire both before and after playing.

5.1 Statistical Analysis

To analyze the existence of statistically significant differences in the pre-test and post-tests measurements of the two groups, the patients was utilized for independent samples. This was complemented by an effect size with a statistic that corresponded, which in this case was Cohen's d. In a second analysis, the posttest and pre-test measurements of each group were compared using the Student's t for related samples. The third analysis involved the use of Cohen's d to assess the magnitude of the change produced during the experience. Finally, the descriptive statistics were utilized which were gathered from the participants 'evaluations and answers. The analysis was carried out using the statistics program SPSS 11. As can be observed in Table 3, the average difference test between the pre-test measurements of the experimental group and the control group did not reveal the existence of statistically significant differences between the two with the variables analyzed. However, there were statistically significant differences between the two groups for all the variables evaluated following the intervention.

By using Cohen's d, it was con- firmed that the differences between the groups after the activity were moderate. The means and standard deviations of the variables in the study that correspond to the experimental and control groups for each study phase are displayed in Table 4. The analysis of the post-test–pre-test scores of the control group revealed no statistically significant differences with respect to any of the variables evaluated, as can be observed in Table 4.

However, significant differences were found in the same analysis for the scores of the experimental group, both in the total of the questionnaire score in relation to serious game and its two factors, namely, facial recognition and social dis- abilities. Consequently, it can be seen how the Asperger's problems decreased among the people who participated in the serious game program.

About the scope of the effect, it was observed that the program had a strong impact on reducing Asperger problems in facial recognition but was weak in terms of affecting other social disabilities.

Regarding the assessment carried out by the participants, the program was given a high score for usefulness (7.2 average) and a slightly lower average score for interest (6.1). 75% of the participants said they would recommend playing serious game to a relative or friend. With regards to the characteristics of the different stories in the game, it can be observed that participants easily identified that said stories were not emotionally well.

Table 3 Average difference test between the pre-test measurements of the experimental group and the control group.

	Pre-test			Post-test		
	t	p	d	t	p	d
Facial Recognition	0.833	0.405	0.127	-3.477	0.001	-0.481
Social Disabilities	-1.146	0.252	-0.280	-3.815	0.000	-0.533
Total	-1.069	0.286	-0.165	-3.894	0.000	-0.597

Table 4 Means and standard deviations of the variables in the study that correspond to the experimental and control groups for each study phase.

	Experimental					Control				
	Pre-test	Post-test	Pre-post			Pre-test	Post-test	Pre-post		
	M (SD)	M (SD)	t	p	d	M (SD)	M (SD)	t	p	d
Facial Recognition	5.11 (2.13)	3.55 (2.56)	16.899	0.000	0.662	4.86 (1.79)	4.66 (2.02)	1.244	0.220	0.104
Social Disabilities	4.41 (3.29)	3.96 (3.33)	3.301	0.001	0.136	5.48 (4.27)	6.00 (4.27)	-1.748	0.088	-0.121
Total	9.54 (4.78)	7.54 (5.24)	10.406	0.000	0.399	10.35 (5.03)	10.64(5.15)	-0.768	0.447	-0.056

6 Discussion

The results of the application of serious game with Asperger patients are presented. Firstly, it is worth highlighting that with regard to Asperger, evaluated using the Questionnaire [20] on patient's attitudes toward Asperger's, there was a considerable decrease among the participants who used the serious game. This was not the case of the participants who used other video games not related to Asperger's. As a result, the game's effectiveness in reducing misconceptions is clearly observed, particularly in relation to a friendly therapy. It must be taken into account that the preconceived notion of violence or danger associated with people with severe mental. The general assessment of serious game by the patients revealed interesting results as well. The participants scored close to eight points (7.2) for the game's usefulness and a slightly lower score for interest (average score of 6.1). Some of the patients' comments made about the game were the following: "everyday examples help to under- stand this disorder a little better," "it is fundamental to help and accept any individual, and to be patient," "people with Asperger as not necessarily an alien", and "I have learnt that I should treat people with same problems naturally." Additionally, 75% stated they would recommend trying this game to a relative or a friend. In light of the results obtained, serious game has proved to be an effective tool for raising awareness among people about Asperger's and for providing information about these disorder, which makes it possible to dispel myths.

References

1. Thompson, D., Baranowsky, T., Buday, R., Baranowsky, J., Thomson, V, Jago Russell, Griffith, M.J.: Serious Video Games for Health: How Behavioral Science Guided the Development of a Serious Video Game (2010)
2. Elisabeth, M., White, J., Smyth, M., Scherf, S.K.: Serious Game Designing Interventions for Individuals with Autism (2014)
3. Murillo, J.B., Collado, L.J., Garcia Garcia-Magariño, I.: Implementation Strategy Games with Evolutiva Programming (2005)
4. Marcano, B.: Serious Games and Training. Theory Digital Electronics Education Society. Education and Culture in the Information Society 9(3) (2008)
5. Author Pictograms: Sergio Palao Hometown: ARASAAC (http://arasaac.org) License: CC (BY-NC-SA) Property: Government of Aragón
6. JTirapu-Ustárroz, J., Pérez-Sayes, G., Erekatxo-Bilbao, M., Pelegrín-Valero, C.: What is the theory of mind? Rev Neurol 44, 479–89 (2007)
7. Vazquez-Nuñez, A.E., Fernandez-Leiva, A.J.: Computer Ephemeral: identifying research challenges in video games (2016)
8. Turiégano, C., López, J.: Evolutionary design new techniques for fighting games (2015)
9. Boluda Cuesta, D., Cuesta Boluda, G., Rodriguez Osorio Jimenez J.: Producing a multiplayer video game genres Unity Combining MOBA and RTS (2014)
10. Wikipedia: Flow (psychology). Retrieved on August 31 (2016)
11. González Sánchez, J.L., Padilla Zea, N., Gutierrez, F.L., Cabrera, M.J.: Usability gameplay: Game Design Centered Player (2012)
12. Unzueta Arce, J., García García, R.: Facial Processing Deficit Disorders Autism spectrum (2002)

13. Mariscal Márquez, C.: Asperger Syndrome (2009)
14. Pérez Aguilera, M.D.C.: Comparative study of social skills of children with Asperger syndrome with children from ages typically developing 6 to 8 years in primary education integrated (2013)
15. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders. Fifth Edition (2013)
16. Olmos-Sánchez, K., Rodas-Osollo, J., Fernández-Martínez, L., Morales-Rocha, V.: Requirements engineering based on knowledge: a comparative case study of the KMoS-RE strategy and the DMS process (2015)
17. Fleming, T.M., Bavin, L., Stasiak, K., Hermansson-Webb, E., Merry, S.N., Cheek, C., Lucassen, M., Lau, H.M., Pollmuller, B., Hetrick, S.: Serious Gamesand Gamification for Mental Health: Current Status and Promising Directions. Front. Psychiatry 7, p. 215. doi: 10.3389/fpsyt.2016.00215 (2017)
18. Vallejo, V., Wyss, P., Rampa, L., Mitache, A.V., Müri, R.M., Mosimann, U.P., et al.: Evaluation of a novel Serious Game based assessment tool for patients with Alzheimer's disease. PLoS ONE 12(5): e0175999 (2017)
19. Eichenberg, C., Grabmayer, G., Green, N.: Acceptance of Serious Games in Psychotherapy: An Inquiry into the Stance of Therapists and Patients (2016)
20. Scott, F.J., Baron-Cohen, S., Bolton, P., Brayne, C.: The CAST (Childhood Asperger Syndrome Test): Preliminary Development of a UK Screen for Mainstream Primary-School-Age Children. Autism 6(1), 9 (2002)

ISSN 1870-4069

Forward Kinematics for 2 DOF Planar Robot using Linear Genetic Programming

Humberto Velasco Arellano, Martín Montes Rivera

Universidad Politécnica de Aguascalientes, Maestría en Ciencias en Ingeniería, Mexico
mc160006@alumnos.upa.edu.mx, martin.montes@upa.edu.mx

Abstract. In the field of robotics, forward kinematics is an activity that allows finding a mathematical model for the resulting position in the final effector based on the robot joints position, a popular alternative for determining this model is defined by the Denavit Hartenberg convention, nevertheless, this method requires knowledge about linear algebra and three-dimensional spatial kinematics. Machine learning uses specific computational methodologies to solving similar problems in several areas, so it could be a viable answer for automatic determining of forwarding kinematics. In this work we propose the use of genetic programming as a machine learning algorithm for finding the forward kinematics of a 2 degrees of freedom robot, getting a satisfactory outcome obtaining a satisfactory result with blocks that describe the expected solution, validating the capacity of the genetic programming in order to validate this algorithm for later work with more complex robots.

Keywords: forward kinematics, automatic robot modeling, linear genetic programming.

1 Paper Structure

In the introduction section, a general description is presented where the problem is discussed, the proposed solution, and the scope sought.

The related work presents all the previous investigation that supports the problem as an open issue, the applications that validate the genetic programming as a solution and its characteristics.

In the theoretical framework, the emphasis is placed on the problem, its origin, and the conventional solution and then is described all the related to the proposed solution.

The methodology presents a detailed description of the problem, as well as the solution process used, its behavior and the tests carried out.

The results section shows the data obtained from the algorithm as well as its performance, the validation applied to the results and the discussion about the impact of them with respect to those found in the literature.

Finally, the conclusions describe the contributions, the scope of the results and the future work arising from it.

pp. 123–133; rec. 2018-09-08; acc. 2018-10-07

2 Introduction

Robotics remains to be a growing research area, which has become relevant since the presentation of the first industrial robot in 1954 [1].

One of the problems present in robotics is the forward kinematics (FK onwards) model, which relates the positions of the joints to the Cartesian positions of the final effector, for which there are several deterministic methods of solution, but the problem requires previous knowledge in different mathematical areas, like linear algebra and vector spaces [2].

Machine learning allows the solution of complex problems without the intervention of human beings that is why represents an alternative for solving the FK problem. Evolutionary algorithms (EAs onwards) are machine learning optimizing algorithms that can be classified in single or multi-objective algorithms, in general, EAs algorithms follow specific operators which include, randomly initialized population, evaluation, fitness assignment, selection and reproduction (Fig**.** 1) [3,4].

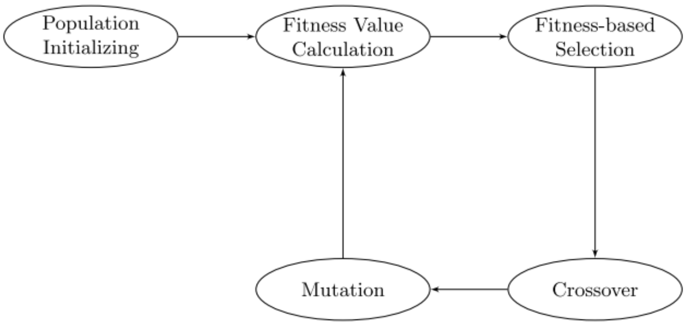


Fig. 1. EA main operators algorithm taken from [3].

EAs have presented important advances in various branches of science [4–8], where they are presented as an alternative method to problems in which the deterministic solution is not successful[9].

The most used EAs are Genetic algorithms (GAs onwards) and Genetic Programming (GP onwards), both based on natural selection principles proposed by Charles Drawing, but applied in different situations. GAs perform numerical optimization on known structures or programs [10], on the other hand, GP performs structural optimization on unknown structures or programs allowing to determine mathematical models, like those required in FK. [3,11].

EAs have been used for various problems in robotics[12], such as tracking people [13], making decisions robots according to their urgency [14], among others, but genetic algorithms are restricted to numerical optimization for known structures [15], Therefore FK cannot be applied directly with GAs, because whit that solution the transformation matrix must be determined for each given position which would require high computational power. Nevertheless, if mathematical equations in those matrixes are determined instead of those numbers then the solution model would be found with GP as an optimizing algorithm but only requiring one initial run [16].

2.1 Related Works

In recent years FK problem has been studied by different methodologies, some of which use geometric analysis to find a simple solution applied to parallel robots [17], and with this technique, they look for solutions for anthropomorphic mechatronic systems where are used 4x4 matrixes that work as operators for solving FK problem [18].

Other investigations show models that are based on the length of the links for transmitting the movement in a hybrid robot [19] and using this research in [20] models are explored based on flexible robots which are not considered in the aforementioned convention.

Another tool used is the quaternions, which use an extension of the real numbers and for this case allow to describe the movement of the coordinates along a kinematic chain [21].

On the other hand, there is the use of computational tools, which have been explored less frequently, within the highlighted works an hybrid algorithm was used between the search with particle swarming and a differential evolution applied to solve FK problem with remote manipulators [22].

Likewise, the use of neural networks has been recurrent for autonomous FK solution, like when applying complex networks to solve FK with parallel robots in [23], and redundant robots in [24].

In the case of GAs, they have been applied with hybrid processes that take advantage of other algorithms. GAs have been tested working in conjunction with simulated annealing in [25] and have been used with geometric similarities for optimizing its numerical parameters [26].

In this work, it is proposed to find the Denavit Hartenberg (DH onwards) parameters with mathematical expressions obtained with a GP algorithm in a robot with 2 degrees of freedom.

3 Theoretical Framework

Kinematics is the branch of physics that analyzes movements without considering the forces that cause them [27]; applied to a robot, the kinematic models describe the relationship between each of the robot joints and its final actuator position [2], where it is possible to determine FK forward or backward depending from if the final actuator position is found or the joints position respectively.

3.1 Forward Kinematics

Forward kinematics is the mathematical model that relates the known positions of each of the joints and their relationship with the Cartesian axes [2].

The DH convention describes the behavior of the l_i link with respect to the l_{i-1} link, with 4 important transformations described with the below parameters [28].

- Θ_i, angle of the joint, with which the orthogonal plane is projected with respect to the normal anterior plane.
- d_i, displacement of the joint, length between the links with respect to their joints.

- a_i, length of the link between the common perpendiculars.
- α_i, torsion angle between the orthogonal projections of the Z axis in a perpendicular plane.

The homogeneous matrix that represents those 4 transformations previous parameters is shown in equation (1):

$$A_1^0 = \begin{bmatrix} \cos(\theta) & -sen(\theta) & sen(\theta) & a * \cos(\theta) \\ sen(\theta) & \cos(\theta) & -\cos(\theta) & a * sen(\theta) \\ 0 & sen(\alpha) & \cos(\alpha) & d \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{1}$$

Following the DH rules are possible to find a transformation matrix that contains the equations that describe the position and orientation of the robots in its final effector with respect to the values given by its joints [1].

3.2 Linear Genetic Programming

This algorithm is considered within the meta-heuristics, since it is an EA that performs a soft search to determine the structure that contains a possible solution [29].

The form of work proposed is that described in Algorithm 1, which specifies the evolutionary cycle that allows a population of possible responses to interact to generate an increasingly better-adapted offspring to the given problem.

```
GP Program (Best Individual)
  Initial values;
  Begin
    Start_population();
    Evaluate_aptitude();
    Repeat
      Select();
      Cross();
      Evaluate_aptitude();
      Mutation();
      Evaluate_aptitude();
    Until aptitude = Expected OR End_generations
End.
```

Algorithm 1. Fitness evaluation algorithm.

Population randomly initialized generates possible individuals, which are evaluated in the problem to be solved, once the cycle has begun, the individuals who may have children in each generation are selected, the children are created from the parents and in some cases have mutations, finally the cycle is evaluated and repeated until the desired fitness value is reached, or the desired generations are achieved [30].

4 Methodology

This section explains how the GP tool was built and the problem of finding DH parameters for a robot automatically and how the proximity to the expected result was evaluated.

4.1 Design of Experiment

For this work, a 2-degree freedom planar robot with two rotational joints is used, its schematic is shown in **Fig. 2**.

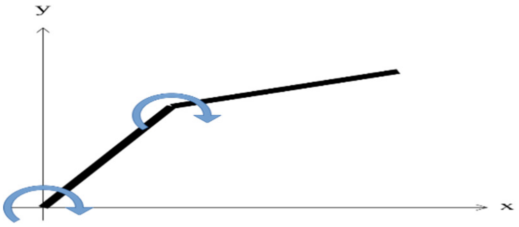


Fig. 2. Planar robot diagram.

It was decided to make use of this robot for its ease and because the search method used has no precedent, it was initialized with a known robot, which, by modifying any of its joints, moves its final effector along the x and y axes.

For simulation MATLAB 2015b was used since the algorithm was worked as a test, for the moment no special command was used, everything was worked with structured code.

4.2 Objective Function

In order to evaluate the suitability of the individuals in FK, the results of the GP were taken and placed in a 2x4 matrix and used to assemble the transformation function according to the DH rules, taking the first column in the matrix as the joint angle, the second as the joint displacement, the third as the link size and the last of them as the angle between the z-axis projection and the next plane.

These values were evaluated in a robot of 300 and 400 cm per link in the position of $\frac{\pi}{3}$ and $\frac{\pi}{4}$ in each joint respectively, the positions selected here considered that the size of the operation could not exchange its values, in addition to working within the first quadrant of work, thus accumulating the absolute value of the sub-traction between the known solution of this robot following the DH convention and those obtained by the GP as is shown at equation (2).

$$\mathrm{Error} = \mathrm{abs}(X_{DH} - X_{GP}) + \mathrm{abs}(Y_{DH} - Y_{GP}) + \mathrm{abs}(Z_{DH} - Z_{GP}). \tag{2}$$

4.3 GP Algorithm

This algorithm was executed with the initial conditions described in Table 1, which were selected according to the suggested rules in the literature, where decisions are made for increasing or decreasing each term based on the results obtained and its behavior [30].

Table 1. Initial values of GP.

Initial condition	Value
Seed	1
Population size	200 individuals
Genes number	8
Alleles number	4
Tournament size	3
Generations	1500
Mutation percentage	10%
Mutation numbers	1800
Mutations per generation	12

After this a basic GP code was assembled following algorithm 1, where individuals with 8 blocks were used each block describing a DH parameter in the table, those blocks are functions of q_i and l_i having the structure shown in equation (2):

$$B_i = f(q_i, l_i)\,. \tag{3}$$

These blocks are assembled into a 2 x 4 matrix and subtracted against the previously found DH parameter matrix.

5 Results

First of all, the DH parameters were obtained following the rules that this convention requires, in this way the results shown by the algorithm will be confirmed. In this way, the parameters of Table 2 were obtained.

Table 2. Planar robot DH parameters.

Transformation	θ	d	a	α
${}^{0}A_1$	q_1	0	l_1	0
${}^{1}A_2$	q_2	0	l_2	0

The parameters obtained were simulated to visualize the behavior of the robot, and to obtain the graph shown in Fig. 3. Where an angle was used in the first joint of $\frac{\pi}{3}$ and in the second joint with $\frac{\pi}{4}$.

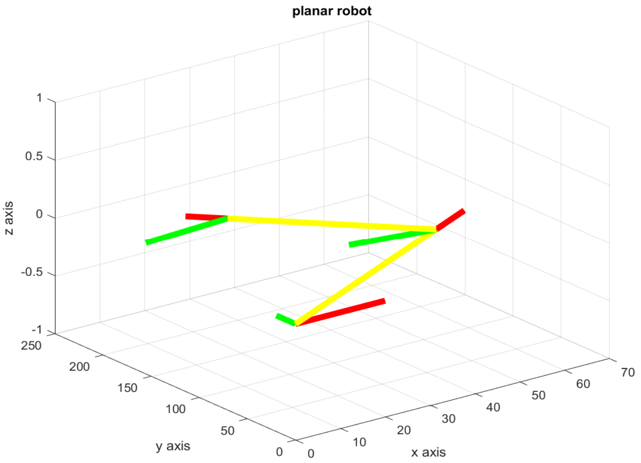


Fig. 3. Planar robot graph.

The next step was to take the results of the GP, which formed the matrix shown in Table 3. These have a larger number of parameters, as this algorithm required more space to increase the search diversity.

Table 3. Best element obtained by the GP.

Transformation	Θ	d	a	α
${}^{0}A_{1}$	$q_1 + \frac{2}{l_2}$	0	$l_1 - 1$	$1 - \frac{3}{l_2}$
${}^{1}A_{2}$	q_2	$\frac{2}{l_2}$	l_2	0

All the solutions as expected have good fitness i.e. equations satisfy FK, but the representation shown by the result requires to be cleaned or that unused extra data be simplified, as was presented at the literature, all the results take the structure to where the correct data is [29].

The graph shown in Fig. 4 shows the position of the robot at the angles mentioned above, where the similarity between the positions obtained can be seen.

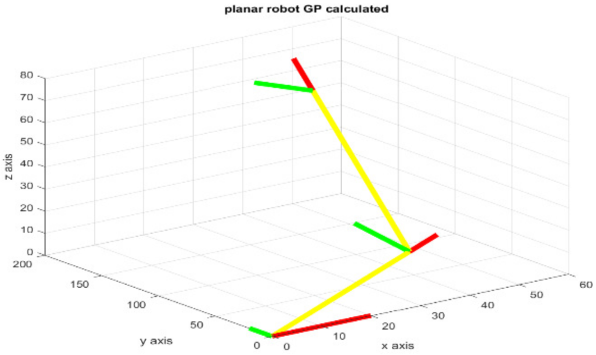


Fig. 4. Planar robot graph obtained by the GP.

The evolution of fitness of the GP can be seen in Fig. 5, which shows a desired behavior that favors the search for the correct result.

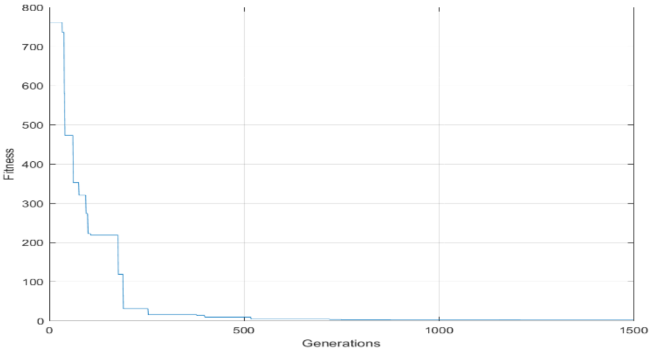


Fig. 5. Population fitness.

At the same time, Fig. 6 shows the best and worst element, where the difference between the two shows the diversity of the population, which disappears when it reaches the best value found.

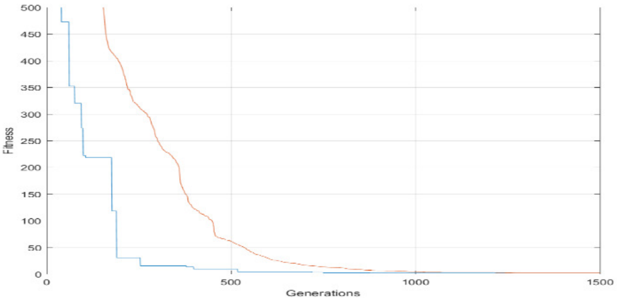


Fig. 6. Behavior of diversity.

The results shown above are only visual, but in order to corroborate mathematically the results, the equation (4) shows the expected evaluated matrix for the planar robot when substituting the links $l_1 = 300$ and $l_2 = 400$, and joints value with $q_1 = \frac{\pi}{3}$ and $q_2 = \frac{\pi}{4}$:

$$DH = \begin{bmatrix} 1.0672 & 0 & 100 & 0 \\ 0.7854 & 0 & 100 & 0 \end{bmatrix}. \tag{4}$$

In comparison with the equation (4), the equation (5) shows the GP algorithm solution in the robot using same values of equation (4) for links and joints, it can be seen that most of them are similar, in addition to the symbolic results are shown:

$$DH_{GP} = \begin{bmatrix} 1.0672 & 0 & 99 & 0.97 \\ 0.7854 & 0.02 & 100 & 0 \end{bmatrix}. \tag{5}$$

6 Conclusion

The relative error in the GP result is over 6.57%, this error comparing with the neuronal network develop is higher [23], [24], where the error has a 6% of difference, the same case happened using a PSO algorithm [25], with a 6.5% of error, that gets a lower difference in the absolute error but have an increase in execution time because the PSO use 0.5 milliseconds less by execution, finally the genetic algorithm [26] just have a 4% of error. The performance of GP is lower because the kind of algorithm used optimizes structure but not numerical parameters, i.e. GP is a soft search that just finds how the answer looks like and giving a really close approximation comparing to other three optimization algorithms, however if numerical parameters are optimized using other numerical optimization algorithm the solution could be improved over other works.

In this way it can be concluded that the problem posed by the introduction has a solution by the method selected, and that there is a result achievable by the GP for this problem, but numerical parameters must be optimized after that and structural solution is obtained.

The above mentioned gives rise to continuing to work on the search for the model and to generalize it, as well as to propose a solution to mathematically complex structures as preliminary results, the aim is to be able to add to the algorithm presented an automatic optimization process, this to compensate for the limitations of the GP; the code will be adapted to work parallel to the cycles reducing the execution time; tests will be carried out, which include the use of a redundant parallel robot that is the main problem of the FK; once satisfactory results have been found in the structural solution other numerical optimizing algorithm must be used for optimizing numerical parameters in the obtained structure, after that, the problem of inverse kinematics will be tackled, with the aim of finding a general computational method for the mathematical models of the robot.

References

1. Baturone, A.O.: Robótica: manipuladores y robots móviles. Marcombo (2005)
2. Saha, S.K.: Introduction to robotics. Tata McGraw-Hill Education (2014)
3. Montes, M., Padilla, A., Canul, J., Ponce, J., Ochoa, A.: Comparative of Effectiveness When Classifying Colors Using RGB Image Representation with PSO with Time Decreasing Inertial Coefficient and GA Algorithms as Classifiers. In: Fuzzy Logic Augmentation of Neural and Optimization Algorithms: Theoretical Aspects and Real Applications, pp. 527–546. Springer (2018)
4. Rivera, M.M., Justo, M.O.A., Zezzatti, A.O.: Equations for Describing Behavior Tables in Thermodynamics Using Genetic Programming: Synthesizing the Saturated Water and Steam Table. Res. Comput. Sci. 122, 9–23 (2016)
5. Li, M.: Multi-Objective Evolutionary Algorithms with Immunity for SLAM. Adv. Artif. Intell. 26, 27–36 (2006)
6. Duque, T., Delbem, A.: An Evolvable Hardware Approach. Adv. Artif. Intell. 37 (2006)
7. Ortiz, H.A.T., Montejano, C.A.P., Villegas-Cortez, J., Avilés-Cruz, C.: Evolución de descriptores estadísticos de superficie de imágenes por programación genética para el reconocimiento de imágenes por CBIR: una primera aproximación. Res. Comput. Sci. 116, 125–134 (2016)
8. Enríquez, G.B., Sánchez-Partida, D., Morales, S.O.C.: Algoritmo genético para el problema logístico de asignación de la mochila (Knapsack Problem). Res. Comput. Sci. 137, 157–168 (2017)
9. Vázquez-Castillo, V., Hernández-Lara, D., Merchán-Cruz, E.A., Rodríguez-Cañizo, R.G., Portilla-Flores, E.A.: Implementación de algoritmos genéticos para el diseño, optimización y selección de vigas. Res. Comput. Sci. 137, 121–134 (2017)
10. Hernández, M., Velasco-Arellano, H., Ubach-González, D., Montes-Rivera, M., Aguilar-Justo, M.O.: Steering Wheel Control in Lane Departure Warning System. Res. Comput. Sci. 147, 9–21 (2017)
11. Weise, T.: Global optimization algorithms-theory and application. Self-Publ. 2 (2009)
12. Ramirez, D., Ortiz, J., Ponce, P., Molina, A.: Sistemas inmunes artificiales aplicados a la robótica: agarre de objetos. Res. Comput. Sci. 135, 55–70 (2017)
13. Soriano, F.G., Montero, A.R., López, A.S.: Seguimiento autónomo de personas con un robot aéreo no tripulado. Res. Comput. Sci. 135, 71–84 (2017)
14. González, F.M.M., Hernández, A.M., Figueroa, H.R.: An effective robotic model of action selection. In: Conference of the Spanish Association for Artificial Intelligence (2005)
15. Coppin, B.: Artificial intelligence illuminated. Jones and Bartlett Publishers (2004)
16. Sivanandam, S., Deepa, S.: Evolutionary Computation. In: Introduction to Genetic Algorithms. pp. 1–13. Springer (2008)
17. Liu, Y., Kong, M., Wan, N., Ben-Tzvi, P.: A Geometric Approach to Obtain the Closed-Form Forward Kinematics of H4 Parallel Robot. J. Mech. Robot. 10, 051013 (2018)
18. Petrescu, R.V., Aversa, R., Akash, B., Berto, F., Apicella, A., Petrescu, F.I.: Geometry and direct kinematics to MP3R with 4× 4 operators (2017)
19. Merlet, J.-P.: Direct kinematics of CDPR with extra cable orientation sensors: the 2 and 3 cables case with perfect measurement and ideal or elastic cables. In: Cable-Driven Parallel Robots. pp. 180–191. Springer (2018)
20. Faulkner, J., Dirven, S.: A generalised, modular, approach for the forward kinematics of continuum soft robots with sections of constant curvature. In: Mechatronics and Machine Vision in Practice (M2VIP), 2017 24th International Conference on (2017)

21. Yang, X., Wu, H., Li, Y., Chen, B.: A dual quaternion solution to the forward kinematics of a class of six-DOF parallel robots with full or reductant actuation. Mech. Mach. Theory. 107, 27–36 (2017)
22. Mao, B., Xie, Z., Wang, Y., Handroos, H., Wu, H., Shi, S.: A hybrid differential evolution and particle swarm optimization algorithm for numerical kinematics solution of remote maintenance manipulators. Fusion Eng. Des. 124, 587–590 (2017)
23. Geng, Z., Haynes, L.: Neural network solution for the forward kinematics problem of a Stewart platform. In: Robotics and Automation 1991. Proceedings., 1991 IEEE International Conference on (1991)
24. Sadjadian, H., Taghirad, H., Fatehi, A.: Neural networks approaches for computing the forward kinematics of a redundant parallel manipulator. Int. J. Comput. Intell. 2, 40–47 (2005)
25. Chandra, R., Rolland, L.: On solving the forward kinematics of 3RPR planar parallel manipulator using hybrid metaheuristics. Appl. Math. Comput. 217, 8997–9008 (2011)
26. Boudreau, R., Turkkan, N.: Solving the forward kinematics of parallel manipulators with a genetic algorithm. J. Robot. Syst. 13, 111–125 (1996)
27. Hibbeler, R.C.: Mecánica vectorial para ingenieros: dinámica. Pearson Educación (2004)
28. Barrientos, A.: Fundamentos de robótica. e-libro, Corp. (2007)
29. Howard, L.M., D’Angelo, D.J.: The GA-P: A genetic algorithm and genetic programming hybrid. IEEE Expert. 10, 11–15 (1995)
30. McPhee, N.F., Poli, R., Langdon, W.B.: Field guide to genetic programming. (2008)

ISSN 1870-4069

Happiness and its Socio-demographic Determinants Analyzed with Datamining: the Case of a Community at the North of the Border of Mexico

Karla Erika Donjuan Callejo[1], Mario Ricardo Sotomayor[2], Alberto Ochoa-Zezzatti[1]

[1] Universidad Autónoma de Ciudad Juárez , Mexico
[2] Universidad TecMilenio, Mexico
erika@corporacionaem.com, ricardo@corporacionaem.com

Abstract. The topic of happiness and subjective welfare has taken a significant relevance inside the theories of several disciplines such as psychology, economics and politics; the importance of the individual subjective welfare. Happiness is a matter seen throughout many angles and it has an impact in the quality of life, therefore, we can definitely consider happiness multidimensional, as the variables affecting it are complex and diverse. This article analyzes which of the socio-demographic variables such as age, gender, occupation, scholarships, migration intention, and available family income are correlated to the level of Declared Happiness of the people who live in the community at the north of the border of Mexico: Juarez. Making use of the tools from datamining: the classification tree of the type CRT in order to segment the subjects and to correlate the main variable (Declared happiness) with the dependent variables, thus, generating a model able to predict the level of happiness.

Keywords: happiness, subjective welfare, quality of life, classification trees, datamining.

1 Introduction

The analysis and discussion of the quality of life during recent years has increased the number of theorists and investigations focused on the subjective welfare and overall happiness, this represents a boom in such topics around the world. No doubt, this matter remounts to Edward Diener (born in 1946), distinguished for his investigations on happiness and subjective welfare theories, which gave him the title “father of the study of happiness”.

Diener is the author of three scales that aid researchers in the evaluation of welfare: The satisfaction with life scale (SWLS) that measures cognitive global judgements of satisfaction with life (Diener, E.; Emmons, R. A.; Larsen, R. J.; Griffin, S, 1985) it has been widely used by diverse researchers, scientists and academics. The scale of positive and negative experiences (SPANE) which evaluates the frequency of experimenting a variety of positive and negative emotions. The scale of flowering that measures the subjective perception of success in important areas of life such as relationships, self-steem and optimism (Diener E; Ryan K., 2009). More and more

pp. 135–151; rec. 2018-09-10; acc. 2018-10-07

researchers talk about “welfare” instead of “happiness”, for the word happiness is only related with joy, while “welfare” involves a wide range of complex subjects. Seligman says is more useful to talk about welfare than happiness (Seligman, 2011). Other authors, such as Ed Diener, “father of the study of happiness”, use both words interchangeably.

For some, the study of the happiness means a revolution, understood from a methodological, theoretical and political stand (Bruno, Fay, 2010); Fay established the relation between happiness, utility, unemployment and inflation, as well as marriage and self-employment increase, he also proposed a relation between happiness and institutional policies.

The Advisory Forum Scientific and Technological A.C. published a document and points out that it is necessary to define welfare on the basis of what is relevant for people: it presents a view on what the subjective welfare is and how to measure it, based on six principles: (1) it is based on the welfare reported by the individual; the information is recovered from a survey of one or more questions related to happiness or life satisfaction; (2) as the question is made to an individual, the answer is given by concrete human beings; (3) recognizes that the welfare is essentially subjective for it is a self-experience of each individual; (4) the focus implies that each individual has the responsibility of playing the decisive role on happiness; (5) once it is accepted that there is relevant information on the welfare report, the focus follows a quantitative methodology to identify the important factors of the welfare of humans; (6) it is required an transdisciplinary approach or at least an interdisciplinary effort to understand the happiness reported by human beings (Foro Consultivo Cientifico y Tecnologico A.C., October 2012).

In the specific case of this paper, the following topic is developed: happiness focused on the case of a border city at the north of Mexico: Juarez City, municipality in the state of Chihuahua; where a non-lucrative and apolitical organization named *Plan Estrategico de Juarez* has the objective of constructing a citizen force invested and participative on public topics, proposing and demanding a better city (Plan Estratégico de Juárez A.C., 2018); one of its projects has the target of generating a system of city information to measure the quality of life of the citizens of Juarez: the System of Indicators *Asi Estamos Juarez (AEJ) or "This is how we are Juarez".*

According to the organization, *AEJ* has the purpose of “measuring the quality of life in Juarez, creating a system of data that both general society and the government can use to have a view of what the current situation is, then, define a path to follow and what we ought to do in order to get to an ideal target”. This system of data has information from official sources as well as a survey of perception made yearly since 2011 with an statistic representative sample and with a sampling technique that allows to generalize the results to the total population.

Each year, the survey of perception by *AEJ* has a sample of 1500 up to 1600 observations, through an structured survey, which are validated an applied face to face by an interviewer who is previously trained and supervised. The survey contains various topics impacting the quality of life and the subjective welfare of the juarenses, including education, health, corruption, security, citizen participation, among others such as variables of subjective welfare and happiness.

This article focuses in one of the variables: the Declared Happiness as a dependent variable and the correlation between the levels of happiness with socio-demographic determinants that are provided by the survey such as age, gender, occupation, scholarship, migration possibilities and familiar income. The objective is to find which of those demographic characteristics have an impact on the happiness of the citizens. Definitely the topic of happiness is complex and must be studied from different points of view, therefore, the limits of this particular study are accepted, however, the study pretends to find, firstly, the correlations of the happiness of the juarenses starting with some of their socio-demographic aspects with the possibility of incorporate more variables to the analysis in future studies.

For the analysis of the socio-demographic determinants that may define happiness, a tool of datamining called decision trees is used, this tool creates a model of classification based on flow diagrams whose objective is to classify the cases in groups or prognosticated values of a dependent variable (criteria) from independent variables (predictors). The decision trees are a statistical technique for the segmentation, stratification, prediction, reduction of data and filtering of variables, identification of interactions, categories fusion and discretization of continuous variables (Vaneza Berlanga Silvente; María José Rubio Hurtado; Ruth Vila Baños, 2013). Decision Trees are among the most used algorithms for solving supervised classification problems (Franco-Arcega, Carrasco-Ochoa, & Martínez-Trinidad, 2013). For example, a body shape predictor where they show how an ensemble of regression trees can be used to estimate face landmark positions directly from a sparse subset of pixel intensities, achieving real-time performance with high quality predictions. Similarly, they have trained a shape model by estimating body landmark locations to evaluate (Trejo & Angulo, 2016).

Since the objective is predicting the levels of happiness of a juarense, given the demographic characteristics, the classification and regression trees (CRT) were used, this trees consist of algorithms made from complete binary trees that make partitions of the data and generates accurate and homogeneous subsets; CRT divides the data in segments so it is as homogenous as possible with respect to the dependent variable.

2 Related Works

At present, there are many efforts and works to measure happiness and subjective well-being in the world. Examples of this are:

a) The World Happiness Report is a landmark survey about the state of global happiness. The World Happiness Report 2018, which ranks 156 countries by their happiness levels, and 117 countries by the happiness of their immigrants. was released on March 14th during a launching event at the Pontifical Academy of Sciences in the Vatican. A launch event was also held on March 20th, celebrating International Day of Happiness at the United Nations (Helliwell, Layard, & Sachs, 2018).

The average ladder score (the average answer to the Cantril ladder question, asking people to evaluate the quality of their current lives on a scale of 0 to 10) for each country, based on the average happiness of the targeted countries over the span of the following years 2015-2017.

Furthermore, the Happiness Report correlates and explains the average happiness ranking by country through 6 variables: GDP per capita, social support, healthy life expectancy, social freedom, generosity, and absence of corruption. In their findings regarding world happiness, Mexico can be found in the ranks of the first 25 countries, being positioned place 24th out of 156 countries.

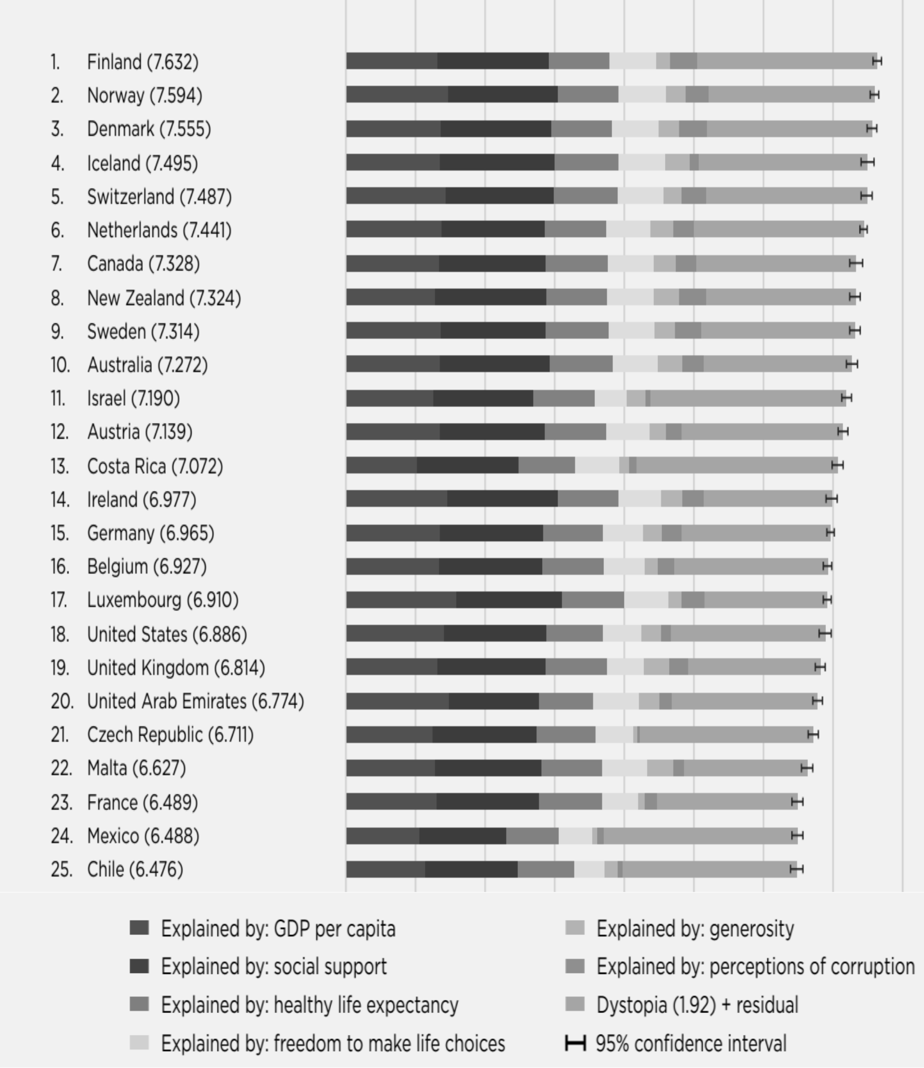
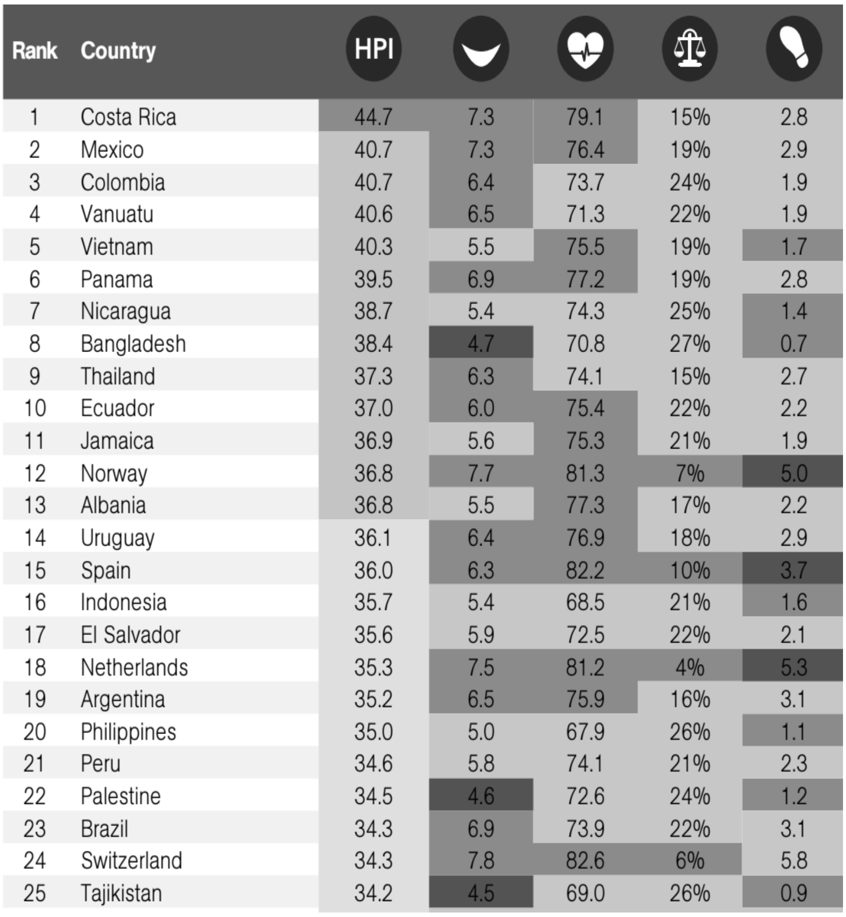


Fig. 1. Ranking of Happiness (Top 25). Source: The World Happiness Report 2015-2017.

b) The Happy Planet Index measures what matters: sustainable wellbeing for all. It tells us how well nations are doing at achieving long, happy, sustainable lives. Wealthy Western countries, often seen as the standard of success, do not rank highly on the Happy Planet Index. Instead, several countries in Latin America and the Asia

Pacific region lead the way by achieving high life expectancy and wellbeing with much smaller Ecological Footprints. The Happy Planet Index provides a compass to guide nations, and shows that it is possible to live good lives without costing the Earth (New Economics Foundation (NEF), 2016). The aforementioned index takes into consideration four determinants: (1)Wellbeing: How satisfied the residents of each country feel with life overall, on a scale from zero to ten. (2) Life expectancy: The average number of years a person is expected to live. (3) Inequality of outcomes: The inequalities between people within a country in terms of how long they live, and how happy they feel. (4) Ecological Footprint: The average impact that each resident of a country places on the environment. In its results about world happiness, Mexico can be found second place out of 140 countries of which it is compared.



Rank	Country	HPI	⌣	♥	⚖	👣
1	Costa Rica	44.7	7.3	79.1	15%	2.8
2	Mexico	40.7	7.3	76.4	19%	2.9
3	Colombia	40.7	6.4	73.7	24%	1.9
4	Vanuatu	40.6	6.5	71.3	22%	1.9
5	Vietnam	40.3	5.5	75.5	19%	1.7
6	Panama	39.5	6.9	77.2	19%	2.8
7	Nicaragua	38.7	5.4	74.3	25%	1.4
8	Bangladesh	38.4	4.7	70.8	27%	0.7
9	Thailand	37.3	6.3	74.1	15%	2.7
10	Ecuador	37.0	6.0	75.4	22%	2.2
11	Jamaica	36.9	5.6	75.3	21%	1.9
12	Norway	36.8	7.7	81.3	7%	5.0
13	Albania	36.8	5.5	77.3	17%	2.2
14	Uruguay	36.1	6.4	76.9	18%	2.9
15	Spain	36.0	6.3	82.2	10%	3.7
16	Indonesia	35.7	5.4	68.5	21%	1.6
17	El Salvador	35.6	5.9	72.5	22%	2.1
18	Netherlands	35.3	7.5	81.2	4%	5.3
19	Argentina	35.2	6.5	75.9	16%	3.1
20	Philippines	35.0	5.0	67.9	26%	1.1
21	Peru	34.6	5.8	74.1	21%	2.3
22	Palestine	34.5	4.6	72.6	24%	1.2
23	Brazil	34.3	6.9	73.9	22%	3.1
24	Switzerland	34.3	7.8	82.6	6%	5.8
25	Tajikistan	34.2	4.5	69.0	26%	0.9

Fig. 2. Happy Planet Index. Source: New Economics Foundation. The Happy Planet Index 2016. A Globlan index of sustainable wellbeing.

c) Better Life Index. This Index allows you to compare wellbeing across countries, based on 11 topics the OECD has identified as essential, in the areas of material living conditions and quality of life: Housing, Income, Jobs, Community, Education, Enviroment, Civic Engagement, Healt, Life Satisfaction, Safety, Work-Life Balance. in its results, Mexico obtained a grade bellow 5, on a scale from 0 to 10.

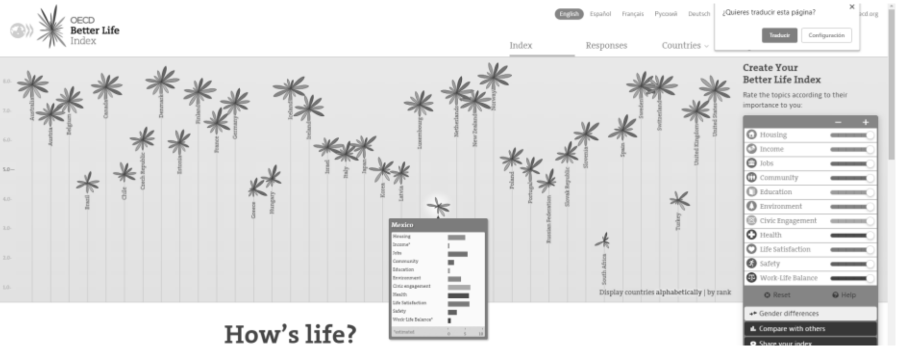


Fig. 3. Better Life Index. Source: OCDE, Better Life Index.

d) In 2007 Moyano and Ramos wrote about Subjective wellbeing: measuring life satisfaction, happiness and health in the Chilean population of the Maule Region. The purpose is to evaluate the subjective wellbeing through its cognitive components, measured as general satisfaction and its domains, and its affective components, measured as happiness, and analyze the relationship with sociodemographic variables. The sample was formed by 927 people, workers and students, between 17 and 77 years old, who answered three instruments.

The results indicate that the persons are satisfied with their life, being their family the principal source of happiness. Married people are happier and more satisfied than singles, and younger persons have lower level of happiness and satisfaction than older persons. Regarding to self-perceived health, men and women did not differ in their perception on general or physical health, although women reported a more negative perception about their mental health than men. Neither men nor women would stop working out of home even if they could do it. A direct and significant relation was found between happiness, self-reported health and vital satisfaction, which is supported by other studies in the field. Finally, there has been a positive correlation between income and welfare (Moyano Díaz & Ramos Alvarado, 2007).

3 Method

Happiness is a subjective concept that depends on diverse factors; for effects of this study, happiness is correlated with socio-demographic characteristics. The objective is to create a predictive model that allows to find out which of the 6 socio-demographic variables contained in the *AEJ* survey are correlated to the Declared Happiness by the juarenses. The model used is a decision tree that allows to classify the cases in groups to prognosticate the levels of happiness (criteria variable) based on the values of the independent variables (predictors). Thus, the classification and regression tree is accurate for the study.

The data comes from an statistical sampling of 1526 individuals, men and women living in Juarez municipality and with legal age, this means older than 18 years. The population (N) and the sample (n) are detailed in the following table:

Table 1. Sample.

Population (1)	Sample (2)	Confidence level and margin of sampling error (2)
1,391,180	1526	- 2.5% statistical error - 95% level of confidence

Source: own elaboration with information of (1) The National Institute of Geography, 2015 intercensal survey and information of *Plan Estrategico de Juarez A.C.* of its system of indicators *Asi Estamos Juarez*, Citizen Perception Survey 2016.

The population proportions statistical formula is used for the calculation of samples.

$$n = \frac{Z^2 * (p * q)}{e^2 + \frac{Z^2 * (p * q)}{N}} \tag{1}$$

in which:
n=Sample,
Z2=Sampling frame,
p=Success probability (Proportion of the population with the desired characteristics),
q=Failure probability (Proportion of the population without the desired characteristics),
e=error range (Level of error whiling to take),
N= Universe or population size.

The sampling technique used was probabilistic and multi-stage, surveys were applied face to face at the interviewee home address, distributed across multiple areas of the city.

4 Results: The Happiness of Juarenses and its Correlation with their Socio-demographic Determinants

The questions and scales used for each one of the variables are listed as follows:

1. Dependent Variable: the Declared Happiness

We've called this variable "Declares Happiness", since it implies the person's capacity to declare how much happy they are in a determined moment, that is to say, it affirms people's capacity of manifesting their happiness and adjust it to a determined scale.

The measuring of subjective welfare of the juarenses is made considering diverse factors, nonetheless, there is an specific question useful for the effects of the study: how happy are you? This question is answered using an scale of 1 to 10, where the extreme values 1 and 10, correspond to "not happy at all" and "very happy". This numeric scale allows the individuals to give a value to a subjective concept such as happiness. Other benefit of this scale is how comprehensible it is to the interviewed individuals thanks to the use of a similar scale in schools at Mexico where a 10 means the maximum scholar grade and 5 implies failing a class or assignment. The third advantage of the scale is it allows to calculate descriptive statistics: mean, mode and median of the Declared Happiness. Finally the problem of the measuring Semantic

Textual Similarity (STS), between words/ terms, sentences, paragraph and document plays an important role in computer science and computational linguistic is avoided (Majumder, Pakray, Gelbukh, & Pinto, 2016).

Table 2. Numeric scale (Declared Happiness).

1 2	**3 4**	**5 6**	**7 8**	**9 10**
Not happy at all	**A little happy**	**Somewhat happy**	**happy**	**Very happy**

The next tables show the frequency of answers, the result is an average of 8.27 in the Declared happiness of the people of Juarez.

Table 3. Declared Happiness.

Declared Happiness					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	17	1.1	1.1	1.1
	2	4	.3	.3	1.4
	3	14	.9	.9	2.3
	4	32	2.1	2.1	4.4
	5	75	4.9	5.0	9.4
	6	59	3.9	3.9	13.3
	7	133	8.7	8.8	22.0
	8	404	26.5	26.7	48.7
	9	329	21.6	21.7	70.4
	10	448	29.4	29.6	100.0
	Total	1515	99.3	100.0	
Missing	System	11	.7		
Total		1526	100.0		

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

Table 4. Descriptive Statistics Declared Happiness.

Declared Happiness		
N	Valid	1515
	Missing	11
Mean		8.27
Median		9.00
Mode		10

Source of Table 4: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

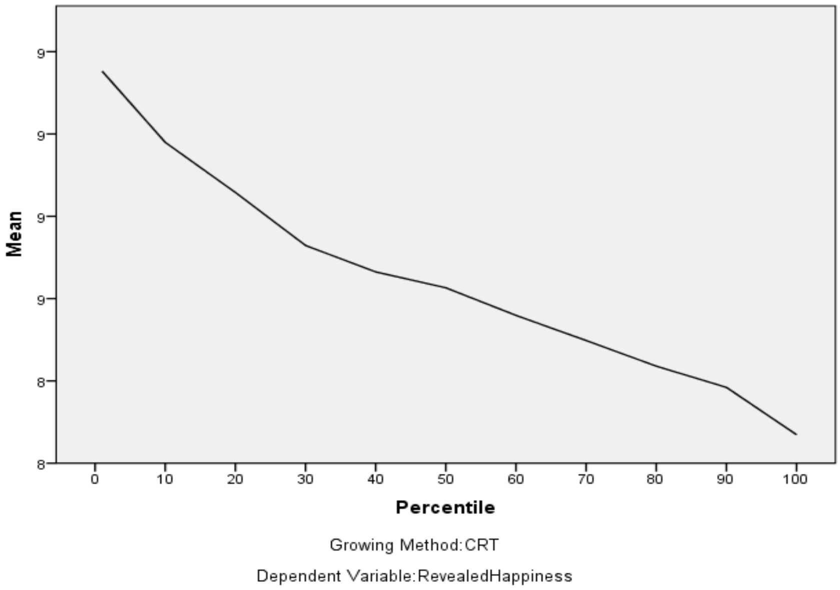


Fig. 1. Mean and Percentile (Dependent variable: Declared Happinness).

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

2 Independent Variables

As mentioned before, the study pretends to identify which variables have a correlation with the dependent variable. The socio-demographic data from the survey that belongs to the juarenses and will contribute to finding the Declared happiness is shown in the next table:

Table 5. Independent variables.

#	Independent variable	Observations
1	Age	Numerical variable. Considering the survey is answered by legal citizens, this means people with over 18 years of age.
2	Gender	Codified variables: 1. Woman 2. Man
3	Principal occupation	Codified variables: 1. Housekeeper 2. Employee of private company 3. Employee of the government 4. Unemployed 5. Student 6. Student who works 7. Employer 8. Retired 9. Self-employment

#	Independent variable	Observations
4	Scholarship	Codified variables: 1. Illiterate 2. Complete elementary school 3. Incomplete elementary school 4. Complete middle school 5. Incomplete middle school 6. Complete high school 7. Incomplete high school 8. Complete degree (college) 9. Incomplete degree (college) 10. Postgraduate 11. Complete technical career 12. Incomplete technical career 13. Knows how to read and to write but did not went to school
5	Migration intention	This variable is measured with a scale of four points that aims to size how much the population have thought of moving out of the city. The specific question is: In the last year, how often did you thought of moving out of the city? And the codified variables are: – A Lot of times – Sometimes – Few times – Never
6	Familiar income	This variable is measured with a scale of four points as well and is about the familiar income and how much this is or is not enough for the consumption and savings of the families in the city. The answers are listed as follows: With the total of the familiar income, you would say that… 1. It is more than enough and we can save money 2. It is just enough with no difficulties 3. It is not enough and have some difficulties 4. It is not enough and have a lot of difficulties 0. Does not know / Did not answer

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

The following tables contain the frequencies in the responses of the Juarenses, for each of the independent variables.

Table 6. Independent variable: Gender.

Gender					
Code		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	869	56.9	60.1	60.1
	2	577	37.8	39.9	100.0
	Total	1446	94.8	100.0	
Missing	System	80	5.2		
Total		1526	100.0		

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

Table 7. Independent variable: Principal Occupation.

Occupation					
Code		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	456	29.9	30.5	30.5
	2	378	24.8	25.3	55.8
	3	159	10.4	10.6	66.5
	4	55	3.6	3.7	70.1
	5	60	3.9	4.0	74.2
	6	54	3.5	3.6	77.8
	7	52	3.4	3.5	81.3
	8	138	9.0	9.2	90.5
	9	142	9.3	9.5	100.0
	Total	1494	97.9	100.0	
Missing	System	32	2.1		
Total		1526	100.0		

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

Table 8. Independent variable: Scholarship.

Scholarship					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	26	1.7	1.7	1.7
	2	248	16.3	16.4	18.1
	3	102	6.7	6.7	24.9
	4	372	24.4	24.6	49.5
	5	79	5.2	5.2	54.7
	6	184	12.1	12.2	66.9
	7	62	4.1	4.1	71.0
	8	158	10.4	10.4	81.4
	9	139	9.1	9.2	90.6
	10	25	1.6	1.7	92.3
	11	91	6.0	6.0	98.3
	12	9	.6	.6	98.9
	13	17	1.1	1.1	100.0
	Total	1512	99.1	100.0	
Missing	System	14	.9		
Total		1526	100.0		

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

Results of the Model: Classification Tree

Fig. 4 shows the resulting classification tree.

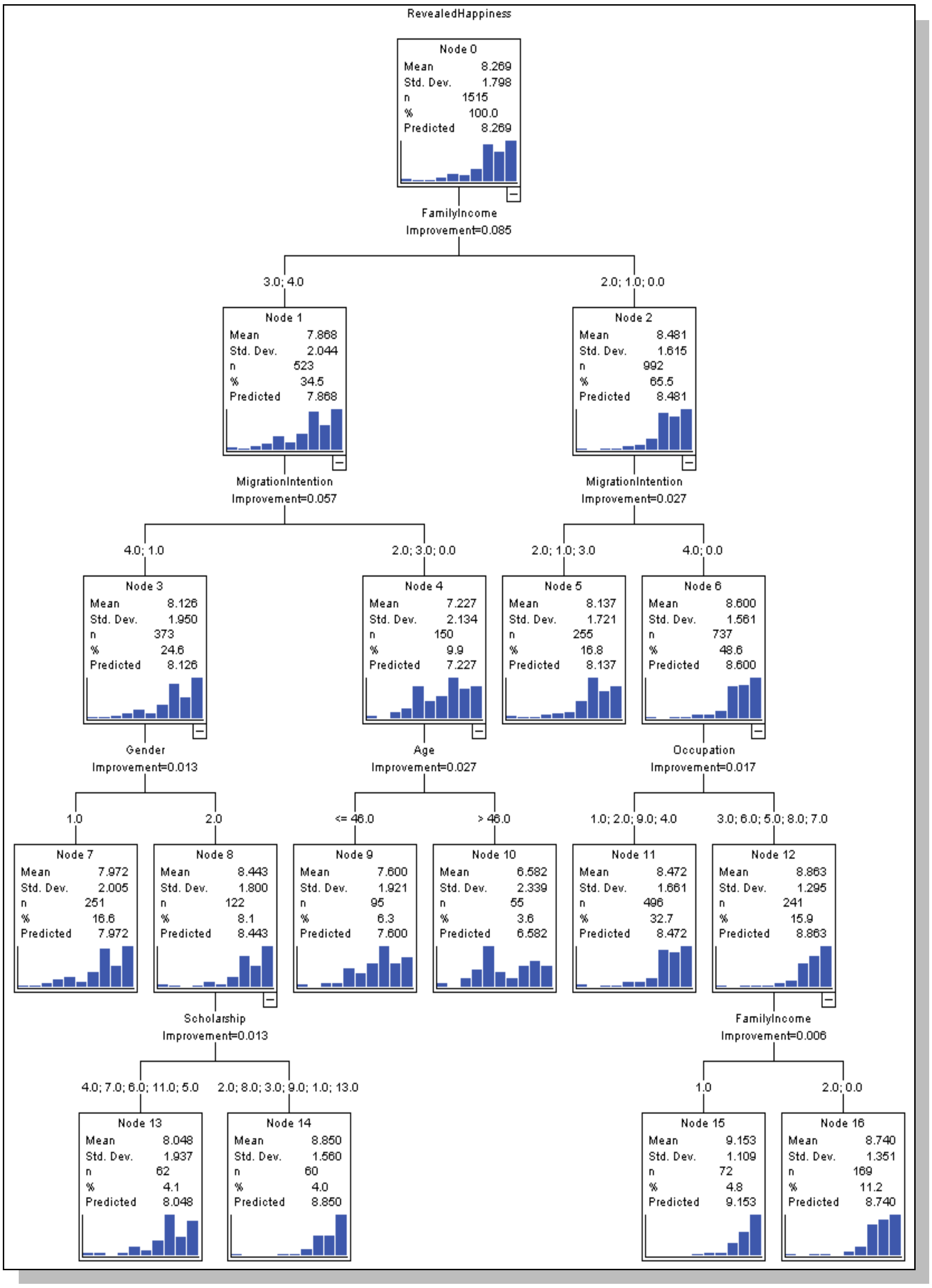


Fig. 4. Classification Tree.

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

Table 9. Independent variable: Migration intention.

Migration Intention					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	24	1.6	1.6	1.6
	1	94	6.2	6.2	7.7
	2	192	12.6	12.6	20.3
	3	161	10.6	10.6	30.9
	4	1055	69.1	69.1	100.0
	Total	1526	100.0	100.0	

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

Table 10. Independent variable: Family Income.

FamilyIncome					
Code		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	18	1.2	1.2	1.2
	1	189	12.4	12.4	13.6
	2	793	52.0	52.0	65.5
	3	465	30.5	30.5	96.0
	4	61	4.0	4.0	100.0
	Total	1526	100.0	100.0	

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

Interpretation of the classification tree

When the model is tested, the variables of gender, occupation and scholarship are discarded as the main predictive variables of happiness. The two most important predictors of Declared Happiness are: family income and the intention to migrate.

1. The node 0 represents the independent variable which is the happiness of the individual, from 1515 observed cases, the average of happiness was 8.269.
2. Principal predictive variable: Therefore is observed that the dependent variable (happiness) branches into two nodes. The node 1 and node 2 belong to the variable of familiar income, so the classification says this is the principal predictive variable. If the results are analyzed, it is observed that the group who answered their income is “more than enough and we can save” is happier as well as the group of those who said their income is “just enough with no difficulties”, both with an average of happiness of 8.48 (this group is the node 2). As for the node 1, the groups whose income is “not enough and have some difficulties” and “not enough and have a lot of difficulties” have an average of happiness of 7.86 out of 10 points.
3. Second predictive variable: The following structure of the tree branches into 4 more nodes, all of them from the second predictive variable: has thought of moving out of the city. From the node 1 are given the nodes 3 and 4, even though in both cases it is observed a lower average of happiness than the

nodes 5 and 6, coincide that the juarenses who have never thought of leaving the city are, in fact, the ones with a higher mean of happiness. The nodes 5 and 6 contain the juarenses with better familiar income and when they are reclassified, they are also the people who have never thought of leaving the city.

4. Third predictive variable. node 3 divides into the third predictive variable which is the gender of the juarense, and, node 4 divides into the predictive variable which is the age. From the node 3 are divided the nodes 7 and 8, the results of predictive variable of the gender say men are happier than women . While from the node 4 are given the nodes 9 and 10, the results of the predictive variable of the age say the happier people have less than 46 years. From the node 5 there are no more predictive variables, and, from de node 6 is divide the predictive variable is de occupation.
5. The last predictive variable from the node 8 is Scholarship and from node 12 the family income.
6. The independent variables importance with respect to the dependent (Declared happiness) sum together 28.7%, that is, happiness is determined by them in that percentage. That is the impact of sociodemographic characteristics on the happiness of the people of Juarez.

Table 11. Independent Variable Importance.

Independent Variable Importance		
Independent Variable	Importance	Normalized Importance
FamilyIncome	.104	100.0%
MigrationIntention	.085	82.2%
Age	.038	36.5%
Occupation	.027	26.5%
Scholarship	.020	19.6%
Gender	.013	12.8%

Growing Method: CRT
Dependent Variable: DeclaredHappiness

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

Table 12. Risk.

Risk		
Method	Estimate	Std. Error
Resubstitution	2.991	.173
Cross-Validation	3.202	.181

Growing Method: CRT
Dependent Variable: DeclaredHappiness

Source: own elaboration with information of Plan Estrategico de Juarez A.C. of its system of indicators Asi Estamos Juarez, Citizen Perception Survey 2016.

5 Discussion and Conclusions

With the model of classification regression tree is established that, in the case of the juarense community and the Declared happiness of the year 2016, it is found that gender, occupation and scholarship are not the main variables that influence happiness. Thus, it is established nodes that define the socio-demographic profile for happiness (variables that influence in more happiness) are the variables of familiar income as principal predictive variable as far as the income is more than enough that people may even save money; that the possibility of leaving the city is not in the mind of the individuals and the in terms of age, the younger they are, the happier. That is to say, it is possible to predict that an individual with enough income to satisfy their necessities and to save, that has never thought of leaving the city and has an age less than 46 years, is more likely to have higher levels of happiness than an individual with not enough income and difficulties that also has thought of leaving the city.

However, it is important to note, although family income was the main predictable variable of happiness, there are multiple authors who claim the aforementioned statement is only true when talking about poor and third world countries, since they have found a close to zero correlation between income and happiness in rich countries, they also have found that the relation between income and wellbeing have relevant effects in situations of extreme poverty and that individuals adapt to their corresponding economical level, hence, a lost of economic status can lead to unhappiness (Veenhoven, 1994; Dinner, 1994; Dinner E, Suh, E., , Lucas, R., & Smith, H., 1999).

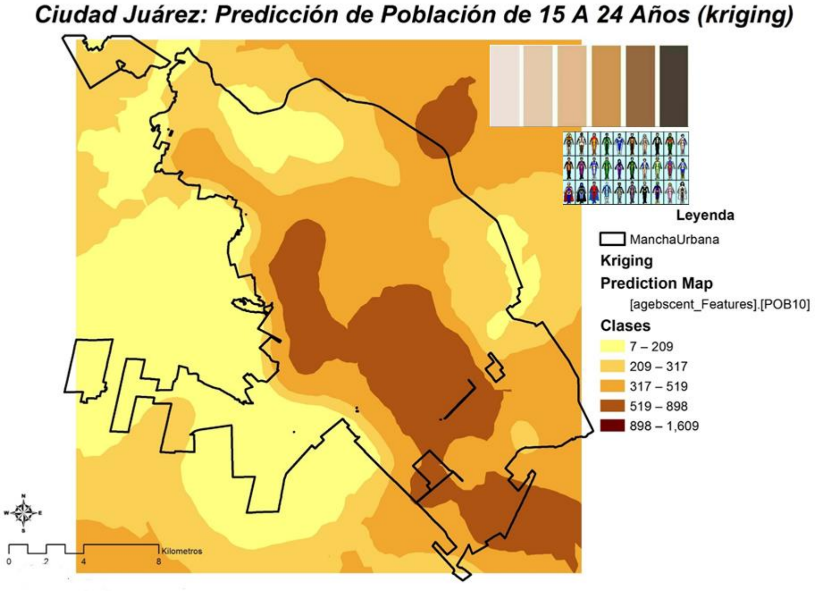


Fig. 4. Prediction model to 2030 including distribution of young people and a future opportunities model. (Source: Adapted of a map using Kriging Model generated in Colech by Luis Cervera.)

Definitely the subject of happiness is complex and is determined by a lot of variables as well as the disciplines that study it. So this study is limited to trying to find out a socio-demographic profile with 6 variables in order to aid in the improvement of happiness research. There is still a lot of work to do to understand everything we need about this topic in order to change the subjective welfare of people and to improve the quality of life.

With the model of classification regression tree is established that, in the case of the juarense community and the Declared happiness in the year 2017, was not found that gender, occupation and scholarships are the main variables that influence happiness. Thus, it is established nodes that define the socio-demographic profile for happiness (variables that influence in more happiness) are the variables of familiar income as principal predictive variable as far as the income is more than enough that people may even save money; that the possibility of leaving the city is not in the mind of the individuals and the age, the younger they are, the happier. That is to say, it is possible to predict that an individual with enough income to satisfy their necessities and to save, that has never thought of leaving the city and has an age less than 46 years, is more likely to have higher levels of happiness than an individual with not enough income and difficulties that also has thought of leaving the city.

Definitely the theme of happiness is complex and is determined by a lot of variables as well as the disciplines that study it. So this study is limited to try finding out a socio-demographic profile with 6 variables in order to help improving the research on happiness. There is still a lot of work to do and to understand about this topic to change the subjective welfare of people and to improve the quality of life.

In our future work, we will seek to generate a shared index values by regions in order to adequately model emerging public policies for each segment of the population, considering their geospatial location as can be seen in Figure 4.

References

1. Diener, E; Ryan, K.: Subjective well-being: A general overview. Journal of Psychology, 391-406 (2009)
2. Diener, E., Emmons, R.A., Larsen, R.J., Griffin, S.: The satisfaction with life scale (1985)
3. Dinner, E, Suh, E., Lucas, R., Smith, H.: Subjetive well-being: three decades of progress. Psychological Bulletin, 276-302 (1999)
4. Dinner, E.: El bienestar subjetivo. Intervención psicosocial. Revista sobre igualdad y calidad de vida, 67-113 (1994)
5. Foro Consultivo Científico y Tecnológico A.C.: Medición, Investigación, e Incorporación a la Política Pública del Bienestar Subjetivo: América Latina. Comisión para el Estudio y la Promoción del Bienestar en América Latina, México (Octubre 2012)
6. Franco-Arcega, A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: Decision Tree based Classifiers for Large Datasets. Computación y Sistemas 17(1), 95-102 (2013)
7. Helliwell, J., Layard, R., Sachs, J.: World Happiness Report. Reporte, The World Happiness Report was written by a group of independent experts acting.

Recuperado el Junio de 2018, de https://s3.amazonaws.com/happiness-report/2018/WHR_web.pdf (2018)

8. Majumder, G., Pakray, P., Gelbukh, A., Pinto, D.: Semantic Textual Similarity Methods, Tools, and Applications: A Survey. Computación y Sistemas, 20(4), 647-665 (2016)
9. Moyano-Díaz, E., Ramos-Alvarado, N.: Bienestar subjetivo: midiendo satisfacción vital, felicidad y salud en población chilena de la Región Maule. (Talca, Ed.) Universum, 22(2). Recuperado el Junio de 2018, de https://scielo.conicyt.cl/scielo.php?pid=S0718-23762007000200012& script=sci_arttext&tlng=pt (2007)
10. New Economics Foundation (NEF). The Happy Planet Index. NEF. Recuperado el Junio de 2018, de https://static1.squarespace.com/static/5735c421e321402778ee0ce9/t/57e0052d440243730fdf03f3/1474299185121/Briefing+paper+-+HPI+2016.pdf (2016)
11. Plan Estratégico de Juárez A.C. www.planjuarez.org. Obtenido de www.planjuarez.org: https://www.planjuarez.org/index.php/quienes (Junio de 2018)
12. Seligman, M.: Flourish: A Visionary New Understanding of Happiness and Well-being. Free Press (2011)
13. Trejo, K., Angulo, C.: Single-Camera Automatic Landmarking for People Recognition. Computación y Sistemas, 20(1), 19-28 (2016)
14. Berlanga-Silvente, V., Rubio-Hurtado, M.J., Vila-Baños, R.: Cómo aplicar árboles de decisión en SPSS. REIRE, Revista d´Innovació i Recerca en Educació, 65-79 (2013)
15. Veenhoven, R.: El estudio de la satisfacción con la vida. Intervención Psicosocial, 3, 87-116 (1994)

ISSN 1870-4069

"Interconnection" APP: Proposal of Interaction with a Virtual Agent, Animations and Augmented Reality an Easy Way to Learn the Usage of Sensors in Smart Cities

César Lozano Díaz[1], Adriana Lorena Iñiguez Carrillo[1], Rocio Maciel[2],
Víctor M. Larios [2], Emmanuel Aceves Martínez[1], C. Alberto Ochoa[3],
Edgar G. Cossio Franco[4]

[1] University of Guadalajara, Zapopan, Mexico
clozano@cucea.udg.mx, adriana.carrillo@cusur.udg.mx,
emmanuel.aceves@cucea.udg.mx
[2] University of Guadalajara, Smart Cities Innovation Center, Zapopan, Mexico
rmaciel@cucea.udg.mx, vmlarios@cucea.udg.mx
[3] Universidad Autónoma de Ciudad Juárez, Mexico
alberto.ochoa@uacj.mx
[4] Instituto de Información Estadística y Geográfica de Jalisco, Mexico
kofrran@gmail.com

Abstract. Most of the technological resources used in a Smart Building environment are automated systems where sensors and their usability become a "black box" for their personnel (end user of these systems), giving in them a wrong impression about Who must be taking care of these resources and services? (These users think these services must be only provided and controlled by the own building), losing the culture of the correct usage of the regular services such as artificial climate, water services, electricity services, and the care of plants outside of these buildings. This paradigm of interaction between the user and the maintenance of resources of a smart building still needs improvement. For that reason, this current research is proposing the inter-connection of following resources: use of sensors, communication protocols, collections of MongoDB data, speech to text, text to speech and animations that represents feelings (sadness, happiness, worry) through the usage of Augmented Reality (AR) and IBM Watson conversation (Conversational Artificial Intelligence A. I.). The results show the architecture and favorable results about the viability to connect the services with natural interfaces. This approach helps and increments the user interaction with the places and objects that are around in the context of a Smart Building to understand the importance of the resources that are administrated and create a collective conscience through the correct usage of resources.

Keywords: smart building; interconnection; conversational agent; augmented reality.

1 Introduction

The implementation of technology based on simple interactions, intuitive and satisfactory improves the user experience. For that reason, researches have increased the development of new forms to interact with technology, and the world surrounds us to incorporate more natural forms of interaction for the user like the usage of speech, gestures, and visualization that enriches the visual experience like the Augmented Reality improving the quality of communication. These types of interfaces are called Natural User Interfaces (NUI). This interface aims to reduce the barrier between user and application, getting closer and closer to a way in which people interact with each other in their daily activities. The applications developed for the end user with NUI are capable of interacting with minimal training or null; the use of these principles of design appropriately, it will be easier the usage and execution of some tasks even for those users without experience. Along with the technological development, new tools have emerged and have revolutionized this interaction. For this research, the following tools have been identified.

The speech interaction has been used in diverse context, for instance, to active some instruction in immersed environments: education [1] or accesses to other services as a document, calendar or website. The Virtual and Augmented Reality have used in a wide variety of fields, such as entertainment [2], health [3], inclusion [4], education [5], manufacturing and logistics [6] and transportation [7], as well as the usage of sensors that allows monitoring Smart Buildings [8]. More specific in buildings the AR systems can help to read real-time measures such as temperature, oxygen, sound, movement, or energy consumption. Although, this last one requires friendly interfaces or previous training for the user to be able to understand this information.

Lastly, the growth of technologies for conversational agents with natural language is now used from simple request interfaces and response capability to full sentences, used in a whole range of scenarios, (customer support, air traffic control, or other situations). These technologies provide access to information in a faster way, and the learning curve is relatively short.

All these resources have been utilized in Smart cities environments in several ways, combining them or separately managing services in an automated way in the buildings [7]. However, in several cases, the usage of sensors administration and functionalities become a “black box” for the end users in these buildings. These systems at the building must provide the wrong impression to the society about the self-regulation and administration of the resources, and merely the fix of a sensor or the care of the resource must be managed by an IT expert only. This approach of smart buildings means that if desired to create collective conscience through the correct usage of resources such as water, electricity, the care of plants inside these buildings, it is necessary to ensure a direct and indirect interaction between the users and the improved services.

This paradigm of interaction between the user and the maintenance of resources of a smart building still needs improvement through an easy way to connect to the user with the services and interaction, similar to human or animal contact on a daily basis. It has not been well defined yet, therefore this research is proposing to implement the App develop with the possibility to interconnect the use of sensors, communication protocols, collections of MongoDB data, speech to text, text to speech and animations with AR that represents feelings (sadness, happiness, worry) adjusted to the measured parameters. The information used in sensors from the "interconnection" app through the Augmented Reality and a conversational system using an avatar form to improve these interactions. In this research, architecture and usability testing have exhibited favorable results connecting these technologies. That allows increasing the interaction and perception of the user about places and objects around in a context. For example, into a Smart Building, it is possible to provide the sensation that the plants and services are living beings that can interact in real time with the end user. Enabling the place for the users to ask questions and to receive responses with information of components condition to interact better with the building. The user interface was possible with the help of the conversational agent IBM Watson and the presentation in real time of the information that used in sensors.

2 Background

2.1 Smart Building

The vision of the cognitive building is to connect with the interactive design which operates proactively with the human activity inside them [10]. The exchange between the sensing and the systems occurs through the collection of data and the data process based on advance learning of technologies [11]; many of these buildings attempt to implement rules for different scenarios, for instance, the usage of energy [12]. The data is coming from sensors installed inside the building and systems that manage various functions defined to create or maintain an artificial environment. Another example will be a scene inside a school when a sensor informs the air quality inside the classroom. If the air is not good, the building will activate the ventilation, and the room will improve the comfort and the health of the individual. Also, the building will collect the consumption of energy to detect if the air conditioning system is working correctly [13].

2.2 Augmented Reality

AR allows users to see the real world, with virtual objects, superimpose or composed with another world [14]. Therefore, AR supplements the reality, instead of replacing it entirely. Ideally, it will seem to the user that the virtual and real objects coexist in the same space increasing the vision of the external world with the digital information. The real environment is aligned with the 3D space as a fusion of the content and context. That fusion is called the immerse multisensory environments and the applications of Augmented Reality provide opportunities for new electronic services for the Smart Cities.

In the context of mobile AR apps for smart cities, it can consider the so-called "video-see-through" apps: mobile applications that in real-time overlay information on top of the live video stream. The first and essential mobile AR apps used only the data coming from two sensors from the following devices: GPS receiver (for the location), and a digital compass (for the direction). [15]

Other categories of applications of the Augmented Reality (AR) are the montage, maintenance and fix of complex machines with instructions that are much easier to understand. Such information in AR replaces traditional instructions booklets with or text or images. AR can use 3D draws superposed to the real-time video, using a mobile device (Tablet of Smart Phone) showing step by step the tasks to develop and how to do them on its visual interface.

2.3 Speech Services

In the past years two significative enhancements appeared in systems based on speech. Such systems, for example, are used in customer support and personal agents as well they integrate services into smart devices.

With the miniaturization of modern electronics, the implementation of Internet of things (IoT), the progress of artificial intelligence and the precision of the algorithms of speech recognition (Automatic Speech Recognition-ASR) and speech systems (Text-to-Speech-TTS) it makes possible to use interfaces with the speech with natural language processing integrated into smart environments. Such interfaces are capable of acquiring and apply in knowledge over their population and their environment to adapt and compliance with their objectives with efficiency.

These interfaces must be sufficiently flexible to make use of a variety of devices, services, knowledge sources and user supplies. Also, with current processing capacities, speech interfaces are being able to manage simultaneously multiple tasks covering different aspects of communication, such as commands dialogues and controls, as well as the dialogues for recovering information [16].

2.4 User Experience Assessment

For the citizens in Smart Cities is essential to move freely and have personal or public devices allowing access to the services practically and straightforwardly. Looking to produce the experience of the user-friendly system systems to avoid a digital divide among citizens.

The user experience (UX) merged of the perception, the action, the motivation and the cognition in all the sense that defines the user interaction with a system [17,18]. The usability has been utilized to measure the user experience in the interactive systems. The ISO 9241-11 norm [19] defines the usability as “the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use". There are different methods and tools to measure the usability [20].

In general; three categories are considered: one at inspection methods, two for inquiry and three for testing. However, need to consider measuring the usability of the system with AR and the conventional systems. Also, it is necessary to consider the combination of real and virtual objects in scenarios and the necessity of real-time interaction.

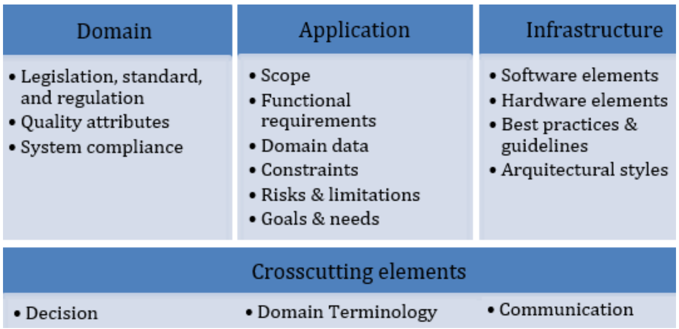
3 Proposed Approach

3.1 Interconnectivity and Model

The growth of new technologies is allowing researchers to experiment with different methods of human-computer interaction. By giving the users an opportunity to understand the environment that surrounds us as an interactive and dynamic way in multimodal interfaces, creating exhaustive user studies to determinate how to implement this technology in relevant scenarios [21] showing its viability to support complex tasks interaction [22]. Proposing to define how an application needs the interaction before its development, this present research is based on the concept of Ambient and Active and Assisted Living (AAL) program, where the organizations need to share their common information coming from different sources and helping provide better assistance [23]. All the elements are interacting and are affected between them in such AAL [24]. An AAL detects and recognizes the actions, activities, and situations inside an environment, which use different sensors to recognize the activities and the comprehension of the behaviors [25].

A reference model is a RAModel [26], in which handles four dimensions to take in consideration: domain, application, infrastructure and transverse elements (Table1).

Table 1. Domain, application, and infrastructure.



Domain	Application	Infrastructure
• Legislation, standard, and regulation • Quality attributes • System compliance	• Scope • Functional requirements • Domain data • Constraints • Risks & limitations • Goals & needs	• Software elements • Hardware elements • Best practices & guidelines • Arquitectural styles

Crosscutting elements		
• Decision	• Domain Terminology	• Communication

The main designs for the AAL systems adapted to the physical and cognitive capabilities, as well as the day to day activities, (ADL) of the user.

4 Architecture of the System

In this current research is being proposed the approaching with every user in the same manner they can access to the information obtained from sensors. The analysis of this information and modality of the system in Audio-Chabot mode through a conversational agent in the user's mobile device; this is done with the purpose of monitoring in real time the sensing results (understanding all the measured parameters), administrate and make them part of a preventive maintenance of the building services.

The document design uses the requirements of the IT managers of the Smart Cities Innovation Center at the University of Guadalajara CUCEA (not directly from the user, so a speculative interface to use is proposed, and improvements of the app expected in the future).

Therefore, the value of this proposal is in this app design "interconnectivity" in which is useful for the administrative services connected in the smart building (sensors that collect measures of temperature, humid, pollution, acoustics, air quality and light conditions), through IoT (Internet of Things) devices. All devices communicate in real time through the MQTT protocol sending information to the server, saving it in a non-relational database MongoDB and taking the data to an application in AR to show the status of a sensor. In this way, the interconnection with the information used in the sensors with a speech to text services with a conversational agent could be “translated” by the agent linked to frequently asked questions. The user could ask (programmed in the Watson Conversation Service), getting responses with negligible value, supreme and ranks. The system can determine the optimal state of the service, errors, preventive actions or also a social interaction simulated (greeting, telling a joke, provide status in general of the room). All this to make the users part of preventive maintenance and they recognize the regulations of the system and create a collective conscience to the correct usage of the resources in a smart building.

In this way, the automation of these buildings it is not a simple action loop/ reaction, or else could be modified based on its most crucial human component. Like Rinaldi [27] proposing to take advantage of the digital capabilities with the power of cognitive computing. Cognitive systems, like IBM Watson ™, the system of questions and answers and derive conclusions of textual content.

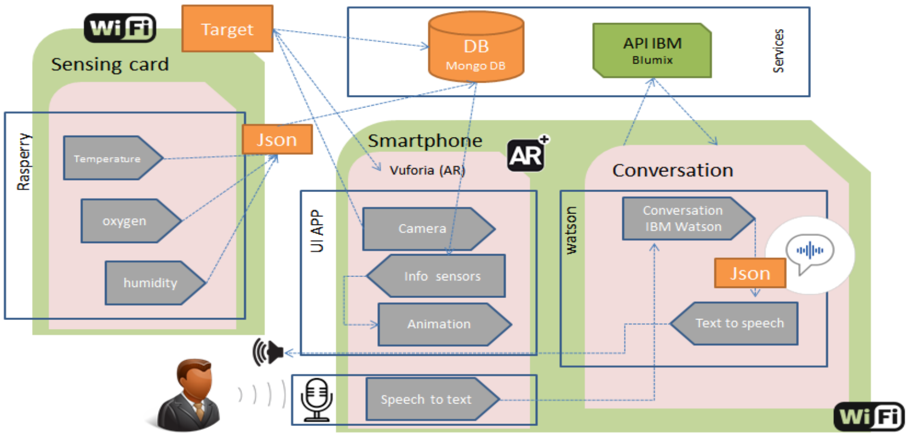


Fig. 1. The app interconnection between services.

Natural speech interaction is used, combined with other interactive modes to devise a multi-modal UI. Hence, researchers are experimenting with different modalities, often combining some of them into mixed interfaces and performing comprehensive user studies to determine their applicability in relevant scenarios and Multi-modal interfaces proved their viability in supporting complex interaction tasks. It is necessary to take into account that the research of a multi-modal interface to design need to have the following services: speech to text, text to speech, Augmented Reality, storing data in Mongo DB database and the service of IBM Blumix Watson conversation. This last service receives user inputs and processes them through the use of machine learning. With the analysis of natural language, it can assign defined intentions to entities. Reference Fig. 1.

The data transfer from the IoT devices is through the WiFi, taking to the database MQTT the records. From Unity, the reception of the sensed data depends on each element of the building through Vuforia in Unity. This target also works as a reference for the conversational of Watson service. The data is stored to open the possibility to do analytics (Fig. 2).

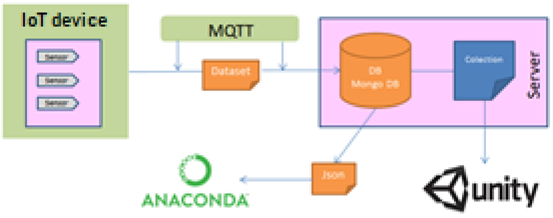


Fig. 2. Connection from Unity.

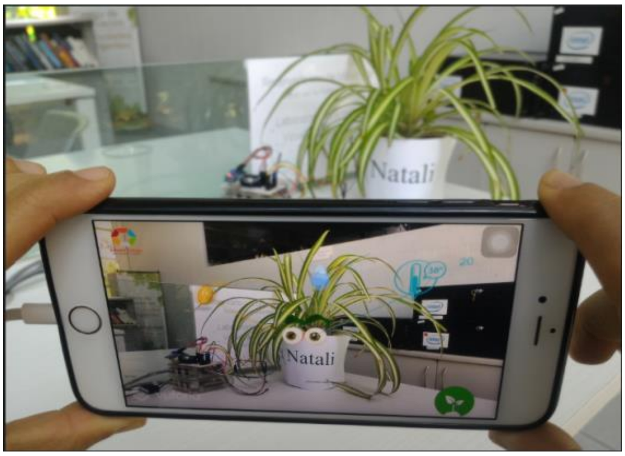


Fig. 3. UI utilized in a plant in the smart cities innovation center.

The data emerge to the UI through floating icons superimposed around the plant which each of the parameters measured (temperature, humid, luminous intensity). The interface also creates an AR with animated eyes and mouth. The animation depends on the "mood" of the plant, this base on the parameters measured, simulating sadness, happiness, worry. The system for speech to text is active in all moment; a button is added to start the conversation, in this case with the plant. Reference Fig. 3, the target is not a QR code else is tracking without marks base on Barroso [28] using the method of level two for AR, the user uses a microphone and speakers from their mobile phones to interact with the plant.

5 Usability Test

The assessed two evaluators, which they have experience in the AR concepts, conversational interfaces, and assessments in user experience. The study comprises an assessment of 14 volunteers (Men = 6, Women = 8) between professors and students of pregrade and postgraduate in the technology area, between ages from 18 to 54.

Individually assessment was done. First, a brief introduction of the app services to each one of the participants. Then, instructions were given to complete a series of activities (Reference Table 2). The participants were encouraged to ask questions and assistance was provided (in case of not completing with a specific task). When concluding the tasks, the participants could explore the application freely. In the end, asked the participants their feedback and they provided comments and suggestions to improve the user experience.

Table 2. Defined tasks.

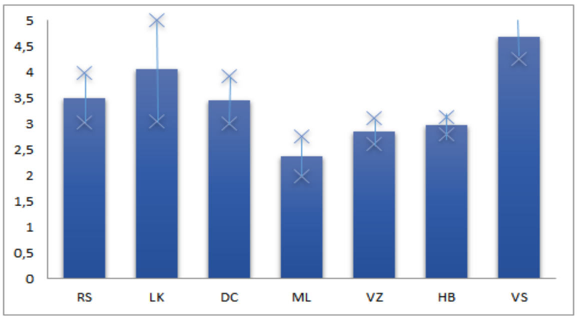
(a) Open the application
(b) Greeting
(c) Knowing what type of plant is
(d) Knowing the plant's mood
(e) Discovering the ideal parameters for the plant
(f) Saying goodbye

The architecture assessment is based on the adaption of the questionnaire of the Subjective assessment of speech-system interfaces (SASSI) [25] and in the heuristics of AR [26]. The factors in the questionnaire SASSI were respected, and visibility factors were added to the AR system. Using a Likert scale of 5 points being (1 = Disagreement until 5 = Totally agree). With a total of 38 items and classified into seven factors, that analyzes the specific aspects of a system with speech recognition and AR graphs:

– A precise response of the system (RS): if the system is precise, reliable, predictable and with minor mistakes.
– Likeability (LK): if the system is useful, pleasant, friendly and easy to use.
– Cognitive Demand (DC): the level of concentration and ease use of the system.
– Annoyance (ML): if the interaction with the system is frustrating, repetitive or irritating.
– Habitability (HB): the certain that user has for the actions to be performed.
– Velocity(VZ): the system responds quickly or slowly.
– Visibility (VS): the images are useful, understandable and appropriate.

The results are shown of the questionnaire below: The participants showed a high level of likeability (M =4.05, SD =0.99), with discreet feeling of frustration of the usage (M = 2.37, SD =0.39). Also, the users required minor effort when using the system (Cognitive Demand: M =3.46, SD =0.46). Although, the participants provided positive comments regarding how easy was the usage of the interaction of speech. It is necessary to increase the dialogues to use because in the RS (Precise response of the system: M = 3.51, SD = 0.48) shows signs that the users moderately agree of the precision with the system. That confirms a control variable without particular problems with the answers and its commands (Velocity: M = 2.86, SD = 0.24). The system behavior seemed to coincide with the conceptual model of the user (Skills: M = 2.96, SD = 0.17). Regarding the AR, the users mentioned the system is useful and appropriate (Visibility: M = 4.68, SD = 0.43). See Fig.4.

In the assessment instrument validation and Cronbach's alpha is obtained with 0.807, for each questionnaire with kurtosis centered in zero and no major of one.



	RS	LK	DC	ML	VZ	HB	VS
MEAN	3,5079	4,0476	3,457	2,3714	2,857	2,964	4,678
S.D	0,4837	0,99351	0,4577	0,39036	0,24398	0,17252	0,42608

Fig. 4. Survey results.

The results of the assessment are showing essential results regarding the usage of AR and speech systems, as a potential benefit for improving the interaction in the systems using sensors in smart cities. The reactions of the participants were very positive, and they enjoyed the interaction with the system. Still, it is necessary to increase the dialogues, when doing a query there are different ways to do it, and as more vocabulary is in the systems, the system used in the plant could have a better understanding with the user.

6 Discussion

The AR applications in mobile devices frequently are handled through tactile controls and voice commands. However, if the possibility of monitoring the sensors in real time and the interaction with a conversational agent is added, it opens a possibility to access information in a faster and efficient way to the users to interact and know more about the results and their parameters. Also, combining the speech services with other forms of interaction to design multi-modal UI. Therefore, the researchers are experimenting with different methods, frequently combining some of them in mixed interfaces and the execution of exhaustive usability to determinate the application in relevant scenarios. The multi-modal interfaces have shown the viability in the tasks that support complex interactions [30].

As future research activities could be considered the interaction user talks combined with an AR management system integrated. Demonstrating an opportunity to take advantage of the IT usage as support of daily living activities. Also, it will interesting involve other groups of potential users that can take advantage of this interaction just using voice systems (such as blind people, or with any physical disability, like arthritis, motor paralysis).

On the other hand, the proposal of interaction here presented is showing a possible way about the form in which it could work: AR manuals used for installation, assembling or fixing a product, with the possibility of asking Frequently Asked Questions (FAQ) to IT assistance in an immediate respond (Being this a conversational agent). Also, it is possible to implement assigned tasks for personal training, plant care and education as learning objects.

7 Conclusion and Future Work

The architecture and usability testing had provided favorable results about the viability to connect the AR speech recognition, here proposed. These technologies allow a significant increase with interaction and user perception about places and objects that are around in a context of a Smart Building giving the sensation that the plants and services are living beings when asking questions and receiving a response about their condition and activity of the building. Since the nature of the virtual component is not about a small component based on an image and audio but multimedia enriched with data of sensors in an animated way. Therefore, this represents a user advantage for the interaction and the easiest way to access the information of the people to be interviewed. New forms of information are provided, not just for the Smart buildings environments but also to deliver tutorials, training, educational resources, install manuals, without the necessity of using the augmented reality, with an essential impact for business and organizations innovation.

It is worth mentioning that the habitability and annoyance results done in usability test depends on some responses programmed in IBM Watson Conversation. In a future through the usability testing, an extensive content adjusted to the requirements will be programmed. Also, exists three areas of opportunity to continue developing this work inside the innovation ecosystems in smart cities:

1) The validation of evaluation instruments for the services of this type (since it does not count as this nature)
2) The implementation of this technology in an urban garden for the care of plants.
3) The incursion in the education with this type of technology to promote the care of the environment, through the care of plants and interaction with students.

References

1. Magal-Royo, T., Laborda, J.G.: Multimodal interactivity in foreign language testing. In: Multimodal Interaction with W3C Standards. pp. 351-365. Springer, Cham. (2017)
2. Magal-Royo, T., Laborda, J.G.: Multimodal interactivity in foreign language testing. In Multimodal Interaction with W3C Standards, pp. 351-365. Springer, Cham. (2017)
3. Matthew, H.G., et al.: Augmented Reality Technology Using Microsoft HoloLens in Anatomic Pathology. Archives of pathology & laboratory medicine 142(5), 638-644 (2018)
4. Cerón, C., Archundia, E., Carcés, A., Beltrán, B., Migliolo, J.: Diseño de escenarios de aprendizaje con interfaces naturales y realidad aumentada para apoyar la inclusión de estudiantes con discapacidad auditiva en la educación media superior. Research in Computing Science 144, pp. 191–201 (2017)

5. Ontiveros-Hernandez, N.J., Perez-Ramirez, M., Hernandez, Y.: Virtual Reality and Affective Computing for Improving Learning. Research in Computing Science 65, pp. 121–131 (2013)
6. Ingemar, K.et al.: Combining augmented reality and simulation-based optimization for decision support in manufacturing. In: Simulation Conference (WSC), 2017 Winter. IEEE, pp. 3988–3999 (2017)
7. Ghada, M., Essam; R., Khaled, H.: Augmented reality vehicle system: Left-turn maneuver study. Transportation research part C: emerging technologies 21(1), 1-16 (2012)
8. Conklin, J.A., Hammond, S.R.: Building integrated photovoltaic devices as smart sensors for intelligent building energy management systems. U.S. Patent No 9,772,260 (2017)
9. Hancke, G.P., et al.: The role of advanced sensing in smart cities. Sensors 13(1), 393–425 (2012)
10. Ciribini, A.L.C., et al.: Tracking users' behaviors through real-time information in BIMs: Workflow for interconnection in the Brescia Smart Campus Demonstrator. Procedia engineering (2017)
11. Rinaldi, S., Depari, A., Flammini, A., Vezzoli, A.: Integrating remote sensors in a smartphone: The project "sensors for ANDROID in embedded systems". In: 2016 IEEE Sensors Applications Symposium (SAS), Catania, Italy, April 20-22, pp. 468–473, ISBN 978-1- 4799-7249-4, DOI 10.1109/SAS.2016.7479892 (2016)
12. Sianaki, O.A., Hussain, O., Dillon, T., Tabesh, A.R.: Intelligent Decision Support System for Including Consumers' Preferences in Residential Energy Consumption in Smart Grid. In: Proc. of International Conference on Computational Intelligence, Modelling and Simulation (CIMSiM), 28-30 Sept., Bali, pp. 154–159 (2010)
13. Rinaldi, S., et al.: Bi-directional interactions between users and cognitive buildings by means of a smartphone app. In: Smart Cities Conference (ISC2), 2016 IEEE International. IEEE, pp. 1–6. (2016)
14. Azuma, R.T.: A survey of augmented reality. Presence: Teleoperators & Virtual Environments 6(4), 355–385 (1997)
15. Graham, M., Zook, M., Boulton, A.: Augmented reality in urban places: contested content and the duplicity of code. Transactions of the Institute of British Geographers 38(3), 464-479 (2013)
16. De la Cruz-Martinez, G., Eslava-Cervantes, A.L., Castañeda-Martínez, R.: Diseño de la Experiencia del Usuario para Espacios Interactivos de Aprendizaje no Formal. Research in Computing Science 89, pp. 53–62 (2015)
17. Heinroth, T., Minker, W.: "Next Generation Intelligent Environments: Ambient Adaptative Systems" Springer Press (2011)
18. Diamantaki, K.: Evaluating the user experience of a mobile user in a smart city context. January https://doi.org/10.1504/IJIEI.2015.069902 (2015)
19. International Standards Organization: ISO 9241-11:2186(en), Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts. https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en
20. Mahrin, M.N., Strooper, P., Carrington, D.: Selecting usability assessment methods for software process descriptions. In: Proceedings of the Asia-Pacific Software Engineering Conference, pp. 523–529, Washington, DC, USA (2009)
21. Crowell, C.: Analysis of Interaction Design and Assessment Methods in Full-Body Interaction for Special Needs. In: 23rd International Conference on Intelligent User Interfaces. ACM, pp. 673–674 (2018)
22. Gutiérrez, M., Thalmann, D., Vexo, F.: Semantic virtual environments with adaptive multimodal interfaces. In: Multimedia Modelling Conference, 2005. MMM 2005. Proceedings of the 11th International. IEEE, pp. 277–283 (2005)
23. Cavallo, F., Aquilano, M., Arvati, M.: An ambient assisted living approach in designing domiciliary services combined with innovative technologies for patients with Alzheimer's

disease: a case study. American Journal of Alzheimer's Disease & Other Dementias 30(1), 69–77 (2015)

24. Feder-Levy, E., Blumenfeld-Liebertal, E., Portugali, J.: The well-informed city: A decentralized, bottom-up model for a smart city service using information and self-organization. In: Smart Cities Conference (ISC2), 2016 IEEE International, pp. 1–4 (2016)
25. El murabet, A., et al.: Ambient Assisted living system's models and architectures: A survey of the state of the art. Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.(2018)
26. Nakagawa, E.Y., Oquendo, F., Becker, M.: RAModel: A Reference Model for reference Architectures. https://doi.org/10.1109/WICSA-ECSA.212.4, pp. 297–301 (2012)
27. Rinaldi, S., et al.: Bi-directional interactions between users and cognitive buildings by means of smartphone app. In: Smart Cities Conference (ISC2), 2016 IEEE International, pp. 1–6 (2016)
28. Ranjan, R., Gan, W.S.: Natural listening over headphones in augmented reality using adaptive filtering techniques. IEEE/ACM Transactions on Audio, Speech, and Language Processing 23(11), 1988–2002 (2015)
29. Barroso Osuna, J.M., Cabero Almenara, J., García-Jiménez, F., Calle-Cardoso, F.M., Gallego-Pérez, Ó., Casado-Parada, I.: Diseño, producción, evaluación y utilización educativa de la realidad aumentada. (2017)
30. Gerard, S.N., Megerian, M.G.: Extracting semantic relationships from table structures in electronic documents. U.S. Patent No. 8,914,419. vol. 180, pp. 1484–1494 (2014)

ISSN 1870-4069

Leisure Organization Models of Young People in the North Mexican Border

Aida Yarira Reyes Escalante, Carlos Alberto Ochoa Ortiz Zezzati,
Edith Vera Bustillos, Alejandra Peña Escarcega

Universidad Autónoma de Ciudad Juárez (UACJ),
Ciudad Juárez, Chihuahua, Mexico
aida.reyes@uacj.mx

Abstract. Throughout an exploratory research design, we approach the study of youth organization in fulfilling leisure as an essential part of everyday life. We use data mining as the technique to drawing conclusions from our vast amount of data, upon extrapolating patterns and associating dependently linked variables and attributes. The structure of this paper is organized by the perspective of youth as a category of study, and conceptualization of human needs to understand the implications of these at the spare time. The organization is approached as a group of people related to systematic arrangement of social relations that give meaning to ideas, and values shared. The analysis is accomplished through data obtained from 300 participants aged 15 to 29, that shed light about categories of youth organization as an articulator of the social coexistence: demographics; preferences by age, gender, type, and frequencies of activities; spending; networks and the before and after-party activities, which counts the strategies they adopt for leisure, in the midst of a complex society in Ciudad Juárez, México. The leisure organizational model is constructed by elements related to various attributes into discernable categories, which can use to draw further conclusions about the theme approached.

Keywords: youth organization; youth and leisure; spare time; data mining; social construction.

1 Human Needs in Spare Time and Leisure

From a positivist perspective, it is considered that human needs have not been a central theme for the economic development theory. While Marx only addressed the mainstream in the economy of the twentieth century, human needs were supported in the anthropology of behavioral roots that offered utilitarianism and marginality, but only addressing consumer preferences [1]. The theme of human needs became an exclusive category of studies on poverty from an economistic point of view [2]. Thus, is important to consider the different disciplinary perspectives on human needs otherwise. In doing so, there is an important chronology of the definition of needs from different disciplines [3], which reveals a gap between a desired state of the person into the motivational state that arises of varying intensity, and a subjective need that can be derived from an action to correct determined situation: From Psychology, the need is the linked condition to the experience or deficiency associated with the effort aimed at suppressing the absence

pp. 165–183; rec. 2018-09-11; acc. 2018-10-07

or scarcity, whose scope and complexity can be variable. While for the field of Education, human needs are link to the planning and educational research: improving, changing, preventing or solving problems of schooling. Regarding the scope of social policies, the needs are associated to the development of the welfare state as a criterion for access to the different dimensions of social protection, public health, and sustainability.

From all the different disciplinary perspectives mentioned above, we highlight the contributions from Psychology to the understanding of human needs, its classification, motivation, and impulses in the satisfaction of these. On the one hand, Maslow [4] exhibited his work in the theory of human motivation, formulating a hierarchy and arguing that as the basic needs (physiological needs) are met, people develop higher needs and desires. On the other hand, McDougall (1971-1938) was one of the first proponents about that intentionality and goal-seeking is what characterized more to human behavior and these depended on different components: cognitive, conative, and affective [6].

The influence of Psychology led globally in the early fifties. Particularly, in the United States was developed important studies of human motivation. For instance, Atkinson Van Nostrand presented a method for evaluating the human reasons (mobile) by analyzing the content of thematic stories and samples of imaginative thinking; concluding that meeting human needs was reduced to the urgency of accumulation, owning goods and services even regardless of its usefulness, emphasizing that the only real need is money [5].

Nevertheless, Doyal and Gough [6], argue that human needs are socially constructed, but also universal. They rejected the aspirations from particular preferences of individuals, instead, their cultural environment can be considered as a need, suggesting a distinction between basic needs and intermediate needs. While Max-Neef, Elizalde and Hopenhayn [7], defined human needs as an interrelated, and interactive system with no hierarchies or priorities between them, but recognizing the presence of a threshold below which the feeling of severe deprivation is recorded.

Related to one of the most important approaches in this study, we considered the perspective from Carosio [8], who states that the expansion and acceleration of consumption as the twentieth-century phenomenon, is part of the behavior to meet human needs that serves and articulates the relations of social coexistence of the consumer society. From a more technical perspective, Locke and Latham [9], incorporate the concept of motivation for studying human needs, choices, and preferences, referring to the existence of both internal factors driving the action, and external factors that can operate as a stimulus, influencing: direction (choice), intensity (effort), and duration (persistence).

2 Leisure: Between Spare Time, Recreation, Amusement or Entertainment

According to the Oxford Dictionary [10] "leisure" is the time when one is not working or occupied; free time. That implies the cessation of occupations for carrying out a recreational activity that becomes an essential part of everyday life. However, it is usually assumed that spare time or leisure is waste of time for not performing a productive task. Historically, leisure was associated with the upper classes of society, while the other population had no chance on enjoyment because they had to work for livelihood. Based on the perspective from Cicero about "leisure" (*otium*), in his speech, *Pro Sestio*, points out its relevance such as "a time to rest the body and recreation of the spirit necessary to tackle the task." In the Roman society, especially in big cities, there was a large focus on entertainment [11].

Throughout the International Worker´s Association (ATI), the fight for the eight hours of labor was declared in the Congress in Geneva, August 1866, demanding that the legal limitation of the working days was a dignified precondition for the improvement and the same emancipation of the working class [12]. Subsequently, in developed societies of the mid-twentieth century, leisure was occupying an important stance in the everyday life of people since the increasing of life expectancy, quality of living conditions, educational level, among other elements. Thus, recreation and leisure previously reserved for a minority become more accepted in large social groups [13].

In addition, we identified other useful contributions analyzing the concept of leisure, such as those from Bock [14], who argue that within the theory of learning and spare time, presuppose decisions and behaviors from people, as a result of their previous experiences by successful repetition of pleasant stimuli, as those who are critical in determining the consumer or buying behavior. Moreover, Nuviala, Ruíz and Garcia [15], define leisure as "… period of time not subject to needs or obligations" (p.13). Meaning that the period of time can be occupied as well in the practice of any kind of activity in the school, home, work or family, which implies not only to have a spare time full of activities, but feeling free of doing or enjoying such activities. Then, leisure can be defined in simple words: the length of time used for activities in which the person has a time of enjoyment and satisfaction. Hence, an overlapping between the concepts of spare time and leisure arises.

Likewise, Munne and Codina [16], refers to spare time as the period of time to do some leisure activity, including the space remaining to do something that person can practice for pleasure after concluding obligatory activities, either fun or tedious. Thus, spare time is simply the length of time dedicated to activities that are made out by simple pleasure and are made without obligation. Similarly, Rodriguez and Agulló [17], argue that leisure is performed in spare time made up for activities and practices that are freely chosen according to the preferences of each individual to meet personal needs, which purpose is relaxation, entertainment, and enhancing of human development: physical, emotional, and cognitive.

In conclusion, leisure also articulates the social life because individuals can build relationships, and interests for enjoyment of recreation [8]. Thus, we recall in this study the relevance of youth organization for leisure as a social construction. In the following section, youth is described as a category of analysis in relation to their context.

3 Leisure Venues

Leisure activities have been classified into three major categories: type, nature, and structure. There are different types of entertainment, whether by water, land, or air, and can be performed individually or by group. The places where recreational activities can be carried out, they tend to be private, public, monitored, and controlled. Some entertainment or recreational activities usually depend on the geography, climate, resources, and culture that can be classified as follows [17]: Outdoor recreational activities, which can be on land, water, and air; activities for entertainment: sports, artistic, and cultural; cultural activities: Artistic activities, arts and crafts, symposia, lectures, oratories, and debates; events or shows: visits to artistic, cultural, historical, social, and sports shows; family activities: birthday parties, XV years, weddings, christenings, family gatherings, meetings, and conventions; games: board games, video games, and all kind of electronic games; specialized activities: communication, motorized, mental, and multiple activities for sensory disabilities; others: visual activities, hobbies, reading, and relaxation.

4 Youth as a Social Group

Youth as a social phenomenon emerged differentiated in Western society from the eighteenth and nineteenth centuries, a historical product that comes from the bourgeois revolutions, and development of capitalism [18]. Hence, youth as social group did not exist before. In a broader perspective, youth is considered as a stage of life mediating between biological and social maturity, and become identified as a social layer that should enjoy of certain privileges but from a period of permissiveness [19].

Contemporary contributions of literature have categorized the stages of the human life into four general stages: childhood, youth, adulthood, and elderly. At the stage of youth, it has several physical, psychological, and social changes that could involve major problems related to identity, confusion, among others issues. Although it is recognized that there are ambiguities in determining the juvenile stage, as defined by the United Nations, UN, youth involves those aged between 15 and 24 years old. While the United Nations Educational, Scientific and Cultural Organization, UNESCO (2016) [20], declares that youth is a heterogeneous group in constant evolution, and the experience of being young varies widely among countries, regions, and localities. However, for other institutions, the juvenile stage involves a combination of the biological and relative social maturity, from the adolescence throughout the independence from family, the autonomy that would define the adult status [21].

In general, youth is spent in the realm of the family. Leaving the family' home and gaining economic independence, make basic milestones for an autonomy that increases thru the traditional establishment with a stable partner, and procreation of the first child. Thus, youth is a stage in which usually responsibilities are more prominent and important than those in childhood, with more complex challenges as being prepared to reach the adult stage. Although, for Masjoan, Planas & Casal [22], there are only peak moments in the transition process of the human being and these may include: the transition from school to find their first job, vocational integration, emotional dependence, relationships, building a family, procreation, and freedom in the use of spare time.

Moreover, youth is a stage that has a number of pressures: meeting family expectations, school accountability, and friends, sentimental experiences having a partner, and all type of norms or beliefs that media and peer group imposes to them [23]. Besides, youth is a stage in the life that represents social status. Being young also implies: a marketing theme, brand, and number of traits that society imposes. The society infringes on the youth certain features that are perfectly market it such as: fashion, esthetic, music, technological, among other elements.

The essence of being young brings benefits: the energy that this stage has, and new experiences that can be delved, are being sold [24]. The prestige and the very meaning of youth brought to market involve signs of expression for the high price and popularity, called as the “juvenilization”, focused on the esthetic of body and the attempt to look younger by incorporating the appearance that characterizes youth models. Several youth signs and elements have been popularized by the mass media and social networks, which recurrently manifest the efforts to achieve legitimacy and valorization through the body. Likewise, Marguilis and Urresti [21], youthfulness can be acquired as cultural imitation and is offered as a service on the market.

Once we have described in this section youth as a category of analysis, we will discuss the elements of leisure. Through leisure, people used to be freed from all of their obligations; therefore, leisure is viewed as a liberator. Young men and women transcend from a stage where they want to experience things for different reasons despite they are not yet ready for it, and because do not have the total freedom required, as they are still dependent on their parents or guardians. This may be related to some limitations youth deal with, because they do not have the resources, the freedom to take their own decisions, or they need the parental consent. In addition, because of the lack of money they may not be able purchase of goods and services of recreation such as event tickets, entertainment venues, and drinks, among others.

Despite the multiple limitations that young people deal with, they intent to fulfill their goals, forge their identity, and socialize in the everyday life. It is to be considered a fortune to the Millennials, a person reaching young adulthood in the early 21st century, that technology in recent decades has taken a strong impact on human beings. For instance, the increasing sociability of young people through the Internet, their social networks, and consumer practices, are the cornerstone for communication in the modern era, despite the loss of personal contact "…the online sociability does not represent a shift in other forms of belonging, but can extend the traditional circuits of meeting and socializing." [25]. Social networking is considered to be updated, fashionable, and encourages young people to seek prompt information. More and more, youth become informed consumers, learning about the market of recreation. However, as well as accessing to different sources of information can provide important benefits, they also deal with certain risks because security issues are involved.

Furthermore, studies about tourism provide data about how young people could be facing different problems or dangers carrying out their recreational activities, which relates the specific context of violence in Ciudad Juárez, because of the culture of illegality and anarchy of the criminal organizations. In this respect, a study about the recreational use of drugs in Mexico, is linked to the evolution and trends of youth leisure associated with a high consumption of alcohol, illegal drugs, and vandalizing; resulting in the emergence of social phenomena known as "the big bottle", and the proliferation of consumption of synthetic drugs [17].

5 Data Mining Applications in Social Aspects

Data Mining, is the extraction of hiding and predictable information inside great data bases, is a powerful new technology with great potential to help to the companies or organizations to focus on the most important information in their Bases of Information (Data Warehouse). Sumathi and Sivanandam [26] indicated the aim of data mining is to extract implicit, previously unknown and potentially useful (or actionable) patterns from data. Data mining consists of many up-to-date techniques such as classification (decision trees, natıve Bayes classifier, k-nearest neighbor and neural networks), clustering (k-means, hierarchical clustering and density-based clustering), association (one-dimensional, multidimensional, multilevel association, constraint-based association). Many years of practice show that data mining is a process, and its successful application requires data preprocessing (dimensionality reduction, cleaning, noise/outlier removal), post processing (understandability, summary and presentation), good understanding of problem domains and domain expertise.

Data mining tools predict future tendencies and behaviors, allowing businesses to make proactive decisions leaded by knowledge-driven information. The automated prospective analyses offered by a product thus go beyond past events provided by retrospective typical tools of decision support systems. Data Mining tools can respond to questions of businesses that traditionally consume too much time to be solved and to which the users of this information almost are not willing to accept. These tools explore the data bases searching for hidden patterns, finding predictable information that sometimes an expert cannot find because this is outside expectations [27].

One of the most transcendental aspects of the use of Data mining is denominated Social Data Mining, which tries to find different patterns in predefined clusters in the network, like the groups of discussion, Use nets, thematic chats among others. Other work has been focused on extracting information about online conversations such as the USENET PHOAKS [28] mining messages in the USENET newsgroup that recommend Web sites. Categorizing the users mentions to create lists of popular Web sites for each group. Where? [29]. It has been analyzed the newsgroup information and the Usenet conversations and if they have been used to create visualizations of the conversations. These visualizations can be used to find conversations with the desirable characteristics, such as equality of participation or regular participants. In Fiore [30], also was extracted information of newsgroups and visualizations of the conversation subject, contributions of individual messages, and the relation among them were designed. Another research has been centered in extracting the information of web user records.

The Log files [31] register information of the users, analyze this to find common connections between Web pages, and they construct diverse visualizations of these data to help user navigation through Web sites. Persecuting the navigation metaphor, some investigators have used the term "social navigation" in order to characterize the work of this nature [32]. Finally, a different technical approach [33] uses the register of activity - e.g., a sequence of visited URLs during a session like the basic unit. Based on this, they have developed techniques to calculate similarities between the trajectories of sequences and to make recommendations – for example, to similar pages to the visited ones.

5.1 Social Data Mining

The motivation to make an approach by means of applications with Data Mining is based on previous works of Social Data Mining in this research area [34]. This research area emphasizes the role of the collective analysis of conduct effort, rather that the individual one. A social tendency results from the decisions of many individuals, joined only in the location in where they choose to coexist, yet this, still it reflects a rough notion of what the researchers of the area find of what could be a correct and valid social tendency [35]. The social tendency reflects the history of the use of a collective behavior, and serves like base to characterize the behavior of future descendants [36]. The Data Mining approaches for social aspects look for analogous situations in the behavior registers [37].

The investigators look for situations where the groups of people are producing computer registers (such as documents, USENET messages, or Web sites and links to groups with a specific profile) like part of its normal activity. The potentially useful information implicit in these files is identified; and the computer techniques to display the results are designed. Thus the computer discovers and makes explicit the "social tendencies through the time" created by a particular type of community.

The systems that analyze social aspects with Data Mining do not require expert users in no new activity, due to this, the investigators in the subject try to explore the information of the user’s preference implicit in the existing activity registers. There is a wide variety of applications of data mining and in various areas of study, for example: Reyes-Nava, Flores-Fuentes, Alejo, and Rendón-Lara made a investigations related to data mining of risk factors in students [38]; Rodríguez-Maya, Lara-Álvarez, May-Tzuc, and Suárez-Carranza write and evaluation of CENEVAL for students [39]; Gonzalez-Marron, Enciso-Gonzalez, Hernandez-Gonzalez, Gutierrez-Franco, Guizar-Barrera, and Marquez-Callejas development and modeling students' dropout in Mexican Universities [40].

6 Context of Youths in Ciudad Juárez

Youth must be legitimated as a social group integrated by individuals that shared ideas and values within a specific context. The context of young people in Ciudad Juárez, is permeated by social, economic, and cultural reality that has influenced violence and poverty. Ciudad Juárez has been classified as a city where the culture of illegality reigns, "abandoned for many years by different levels of government in the most important areas of social policy” [41]. According to Lourdes Almada [41], Ciudad Juárez and their families have experience processes of profound transformation due to three main factors:

a) An unprecedented population growth, constituting the city as a hub of immigration attraction: its population quadrupled between 1960 and 2000.

b) Massive incorporation of women and young people to work; for example, a difference of about 10% is observed between the participation rate of women aged 15 to 39 years in Ciudad Juarez, about the national average rate; and

> c) A policy based on urban growth, land speculation, and the great interests of construction companies, rather than the needs of families and the general population housing. (p. 68)

In the census of 2015 by The National Institute of Geography and Information (INEGI) [42], indicated that in Mexico the segment of the population aged between 15 and 29, is an important social force, economic, political, and cultural. Data shows 38.3 million young people whose average age is 27 years old, which 50.3% are women, and 49.7% men. The population in the state of Chihuahua is about 3‘406,465 people, but a significant 39.10% of the population is located in Ciudad Juárez, the largest city. The average age for men and women was 25 years in the municipality. There are 862,942 young people aged between 15 and 29, which 50.05% are men, and 49.95% women. The population distribution by gender is almost equals (49.97%, 50.03% men and women); therefore, there is a ratio of 99.89 men per 100 women. While at the national level were 48.83% men, and 51.17% women, implying a greater number of women in compare to Ciudad Juárez.

Nowadays, Ciudad Juárez offers different types of entertainment venues for young people. There are alternatives such as variety of conventions, sporting events, music, shopping malls, clubs, casinos, bar and restaurants, as well as, some natural areas for outdoor recreation such as the Dunes of Samalayuca, an important desert area located on the margins of the municipality, among other surrounding areas. The insecurity that citizens from Ciudad Juárez has deal with, influences the choices for entertainment, as from 2008 a wave of violence was recorded, modifying the everyday life and consequently, the ways to be organized for leisure. During the most critical years of violence, 2008-2012, there were massive closures of different types of business, which was critical for entrepreneurs. However, the current offering and demand for entertainment and recreation arise, resurging the social life in Ciudad Juárez. However, the experiences from past years in relation to public insecurity have led to changes in the ways in which young people attempt to use their spare time.

7 Methodology

This is a descriptive and exploratory study through a survey research design. The purpose of this study is to analyze the youth organization in relation to leisure, we utilize the approach of organization as a group of people that build social relations giving meaning to the shared ideas and values [14]. The random sample consists of 300 youth from Ciudad Juárez. The sample was calculated using an online program for sample calculations of finite population, where 862,942 youth and the relational age range 18-29 [42]. With a confidence level of 90% and an error of 5%, the sample is 300. The surveys were conducted in various public places such as parks, sports fields, shopping malls, bars, restaurants, and in educational institutions: high schools and universities, during 2017.

For data collection, it was design and validated a survey instrument designed Ramirez in 2017. The instrument was organized by items related to different dimensions, and indicators or categories of information, adapting five organizational elements about features that characterize the youth organization: social structure, social actors, goals, interests and choices [43]. Through the elements mentioned, help to identify demographics, planning, choices and preferences for leisure.

To ensure proper handling of data and get a better data entry, the responses from the survey were organized by utilizing in the Statistical Package for Social Sciences (SPSS version 23.0). The analysis was carried principally through descriptive statistics, frequencies and ranges of the relationship of the domains, and categories classified with based on the elements of organization adapted. The relation of the items and construction model organization are using Data Social Mining with the Software Applications of Weka 3.9.2, and Tableu 15.0.

8 Analysis

The analysis is realized with based on the classification adopted in order to identify the relational domains that characterize the youth organization: social structure, social actors, goals, interests and choices [43]. The domains described before, were itemize by categories that reveal youth organization on carrying out their leisure activities: demographics and family structure; choices and entertainment preferences by age, gender, and by type and frequencies of activities; the before and after-party activities, the planning process, spending, and companionship. Other categories that emerged are render, the organizational network, and the social construction of youth and leisure in a vulnerable city.

a) Youth and demographics. The demographics of youth participants in the study show results by gender, 55.33% female, and 44.67% male. Regarding the sexual preference of young people, 90.33% heterosexual, 5.67% homosexual, and 3.67% bisexual. The ranges of youth aged 15-20, is 45.67%, followed by the range of 21-25 with 43%, and 26-29 with 11.33%. The level of education of youth is related to the age grouping of participants: 70.67% of respondents have a degree from a university, the 16.33% completed high school, and 5.67% have completed middle school. The remaining, 4.67% have a technician career, and 1.33% with primary or incomplete. Also, the predominant civil status is single, with 87.67%, then, 4.33% is married, and 7% in free union. Only 0.67% is divorced, and widowed 0.33%. (See Figure 1). The family structure of youth is relevant because it mirrors the social structure of Ciudad Juárez. Regarding the structure of family of the young people, these have mostly 64% a traditional family i.e., in this type of family it is present the father, mother, and children. Followed by 16.67% of youth with a single parent family, mostly represented by mother and children. A minority of respondents, 9.67%, has their own family, living with their partner since the marital status. The origin and time of residence are relevant in the discussion of immigration and identity. More than two thirds of young people were born in Ciudad Juárez, a significant 78.26%. While the origin of the remaining of respondents, 21.74%, were from Veracruz, Durango, and Torreon. Regarding the residence in Ciudad Juárez, 47.33% of respondents have an average of 13 to 20 years living in the city. Followed by 39% with 21 years or more, as well as, 7.33% from 6 to 12 years, 4.67% from 2 to 5 years, and 1.67% has one year or less of residence in the city. The area of residence of youth reflects spatial information about the population. For instance, significantly, 47% is located by the south of the city (18.67% by Las Torres, 11.33% by La Cuesta, 9.33% by Camino Real, and 9% by the Airport). As well as, 19.67% is located by the new geographical down town (12% by PRONAF, and 7.67% from the area of Gomez Morin, a resurging area for leisure). In addition, 19.33% of the respondents live in the area located by Tierra Nueva (southwest).

b) Youth organization by choices, preferences, and planning for leisure. The frequency of youth leisure is related to the rate at which recreational activities are carried out, or repeated over a particular period of time in the given sample. The rate varies as follows, revealing the resurging social life of youth in Ciudad Juárez: 35.45%, once a week; 23.75%, more than twice a week; 23.06%, once a month; and 17.73%, every fortnight (2 weeks). On the one hand, young people declared being well informed about events or places for entertainment. Media and social networks play the most relevant resources used for youth about the marketing of leisure offered in the city, such as Facebook, with 70%; followed by the 15% who are informed by direct recommendations from friends; and 5%, informed by other means or people, such as family, different applications, and mass media both electronic and printed.

c) For the organizational resources that youth utilize, Facebook and Messenger are the most utilized networks for communication of planning leisure, 39.60% and 34.56% respectively. Followed by 11.74% who preferred using the groups of people through the WhatsApp, but 8.72% prefer to be organized by face to face communication, and 5.03% through both, phone calling, and texting. The choices for leisure varied by type of recreational or entertainment activities: 20% of youth agreed that special events (massive cultural and artistic events) is the option that they most prefer for leisure, followed by 18.33% of others that prefer other recreational events or activities. While others, 16.67% prefer the enjoyment of recreation in natural areas. Youth also choose outgoing for eat at restaurants, 14.67%, and others prefer to get fun at nightclubs, 10.33%. Likewise, 10.33% of youth prefer going to shopping in the city, while, 9.67% prefer going to shopping in the across neighboring town, El Paso, Texas, U.S. The choices and preferences by gender are showed in the following charts. For instance, 37% of women most likely to go to special events, instead, 27% of men have a preference for recreational events of any kind. The range of youth aged 15 to 20 have a preference for recreational events is the first place. While youth aged 21 to 25, they prefer cultural or artistic events, instead, youth aged 26 to 29 have a preference for leisure in natural areas.

d) Expectations, socializing, and networking for leisure**.** Leisure should be part of everyday life and its purpose bring benefits to the enjoyment and relaxing of mind, body, and spirit. Thus, findings from the survey conducted with youth shows that 13.00% aims to socialize through leisure activities, while 22.67% have the expectation of leisure as the purpose to get a recreational distraction for themselves. Other participants in the study stated that they have specific purposes on target leisure activities, as follows: 20.67% seeks going to dance;

18.33% for going out to eat; and 12.67% attempt to have fun by drinking alcoholic beverages. Remaining respondents stated having other purposes for leisure such relaxing, meditating, playing sports, and meeting new people. When planning to go out to have fun, the entertainment options by gender are interesting. The main objective in women is to be distracted and in the case of the most popular for men was eating. Usually, leisure serves as the bridge to build social life. Accordingly, young people attempt to carry out their leisure activities with a companionship by friends, peers from school or work, family, and their sentimental partners. The most popular option in planning leisure is with friends, 47.33%, and significantly, 38.33% makes plans to going out with their sentimental partners. The reminder is represented by 11.67%, who makes plan for leisure with family, while 1.33% with schoolmates and 1.34% make plans by individual choice.

e) Consumption expenditure in leisure**.** How much youth does is available to invest for leisure? That is an interesting question. To carry out different leisure activities, young people are available to make different expenses about: transportation, clothing, accessories, food, alcohol drinks, and ticket covers for events or entrance to specific nightclubs. Youth influences styles and design of fashion and mode. There are important expenses in which youth incurred about wearing specific styles for feeling great in the looks they love. The style of dressing fashion it depends on the kind of the leisure activity they choose but youth mostly prefer "Casual", 83.95%. Others prefer to be dress up in a more formal style, 8.03%. While some of the youth prefer wearing a rocker style, with 3.34%. Also, 3.01% wear other styles: sporty and thematic.

In relation to the expenditures related to transportation utilized to attend the leisure activities, 44.67% of youth use their own vehicle, followed by taking an "Uber", 37.67%. Also, the 6.67% uses public transportation, while only 5% used to get a ride. While 4.33% of youths is transported by the members of their family, only 1.33% venture to walk 0.33%, and others use taxi services. Regarding the level of consumption in carrying out leisure activities, 69.33% of youth often spend an average of $ 200.00 to $ 500.00 Mexican pesos; followed by 20.33% in spending $ 501.00 to 700.00 pesos. While in a higher level of consumption, there is 8.67% of youth spending between $ 701.00 to $ 1000.00 pesos, and 1.67% only spends about $ 1001.00 to $ 1500.00 pesos in leisure. The category that represents the higher consumption is "food", 48.33%; then, the "Alcohol drinks", 32.67%. Followed by fashion and clothing, 7.67%, ticket covers with 7.00%, and other expenses.

The relationship between the variables allows us to observe how the young people in any amount, always have their main expenses related to food and alcohol drinks, likewise, the lowest amounts (200-500 pesos) are related to own car transports and the use of uber, see figures 3 and 4.

f) Leisure and its context: security strategies and responsibilities. The context of uncertainty because the lack of public safety in Ciudad Juárez, has influenced the social life of citizens, encouraging to young people and their families to organize themselves in order to take care of their physical and emotional integrity. In addition, young people deal with fear since they have had experiences related to the violence in the city. They declared having fears to the following:

Assaults, is the greatest fear of young people, 33.67%. While others fears having a car crash, 31.67%, abuse of police authority from police. Therefore, parent supervision and monitoring become a reinforced strategy as one of the security strategy for youth. In this regard, 45% of youth responded that usually they have to request permission to their parent or guardians for leisure, but 23.33% only ask for permission occasionally. Regarding the request for permission to attend "After-party", 42.86% of youths do it, but 4.84% said they do it occasionally. The fears by gender resulted to be almost similar between men and women. The designated driver is the primary responsibility among young people, especially at night, 63.67%. As well as, 21.33% of youth responded that they also assign a person who is responsible for taking care or their personal belongings, such as purses and money. While 12.33% use to assign a person to be in charge of communication through the group and relatives during the activity, and 2.67%, assigns someone to handle with health contingencies that youth could deal with, such as medications and further instructions for health care. Young people use to complement or extend their leisure experiences, by doing before-after parties' activities, for instance. However, two-thirds of respondents, 64.67%, do not participate on the before and after leisure. Instead, only 15.33% use to carried out the before-after party, and 20.00% do it occasionally, and for these: 52.68% often go to the "After-Party" (32.21% occasionally, and 20.47% regularly), and of these, 62.11% responded that they are normally accompanied by friends. The after party is generally carried out in places like private homes, 77.50%. The set-up of different types of rules or requirements by the group of youth participating in the after-party implies that: 31.08% of youth stated, "Each person carries their consumption". Also, more than half of youth state that choosing the location for the after-party is taken with the utmost caution, 53%. The youth responded that usually are accompanied by friends to return to their homes after leisure, 39.13%. While 26.09% of youth is accompanied to return home by their partners, 22.36% use to return to their homes by themselves, and 9.94% use to be accompanied by their families.

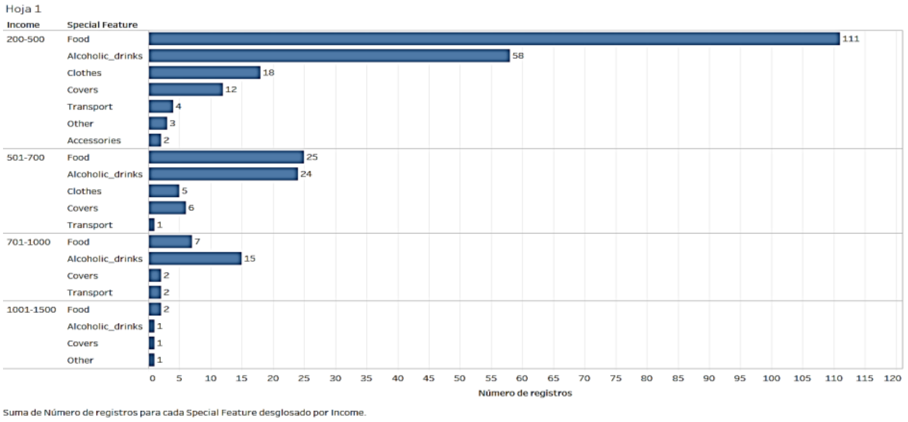


Fig. 3. Income and Special Feature.

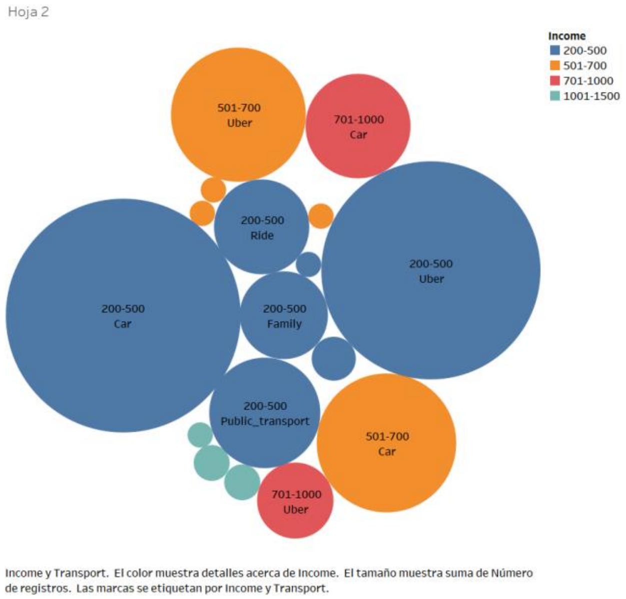


Fig. 4. Consumption expenditure in leisure.

9 Construction Model Elements

Model organization elements are determinate by the principal element related to Social Media on the Facebook are the most important tools to planning to go fun is related to communication with friends and family, the planning process make the construction of objectives a than the clothes, transport, spend and special feature, finally the roles and precautions, see figure 6 and the construction tree relationship items in the figure 5.

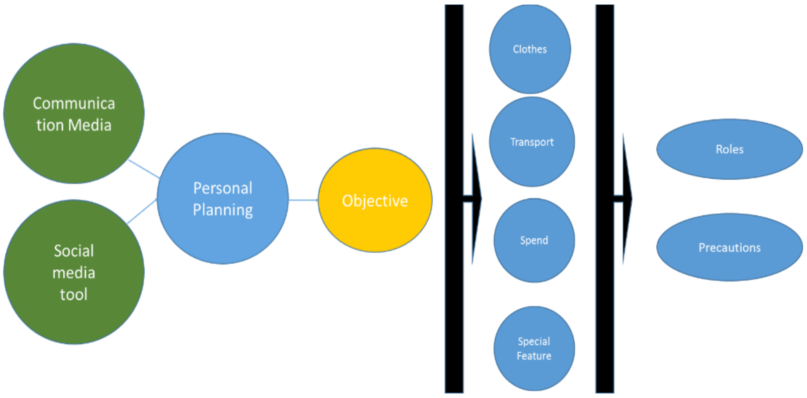


Fig. 5. Leisure Model Organizations of youth people in Ciudad Juarez.

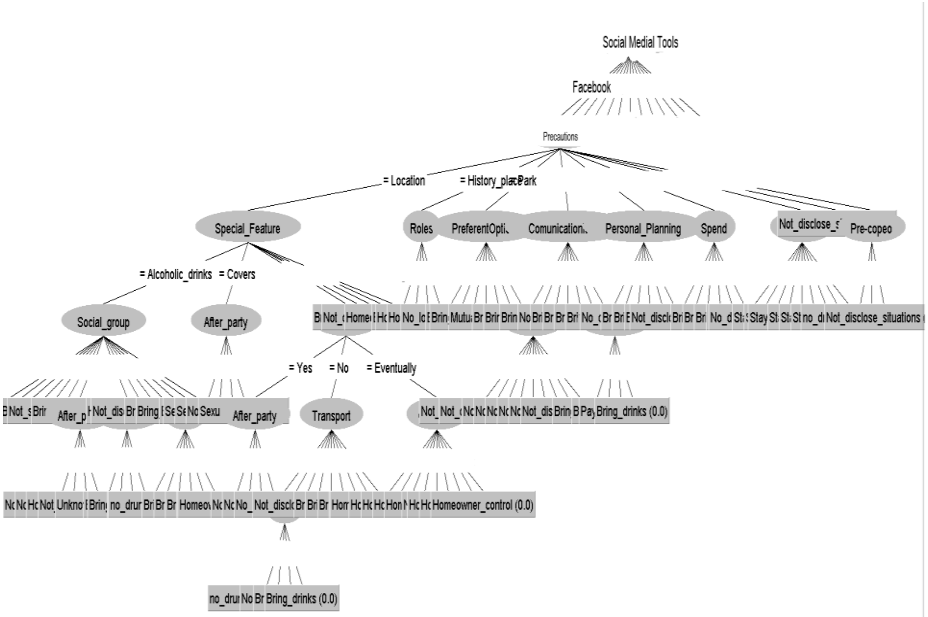


Fig. 6. Construction relation variables tree.

10 Summary of Conclusions

The results allow render important accounts about youth organization in fulfilling leisure. In addition to the descriptive statistics used to analyze the results from the survey conducted with youth of Ciudad Juárez, other categories emerged from the additional comments they did.

As a result of the experiences from the violence infringed in the city, young people deal with important decisions to take. For instance, beside the youth dependence from family they are tied to, accounts confirm how the adoption of different strategies is a constant in the everyday life of youth and their families. For instance, communication and companionship in leisure can be an issue of human security. In that respect, youth have to notify family or acquaintances about the places they go, with who they socialize, and the schedules.

In this regard, a female respondent indicates: "... When I go out, the most important thing I have to take care of, is that I should carry my cellular and it should be charge the battery, second, I have to inform my mother where I go". Likewise, a male youth respondent states, "I am afraid to loose communication when my battery is dead". As well as, other male respondent states this "... carrying money is important, but just the enough, in case I get mugged, I do not lose that much money"; while other young women stated that, "What it worries me the most, is to not be alone somewhere". Other male youth stated, "I'm afraid to lose my friends in the club"; also, a male youth states, "I fear that my car might break down when going out". Thus, youth want to keep company always in order to feel they are secure during leisure: before and after, and returning home.

Moreover, youth manifested their experiences during leisure: "Kidnappers hit me, and now I have scars". While others refer to their experiences regarding the recreational service received, "There are long lines at the clubs", "Waiters charge you more if you do not notice it, and you wait a lot for them to served you". Also, other youth participants in this study perceive that there are all types of imminent dangers that can bring more problems and risks during leisure: "The police are bullies"; "There are occasions when I spend a lot of money, especially when I get drunk "; "Taking care of my drunk friends is like babysitting".

The recurrent category of analysis emerges related to experiencing fear or dread of young people. Fear in humans is considered as a response to a natural instinct. Fear is one of the most basic emotions of humans, and is associated with anxiety, anguish, and panic. Fear can influence individuals and communities, even in complex societies [44]. Jean-Paul Sartre declared that "all men are afraid, all of them. Men who is not afraid is not normal, and that has nothing to do with courage" [45].

Although the decision-making about security strategies that people have to adopt, are presented on daily basis in large cities with high levels of insecurity, the United Nations Development Program, PNUD (2015) [46], has declared the basic conditions for achieving human security: Eliminating fear in the population and eliminate its shortcomings, since feelings of insecurity focuses more on the concerns of everyday life. In addition, PNUD states that human security should be people centered, and the life in conflict or in peace. Hence, it concerns about how people feel within society, the freedom that they have to exercise, the access to market, and to social opportunities. Human security means that people can exercise these choices safely and freely, having relatively confidence that the opportunities they have today, will not disappear completely tomorrow.

However, the options for social opportunities and freedom in the case of Ciudad Juárez, has been an issue, as this statement confirms by female members of the Civil Society Organizations (CSOs), "We don’t even have the right to walk in our streets" [47]. By 2010, a young woman stated, "going out to have fun in Juárez, is like practicing an extreme sport; there are no conditions for young women to be on the streets for going to a bar or nightclub to have fun "(Idem). While the Round Table of Women of Ciudad Juárez has declared that, "... being young and being a woman in this city, put us in a double risk, there is no safe place for young females, since shootings have also happened near to universities"[48]. Moreover, parents have complained to the Mexican state about the disappearance of young women and the fear of families to live in such insecurity, "…seeing our daughters going out, without knowing if they will return, is an issue that affects to everyone, but no one will stop these" (Idem., 2018). In this regard, the civil association founded by relatives and friends of missing and murdered young women in the State of Chihuahua, "Nuestras hijas de regreso a casa" (Our daughters returning home), argued this "Because of the murders of young and poor women in Ciudad Juárez, which are documented since 1993, youth live in fear of becoming victims of femicide", stating that, "We can no longer go out and have fun, but we can be united to demand justice "(Idem.).

However, social life gradually has been recovering in the city by National Geographic-Mexico, [49]. Although CNN México [50], declared that social life, especially the nightlife, has been recovering from the years 2015 to 2018, stating that Ciudad Juárez is recognized worldwide as one of the most famous border cities for its nightlife, and where it can be perceiving a healthy environment in different places. In addition, the opening of different services has increased, such as restaurants and bars, and these located mainly in two areas that have emerged in the new geographical downtown, the area known as the PRONAF loop, and the Gómez Morin. In addition, Ciudad Juárez is considered an attraction for different investors because the strategic location as an international border, the dynamic movement of diverse capital, the purchasing power, and other important socio-economic variables.

Nonetheless, in fulfilling leisure into a context of uncertainty, a social construction surfaces throughout the youth organization as the collective imaginary pursuing the exercise of choices safely ad freely [8] & [6]. Then, the recall of youth organization approach as a group of people related to the systematic arrangement of social relations, give meaning to ideas and values. But it is clear that youth and their families need to retake the beginning of the old atmosphere of night entertainment weekends that brought the fame of Ciudad Juarez fame: from the beginning of the XX century with the alcohol prohibition, until world war II, and later recognized also through the different decades by famous artists, visitors, and its population. After all, who does not remember the song “Noa Noa”?, a song from the decade of the 70’s by the famous deceased singer, Juan Gabriel, whose lyrics describe a nightclub that mimicked life in Ciudad Juárez as a place “where everything is different”, embracing enjoyment, happiness, and a friendly atmosphere. Such a place has been immortalized through that lovely song, whose memories resemble a great nostalgia for those years.

References

1. Sen, A.: Development as freedom. Oxford: Oxford University Press (1999)
2. Ramos Gorostiza, J.L.: La cuestión de las necesidades en el pensamiento económico. ICE, Nº 181, pp. 205–220 (2004)
3. Moreno Cámara, S., Palomino Moral, P.Á., Frías Osuna, A., del Pino Casado, R.: En torno al concepto de necesidad. Index de Enfermería. 24(4), 236–9. Disponible en: http://www.index-f.com/index-enfermeria/v24n4/10017.php (2015)
4. Maslow. A.H.: A Theory of Human Motivation. Originally Published in Psychological Review 50, 370-396 (1943)
5. Elizalde, A., Martí Vilar, M., Martínez Salvá, F.: Una revisión crítica del debate sobre las necesidades humanas desde el enfoque centrado en la persona. Polis. Revista Latinoamericana (2006)
6. Doyal, L., Gough, I.: La teoría de las necesidades humanas. Editorial ICARIA, España (1994)

7. Max-Neff, M.; Elizalde, A., Hopenhayn, M.: Desarrollo a escala humana: una opción para el futuro. Development Dialogue, número especial, CEPAUR y Fundación Dag Hammarskjöld, Uppsala, 2010. Retrieved on September, from: http://habitat.aq.upm.es/deh/adeh.pdf, (2017)
8. Carosio, A.: El Género del Consumo en la Sociedad de Consumo. Revista de Estudios de Género. La ventana [en línea], III. (2008). Retrieved on September 22, 2017, from: <http://www.redalyc.org/articulo.oa?id=88411497006> ISSN 1405-9436
9. Locke, E.A., Latham, G.: What Should We Do About Motivation Theory? Six Recommendations. For The Twenty-first Century. Academy of Management Review 29(3), 388–403 (2004)
10. Oxford Dictionary (s/f). Leisure. Retrieved 14 November, 2017. www.rae.es.
11. Piquero, J.; Cabrero, F., Cordente, V.: Los oficios de la diversión en Roma. Espacio, tiempo y forma. Serie II, Historia antigua, Nº 24, pp. 363-379 (2011)
12. ATI. Estatutos de la Asociación Internacional de Trabajadores, Ginebra, 1866. (2017). Retrieved on September 13, 2017, from: https://anarquismoenpdf1.files.wordpress.com/2015/09/estatutos-de-la-asociacion-internacional-de-trabajadores-ginebra-1866-1.pdf
13. Mora, D.C.: Derecho a la diversión acervo jurídico de la UNAM. Repositorio en línea UNAM (2010)
14. Bock, P.K.: Introducción a la moderna antropología cultural. México: Fondo de cultura económica (1977)
15. Nuviala, A., Ruiz Juan, F., García Montes, M.E.: Tiempo libre, ocio y actividad física en los adolescentes. La influencia de los padres. Retos. Nuevas tendencias en Educación Física, Deporte y Recreación, 2003. Retrieved on September 17, 2017, from: file:///C:/Users/hp%20pc/AppData/Roaming/Mozilla/Firefox/Profiles/d8xc2nm2.default/zotero/storage/3RESAVXU/2282437.pdf
16. Munne, F., Codina, N.: Psicología Social del Ocio y Tiempo Libre. Madrid. (1999). Retrieved on September 17, 2017, from: https://www.researchgate.net/profile/Nuria_Codina/publication/257766145_Psicologia_Social_del_ocio_y_el_tiempo_libre/links/00b7d525d5643621e4000000/Psicologia-Social-del-ocio-y-el-tiempo-libre.pdf.
17. Rodríguez Suárez, J., Agulló Tomás, E.: Estilos de vida, cultura, ocio y tiempo libre de los estudiantes universitarios. Psicothema. (1999). 11(2). Retrieved on October 17, 2017, from: http://digibuo.uniovi.es/dspace/handle/10651/27674.
18. Balardini, Sergio. De los jóvenes, la juventud y las políticas de juventud. Última Década. (2000) Retrieved on May 16, 2017, from: http://www.uacm.kirj.redalyc.redalyc.org/articulo.oa?id=19501301
19. Margulis, M., Urresti, M. (ed.). La construcción social de la condición de juventud. En: Cubides, H.: Viviendo a toda. Jóvenes, territorios culturales y nuevas sensibilidades (1998)
20. UNESCO. Declaración de los derechos humanos (2016)
21. Margulis, M., Ariovich, L.: La juventud es más que una palabra: ensayos sobre cultura y juventud. Editorial Biblos (1996)
22. Masjoan, J.M., Planas, J., Casal, J.: Elementos para un análisis sociológico de la transición a la vida adulta. Política y Sociedad 1(97) (1988). Retrieved may 2018. https://doi.org/10.5209/POSO.31832

23. Londoño, Valencia, Sánchez: Asertividad, resistencia a la presión de grupo y consumo de alcohol en universitarios. Universidad Católica de Colombia ac-tacolombianadepsicología11 (2007) (1):155-162. Available from: https://www.researchgate.net/publication/28242875_Asertividad_resistencia_a_la_presion_de_grupo_y_consumo_de_alcohol_en_universitarios [accessed Sep 28 2018].
24. Lodoño Pérez, C., Valencia Lara, S.C., Sánchez. L., León, V.: Diseño del cuestionario resistencia a la presión de grupo en el consumo de alcohol (CRPG). Suma Psicológica, 2007. Retrieved on September 22, 2017, from: <http://www.redalyc.org/articulo.oa?id=134216871005> ISSN 0121-4381
25. Winocur, R.: Internet en la vida cotidiana de los jóvenes. Revista Mexicana de Sociología 68. Núm. 3 (julio-septiembre, 2006): 551-580. Universidad Autónoma de México. Instituto de Investigaciones Sociales. México, D.F.
26. Sumathi, S., Sivanandam, S.N.: Introduction to Data Mining and its Applications. Studies in Computational Intelligence book series (SCI, volume 29) (2006)
27. Azzalini, A., Scarpa, B.: Data Analysis and Data Mining (2014)
28. Terveen, L.: Using Frequency-of-Mention in Public Conversations for Social Filter-ing. Proceedings CSCW’96 (1996)
29. Viegas, F.: Chat circles. In: Proceedings of CHI’99, ACM Press, pp. 9–16 (1999)
30. Fiore, T.: Visualization Components for persistent Conversations. In: Proceedings of CHI’2001 (2001)
31. Wexelblat, P.: Footprints: History-Rich Tools for Information Foraging. In: Proceedings of CHI’99 (1999)
32. Munro, J., Höök, K.: Social Navigation of Information Space. Springer (1999)
33. Broedbeck, K.: The order of things: Activity-Centered Information Access. In: Proceedings 7thICWWW’98 (1998)
34. Bush, V.: As we may think. The Atlantic Monthly, July (1945)
35. Toriello, A., Hill, W.: Beyond Recommender Systems: Helping People Help Each Other. HCI in the new Millennium, Addison Wesley (2001)
36. Hé, Z., Milodragovich, K.: Discovering chinese descendents in Palé Island using Data Mining. CACCBR; Astana, Kazakhstán (2005)
37. Padméterakiris, A., Gyllenhaal, J., Ochoa, A.: Implementing of a Data Mining Algorithmn for discovering Greek ancestors, using simetry patterns. Central Asia CCBR (Data Mining Workshop); Astana, Kazakhstán (2005)
38. Reyes-Nava, A., Flores-Fuentes, A., Alejo, R., Rendón-Lara, E.: Minería de datos aplicada para la identificación de factores de riesgo en alumnos. Research Computer Science 139, pp. 177–189 (2017)
39. Rodríguez-Maya, N.E, Lara-Álvarez, C., May-Tzuc, O., Suárez-Carranza, B.A.: Modeling Students’ Dropout in Mexican Universities. Research in Computing Science 139, pp. 163–175 (2017)
40. Gonzalez-Marron, D., Enciso-Gonzalez, A., Hernandez-Gonzalez, A.K., Gutierrez-Franco, D., Guizar-Barrera, B., Marquez-Callejas, A.: Evaluación de parámetros de encuesta de ingreso del CENEVAL para alumnos candidatos a ingresar al nivel superior, caso de estudio ITP. Research in Computing Science 139, pp. 135–147 (2017)

41. Almada, L.: Diagnóstico sobre la realidad social, económica y cultural de los entornos locales para el diseño de intervenciones en materia de prevención y erradicación de la violencia en la región norte: el caso de Ciudad Juárez, Chihuahua. CONAVIM (2009)
42. INEGI. Censo de Población, 2015. Retrieved on November 14, from: www.inegi.gob.mx (2012) Censo comercial e industrial. Tipos de industrias. www.inegi.gob
43. Barba Álvarez, A.: Administración, teoría de la organización y estudios organizacionales: tres campos de conocimiento, tres identidades, 2013. Retrieved on February 10, 2017, from: http://zaloamati.azc.uam.mx/handle/11191/2600. Consulta: 10 febrero (2013)
44. Fuentes Gómez, J., Rosado Lugo, M.: La construcción social del miedo y la conformación de imaginarios urbanos maléficos. Iztapalapa, Revista de Ciencias Sociales y Humanidades, México, pp. 64–65 (2008)
45. Delemeau, J. (Ed.): El miedo: reflexiones sobre su dimensión social y cultural. Medellín, Colombia. Corporación Región (2002)
46. PNUD. Informe sobre Desarrollo Humano, 2015. Retrieved on January 10, 2018, from: http://hdr.undp.org/sites/default/files/2015_human_development_report_overview_-_es.pdf
47. CIMAC, Cimacnoticias. 2010. Retrieves 10 sep 2017. From https://www.cimac.org.mx/node/5
48. NUESTRAS HIJAS DE REGRESO A CASA, A.C., Quienes somos, 2018. Retrieved on Febrero 06, 2018, from: https://nuestrashijasderegresoacasa.blogspot.mx/p/quienes-somos.html
49. National Geographic-Mexico. The resurging of social life in Ciudad Juarez. (2017). Retrieved on January 13, 2018, from: https://www.nationalgeographic.com/.../juarez-mexico-border-city-drug-cartels-murder.
50. Wordpress. Vida nocturna en Ciudad Juárez, 2018. Retrieved on January 06, from: https://vidanocturnajuarez.wordpress.com/

ISSN 1870-4069

Modeling a Roof Garden to Buildings in a Smart City using Equation Weight to Calculate Distribution of Load Live and Weight Maximum on a Roof Top

Angel de Jesús Calam Torres[1], Alberto Ochoa-Zezzatti[2],
José Alberto Hernández Aguilar[3], Víctor Antonio Chulin Tec[1]

[1] Instituto Tecnológico de Chetumal, Chetumal Quintana Roo, Mexico
[2] Juarez City University, Chihuahua, Mexico
[3] FCAeI-Universidad Autónoma del Estado de Morelos, Cuernavaca, Mexico
angel.angel002011@gmail.com

Abstract. This research presents an intelligent model related with the modeling of a roof garden in buildings in the center of the country of Mexico and, in general, the buildings are located in Mexican national territory, taking in consideration the legislation in such delimitation. We analyze the behavior and features of a roof slab from the point of view of the constructive conception that was designed to complete the building and that does not have the specifications of a slab of mezzanine that from the beginning are established in the calculation memory of the construction. For that reason, it is important to determine the optimal parameters for the development of the roof garden and thus begin to transform the cities with the characteristics and conditions to be a smart city. The intent of the present research is to apply mathematical tools, computational as well as artificial intelligence software for roof garden modeling, based on a mathematical model that allows to integrate the dead load to live load, and the specific weights of the dry stratum and the wet stratum since the use of smart farming is incorporated in the roof top design.

Keywords: roof top, intelligent garden, buildings in a smart city, load live, smart farming.

1 Introduction

The Roof Garden are born from the idea of using spaces in cities where one of the main problems is lacking of farmyard, reduced to create a garden or at the time natural conditions for developing crops spaces, where the aesthetic and environmental benefits of gardens on rooftops and roofs have been recognized for decades. In this way, the quantification of these benefits has not been investigated deeply in the US, but in other countries such as Germany and Canada [1].

Green roofs provide a large range of benefits from amenity to ecological and technical advantages to financial aspects [2]. The California-based study by Simpson and Machpherson [3] shows that tree shades have potential to reduce annual energy use for cooling 10-50% (200-600 kWh) and peak electrical use up to 23% (0.7 kW).

At present, there is a lot of literature that describes the importance of family gardens for stress prevention, leisure and personal issues and social identity [4].

The inclusion of mathematical tools has been increasing over time but it is a relatively a new topic in which it is being given great importance in different aspects. Kumar and Kaushik (2005) performed a mathematical model to evaluate the cooling potential of garden areas on the roofs of buildings exposed to solar energy [5].

2 Structural Equation Model

The structural equation models show the dependency ratio between the variables. For example, by integrating a series of connections for the electric line for the case of the people who depend on it or in its independent case, the one that is within the same model of the variables that can be independent in the same way they can be dependent on others [6], this is how they become a useful tool [7].

The reason why the Structural Constructive Factors (FCE in Spanish) were taken is because they were considered to be the most important for the analysis of loads on the roof. Given that Ergonomic Environmental Factors (FEA in Spanish) are considered secondary factors, since they represent variations where the most significant is precipitation; this is considered in the specific wet weight of the land for garden. A wind = 0 m/s is taken into account since in conditions to evaluate loads it does not represent a significant value, nevertheless it is for trees of more than 1 meter and mainly for future metal or wood structures that are incorporated above the rooftop.

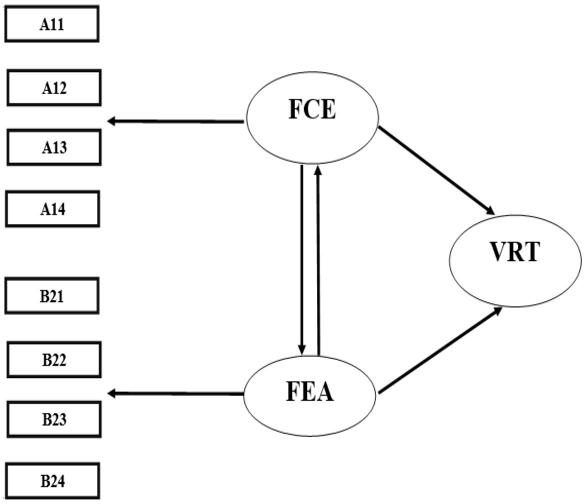


Fig. 1. Relationship of factors

Figure 1 shows the relationship between the factors:
A11 (Reinforced concrete slab),
A12 (Slab and beam slab),
A13 (Live load) and A14 (Dead weight),

B21 (Specific weight of the dry layer),
B22 (Specific weight of the wet layer),
B23 (Wind), and B24 (Precipitation).
where:
FCE means (Structural Constructive Factors) and FEA means (Ergonomic Environmental Factors).

2.1 Structural Constructive Factors

According to a study carried out by the National Chamber of the Clothing Industry published in 2012 (CANAIVE, 2012) shows that Mexican body size and its weight respectively are less than the Americans and Mexican Americans. These measurements were taken with a sample of 17, 364 Mexicans of legal age. The study was carried out in October 2010 to June 2011.

Table 1. Tabulation and normalization of values, given the weight of a Mexican-American equivalent to 81.9 kg average.

No max. Of users	P	Wp
0-24	0.1	1965.6
25-49	0.2	4013.1
50-74	0.3	6060.6
75-99	0.4	8108.1
100-124	0.5	10155.6
125-149	0.6	12203.1
150-174	0.7	14250.6
175-199	0.8	16298.1
200-224	0.9	18345.6
225-250	1.0	20475 kg

2.2 Dead Weight of Concrete Slabs

The calculated deadweight of normal weight concrete slabs cast in place will be increased by 0.2 KN/m² (20 kg / m²). When a layer of normal-weight mortar is placed on a pre-cast or pre-cast slab, the calculated weight of this layer will also increase by 0.2 KN / m² (20 kg / m²) so that the total increase will be 0.4 KN / m² (40 kg / m²) [3].

The value of the resistance of concrete from de f´c=250 kg/cm^2 [15] is used for slabs and columns of houses, social centers and schools. As well as the concrete with f´c=350 kg/cm^2 is used for slabs and columns of buildings [10]. The following table shows the standardized values of the strength for conventional concrete slabs, since the compressive strengths (f'c) is greater than 499 kg /cm^2 are considered as high resistance concretes [11].

Table 2. Standardized values of reinforced concrete roof slabs. Considering: thickness of 10 cm, revoked, flattened, waterproofing and the safety factor.

kg/cm^2	P	W_D
271.5-294.14	0.1	272
294.15-316.79	0.2	297
316.8-339.44	0.3	322
339.45-362.09	0.4	347
362.1-384.74	0.5	372
384.75-407.39	0.6	397
407.4-430.04	0.7	422
430.05-452.69	0.8	447
452.7-475.34	0.9	472
475.35-498	1	497

ɣd	ɣhum
1330 kg	1800 kg

Fig. 2. Specific weight of the dry and wet organic layer respectively (kg / m3).

3 Mathematical Analysis by the Loads Exerted by the Construction Elements on the Roof

By means of the analysis of variables the first equation that allows calculating the total weight that will have, the roof garden, is presented, later presents a second equation whose improvement is a function of the accumulated precipitation for each cubic meter, where the units of kg / m2 and that finally the expected result is expressed in kilograms [13]

$$Z = \frac{M * C}{D}, \tag{1}$$

$$Z = \frac{M * C}{D} + \Delta ɣs, \tag{2}$$

Where:

M= Reinforced concrete slab

C = Live load analysis *Wp*

D = Specific weights of organic layer.

ɣs = Difference of specific weights *ɣd* and *ɣhum*.

Z= Total weight Roof garden

4 Experimentation

Table 3. Value relation Z (in kg).

Minimum Z Values	Maximum Z Values
401.9874	871.9874
896.1584	1366.1584
1467.3032	1937.3032
2115.4216	2585.4216
2840.5137	3310.5137
3642.5795	4112.5795
4521.6189	4991.6189
5477.6321	5947.6321
6510.6189	6980.6189
7651.1842	8121.1842

The above data refer to the possible results that the roof can have, for safety reasons the maximum Z values expressed in kilograms are taken whose last combination exceeds 8 tons.

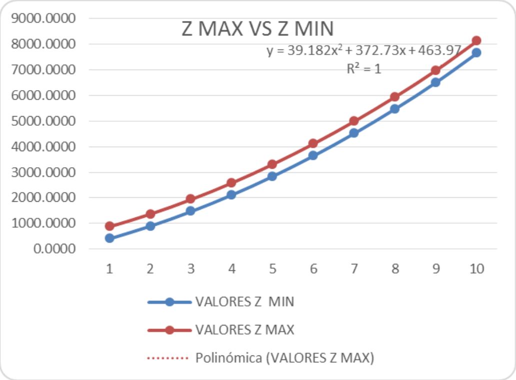


Fig. 3. Minimum and maximum weight (w) comparison.

The results of the Z_{min} values correspond to eq. 1, which contemplates the specific weight of the dry organic layer (ɣd). Consequently, the values of Z_{max} correspond to the equation and the specific weight of the humid organic layer is taken into account, whose equation expresses the addition of the difference of the specific dry and wet weights.

The resulting equation 39.182x2 + 372.73x + 463.97 is a function whose result represents the total load, that can be on the top of the building contemplating the roof slab (W_D), the average weight of people (W_P), and the specific wet weight of the organic layer (ɣhum).

Table 4. Analysis of the maximum weight (w) according to the roof area (m^2).

Surface of buildings house room and department		
m2	P	Support Wmax
100	0.1	19000
200	0.2	38000
300	0.3	57000
400	0.4	76000
500	0.5	95000
600	0.6	114000
700	0.7	133000
800	0.8	152000
900	0.9	171000
1000	**1**	190000

According Building regulation of México City, it establishes the following living loads for Buildings:

Apartments and rooms in houses: 190 kg/m^2,
Meeting places with fixed seats: 350 kg/ m^2.

In relation to the above, the main experiment occurs with the condition of not exceeding 95 tons for example: with the case of the Habitárea Towers in Juriquilla, Querétaro (grupoacerta.com/project/habitarea-towers/), which have an architectural design whose roof area is designed by the following dimensions: 35x14 m=490 m^2 , which works for P=0.5 of the table. Is in this way that the following question arises to which we must answer, for what amount of people is it permissible to add *Wp* load without overloading the roof? and for what amount of area?

5 Analysis of Results

To represent the growth of the loads, it can be observed in this graph that the values are increased in an increasing way, as a result we obtain a polynomial equation of degree 2, which is in function in the data described previously in Table 1.

Table 5. Array with final organic layer and the live load expressed to maximum support.

m^3	Wt (Kg)	No. users	Wp (kg)	Maximum weight
10.55556	19000	231.99023	19000	38000
21.11111	38000	463.98046	38000	76000
31.66667	57000	695.97070	57000	114000
42.22222	76000	927.96093	76000	152000

52.77778	95000	1159.95116	95000	190000
63.33333	114000	1391.94139	114000	228000
73.88889	133000	1623.93162	133000	266000
84.44444	152000	1855.92186	152000	304000
95	171000	2087.91209	171000	342000
105.55556	190000	2319.90232	190000	380000

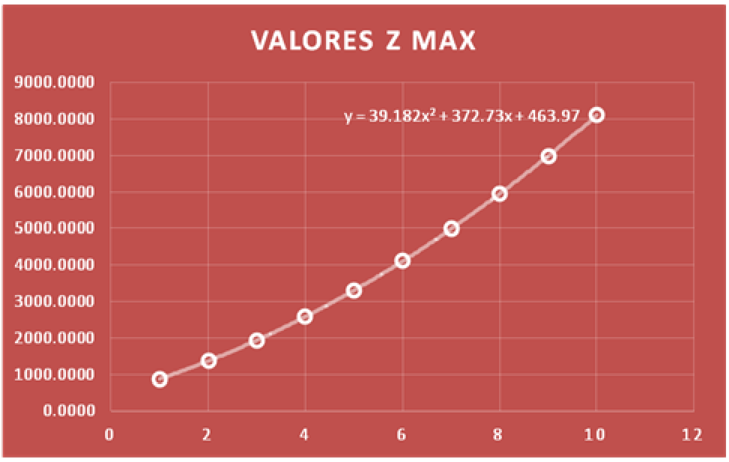


Fig. 4. Polynomial Graphic maximum value z.

In the previous arrangement, in the first column (from left to right) the amount of m3 of organic layer is shown, which is equivalent to the total weight of each value of the second column (Wt) expressed in kg. Similarly, the third column shows the number of users whose equivalences in kg are expressed in the fourth column.

The resulting equation is: 81.9x-1800y = 0, where: x = No. users, y = m3 organic layer.

Table 6. Balance of variables x, y: 50% to 50%.

Approximation	Organic layer	No. Users	Accumulated
-0.0040815	5.27778	115.995115	19000
0.0008370	10.555555	231.99023	38000
-0.0028350	15.833335	347.98535	57000
0.0020835	21.11111	463.980465	76000
-0.0019980	26.38889	579.97558	95000
0.0029205	31.666665	695.970695	114000
-0.0011610	36.944445	811.96581	133000
0.0041670	42.22222	927.96093	152000
0.0000855	47.5	1043.956045	171000
-0.0039960	52.77778	1159.95116	190000

In table 6, 50% of both the organic layer and the number of users are obtained, this with the purpose of achieving a balance between the variables and thereby obtaining the left column of approximations. The results of the left column represent the approximation to 0 that meets the equation 81.9x-1800y = 0; however, the kilograms of the organic layer and the number of users must be rounded to the nearest smaller integer for the purposes of real loads.

Table 7. Equilibrium coefficient to find the optimal point.

Equilibrium coefficient	Rounding Down	
418.5000000	5	115
918.9000000	10	231
1419.3000000	15	347
119.7000000	21	463
620.1000000	26	579
1120.5000000	31	695
1620.9000000	36	811
321.3000000	42	927
821.7000000	47	1043
1322.1000000	52	1159

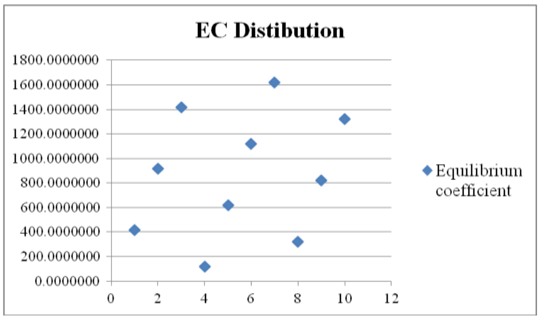


Fig. 5. Dispersion diagram of Equilibrium Coefficient Distribution.

The equilibrium coefficient is obtained after having rounded the variables x, and the nearest integer down. Then, applying the equation 81.9x-1800y = 0 corresponding to the number of users and the weight of the organic layer, we obtain the aforementioned coefficient.

Finally, in the lower part of the diagram we have the lowest point marked with the number 4 which corresponds to the distribution coefficient 119.7, see Table 7.

This indicates the number of people that can be on the roof of a building are 463, see Table 6. To this you can add the own load of the garden up to a maximum limit of 21 tons/m3 (table 7) because the weight of 463 people is 463(81.9) =37,919.7 and 21(1800) =37,800, the sum of the products is 75,719.7 kg does not exceed level 4 (see Table 4). In the table 4 refers the area in m2, where 400 m2 corresponds to 76,000 kg, that´s the reason why that amount of area is chosen.

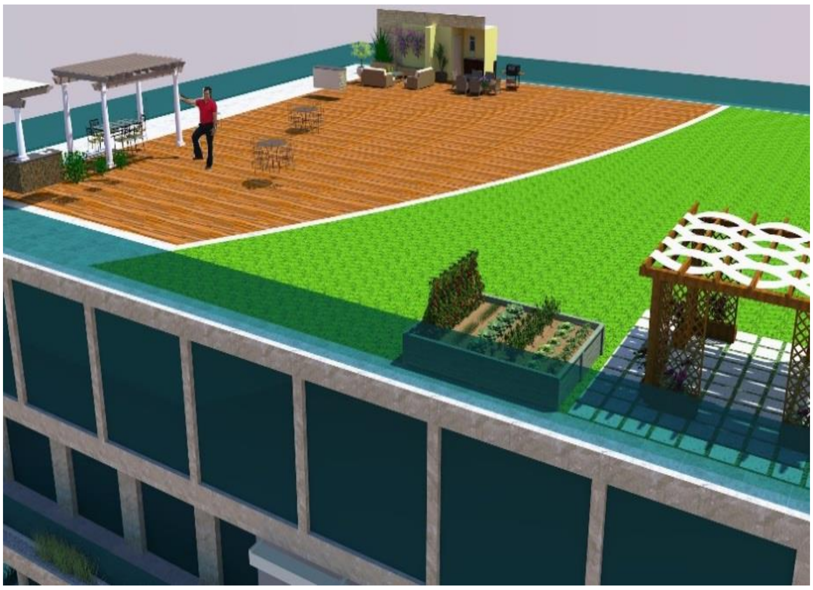


Fig. 6. Roof garden proposed in a Smart City including smart farming and organic layer.

As a result of the research and using Unity software for virtual reality, the prototype was designed according to the data obtained in the previous calculations, which is shown in the Figure 6 (The use of the software is for representative purposes only).

The previous figure is a proposal of a roof slab with roof garden and Smart farming, whose area = 400 m^2, which can hold up to 463 people, which in essence is the optimal point that was sought [12].

6 Conclusions and Future Work

After the experiments it is possible to emphasize the importance of calculating the possible loads that can be had on the roof. That is why it is a high priority to know the maximum number of people that can be occupied without compromising the structural safety of the building. In the study, we reach the conclusion of finding a balance between the variables since they are loads that must be distributed on the slab, otherwise they would become point loads and bring as consequences fracture points, the latter are analyzed in the diagrams at the moment and cutting forces. Is very important to this research integrates a model of virtual reality associate with the final model using virtual reality, in our research we propose a Unity model, as is proposed in Figure 10.

It is necessary to review the calculation memories of the building where the Roof garden is going to be built and specially to emphasize the reinforced concrete elements such as beams and columns. As a last recommendation, you have to review and be sure of the correct distribution of those elements to facilitate the development of the proposal where the live load is balanced with the organic layer.

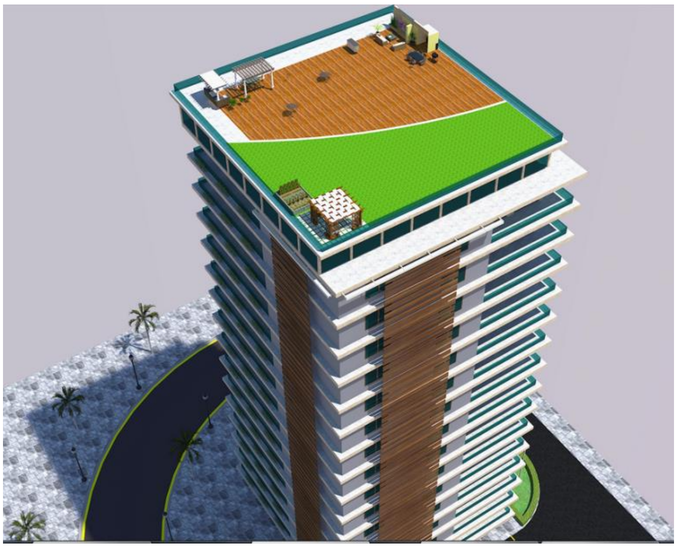


Fig. 7. Roof garden final in a building of 17 levels, 90 m of height and 20m x 20m to dimensions based in final results (400 m^2). It can support 463 people even whit smart farming.

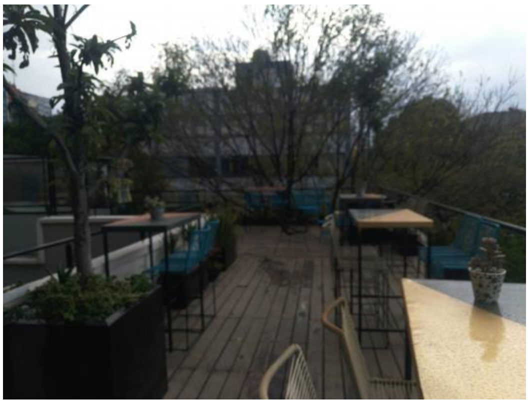


Fig. 8. Proposal of space distribution in a Roof Garden in Averanda, Cuernavaca, Morelos

The number of people at the same time must be carefully analyzed to avoid problems both with the spacing and recreation of the people, and not to affect the group of plant species in it. In our model, 27 different species were chosen, which may exist between them [15].

Another future work is to collect samples of 77 buildings and contemplate those that are under construction or as it is also known as "projection" since these present characteristics that incorporate the category of intelligent buildings, resistant to earthquakes, fires, and with new loads such as the installation of solar panels and intelligent control system.

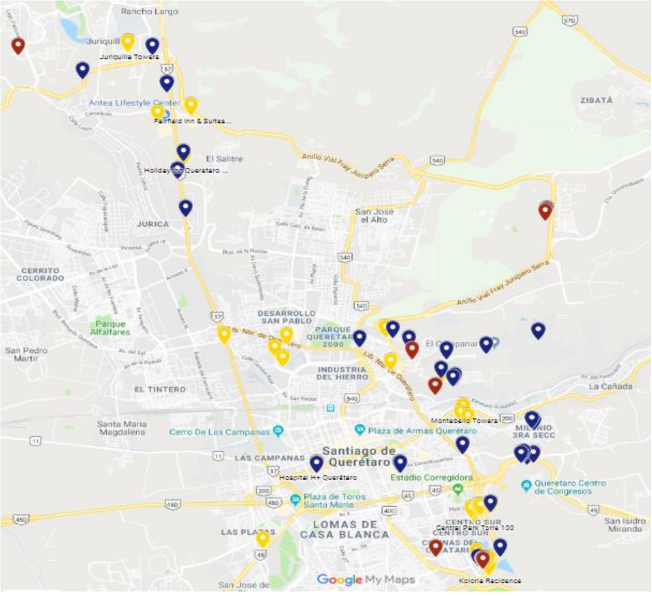


Fig. 9. Map of representative buildings of Querétaro. Where the yellow points represent the buildings constructed, the blue points symbolize the buildings under construction and the red points the buildings projected.

In the city of Querétaro, there has been an ever greater need for corporate offices and housing complexes that is manifested in the current Vertical Construction Boom. There are 28 buildings built exceeding 40 m in height and 2 of the highest are Juriquilla Towers B and Juriquilla Towers A, both with 30 floors, with a height of 116 m and 115 m respectively. There are 30 buildings under construction, where the highest is not strictly the one that has the most floors. The San José Moscati hospital is 130 meters high and 28 levels, while the High Park Corporate 1 is 92 meters high and 29 levels. Finally, there are 19 projected buildings, of which the Westin Querétaro Hotel will be 170m high with a total of 40 floors, this being the tallest building the city will have.

For the Design of Experiments (DOE) we have 77 data and it will be denoted as A = Constructed, B = Under construction, C = Projected to establish a null hypothesis and an alternative hypothesis. With this, a Design of complete blocks can be established at random, 1 block factor and by means of the ANOVA statistical technique with two classification criteria.

Table 8. Higher buildings grouped with two classification criteria: for the height and the number of floors that each building has.

HEIGHT (m)	LEVELS
116 A	30
130 B	28
170 C	40

In this way, it is possible to make a DOE of a factor, first to compare the different levels that each building has and if there are significant differences with respect to height in order to select the buildings that are optimal for the design of a roof garden.

Acknowledgements. We thank the community of Computational Optimization focusing in Smart Cities that for a decade has done research in this area, as well Dr. Alberto Ochoa who gave the Knowledge Set to make reality this research. This research was supported by COQCyT and Instituto Tecnológico de Chetumal.

References

1. Monterusso, M.A., Rowe, D.B., Rugh, C.L.: Establishment and persistence of Sedum spp. And native taxa for green roof applications. Hortscience 40 (2), 391–396 (2005)
2. Johnston, J., Newton, J.: Building green: a guide to using plants on roofs, walls & pavements. London Ecology Unit, London (1995)
3. Simpson, J.R., McPherson, E.G.: Potential of tree shade for reducing residential energy use in California. Journal of Arboriculture 22(1), 23–31 (1996)
4. Syme, S.P., Campbell, E.: Predicting and understanding home garden water use. Landscape and Urban Planning 68(1), 121–128 (2004)
5. Kumar, R., Kaushik, S.C.: Performance evaluation of green Roof and shading for thermal protection of buildings. Building and Environment 40(11), 1505–1511 (2005)
6. Mejía, M., Cornejo, C.: Aplicación del modelo de ecuaciones estructurales a la gestión del conocimiento. In: Arequipa: LACCEI (2010)
7. Casas, M.: Los modelos de ecuaciones estructurales y su aplicación en el índice europeo de satisfacción al cliente. Madrid: Universidad San Pablo-CEU (2002)
8. Ochoa A. et al.: Baharastar–Simulador de Algoritmos Culturales para la Minería de Datos Social. In: Proceedings of COMCEV (2007)
9. Ochoa, A. et al.: Dyoram's Representation Using a Mosaic Image. The International Journal of Virtual Reality (2009)
10. Acme Concretos. Use by resistance. Consulted in: http://www.acmeconcretos.com/ index.php/en/usos-por-resistencia (2018)
11. Concretos, Lima. Concretos Lima S.A.C. (2018)
12. Rudomín, I., Vargas-Solar, G., Espinosa-Oviedo, J., Pérez, H., Zechinelli-Martini: Modelling Crowds in Urban Spaces. Computing and Systems 21(1), (2017)
13. González Barbosa: Construction of an Optimal Solution for a Real-World Routing-Scheduling-Loading Problem. Computing and Systems 13(4), 398-408 (2010)
14. Raisa, S., Micó, M.: Efficient use of Pivots for Approximate Search in Metric Spaces. Computing and Systems 17(4), (2013)
15. Luévanos, A., López, S., Medina, M.: Optimization of Reinforced Concrete Beams for Rectangular Sections with Numerical Experiments. Computing and Systems 22(2), (2018)

ISSN 1870-4069

Patterns of Motivational Orientation and its Relationship with Academic Performance in University Students

María Arely López Garrido[1], Erika Yunuen Morales Mateos[1],
José Alberto Hernández Aguilar[2], Carlos Alberto Ochoa Ortíz[3],
Carolina González Constantino[1], Oscar Alberto González González[1]

[1] Universidad Juárez Autónoma de Tabasco Cunduacán, Tabasco, Mexico
[2] Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos, Mexico
[3] Universidad Autónoma de Ciudad Juárez, Ciudad Juárez, Chihuahua, Mexico
arely.lopez@ujat.mx, erika.morales@ujat.mx,jose_hernandez@uaem.mx,
alberto.ochoa@uacj.mx, carolina.gonzalez@ujat.mx, oscar.gonzalez@ujat.mx

Abstract. This research aims to identify the learning strategies employed by university students; and represent them through graphic techniques of multivariate analysis. The sample consisted of 30 students from the Multidisciplinary Academic Division of the Rivers; of the degrees in Administration and Administrative Computing of the generational cohorts 2010, 2011 and 2012. The instrument used is the Inventory of Strategies of Learning and Motivational Orientation (EDAOM) this instrument consists of a self-assessment that the student makes on the learning strategies that employs. For the implementation of the visual techniques was used the Language R, with these techniques were represented individually to the students with their respective learning strategies. The results show the interaction of the students with the learning strategies and the academic performance, visualization that the students of the Degree in Administrative Computing obtain the lowest values in the strategies of performance and academic performance.

Keywords: apriori, EDAOM, academic performance, visualization, associationrRules.

1 Introduction

The academic performance is an indicator that directly affects the terminal efficiency and the achievement rate, these are indicators that are evaluated in the accreditations of the Educational Programs, therefore for the Institutions of Higher Education it is necessary to increase the academic performance of the students, reduce the disapproval and abandonment to achieve the rates of achievement and terminal efficiency [1], for this reason it is important for Higher Education Institutions to know the factors that affect these indicators and thus implement strategies that help its improvement, likewise the IES perform their administrative procedures and activities through the use of transactional information systems generating a large amount of data, however in many cases these data are not used to generate useful knowledge that supports the decision-making process. the academic administration for the implementation of strategies.

Likewise, since 2000, ANUIES has formulated the document Higher Education in the 21st Century, which includes the Integral Development of Students program, which aims to support students with tutoring and integral development programs to complete their studies in the deadline and achieve the training objectives established in the plans and syllabuses.

Consequently, the IES implemented the Tutoring Programs with the aim of reducing the dropout and failure by improving the use and as a result the terminal efficiency indexes [2], a support tool for tutors is the inventory of Learning Styles and Motivational Guidance. (EDAOM) that is applied to students to identify if students need to reinforce their learning strategies and motivational guidance to study to achieve better academic performance [3].

In this study, the objective is to identify the EDAOM motivational orientation variables that influence academic performance through association rules.

1.1 Related Work

Different studies have been carried out in which data mining techniques are applied to identify the factors associated with academic performance and student desertion. For example, in the Technological Studies of Jocotitlán, a study was conducted to know the factors that cause the School dropout using data mining association rules using data from economic, family and academic backgrounds found various patterns on the factors that determine student desertion [4], likewise Rodríguez-Maya et al. [5] , propose a model to predict the desertion based on the information of the entrance examination and self-reported by students, the obtained model has an accuracy of 86%.

Another study carried out to know the factors that most influence the results obtained from the EXANI-II exam for admission to higher education of CENEVAL, to determine the attributes that most influenced the use of Principal Component Analysis techniques, for the classification methods used algorithms of rules, trees, Bayes, and lazy algorithms and metaalgorithms, obtained as a result that the attributes that most influence the performance of the exam are hours worked, high school average, year of birth of the student, number of books in the student's home and schooling both parents [6].

Márquez et al [7] conducted a study using data mining classification techniques to detect the factors that most influence secondary school students to drop out. They conclude that classification algorithms can be used to predict performance.

On the other hand, advances in the development of automatic information processing tools have provided the creation and use of data analysis techniques [8], generating increasingly sophisticated, dynamic and interactive graphics to visualize data or models [9] , likewise, through visual representation, the relationships described by the graphs are easily understood and easy to remember [10], in this context of data visualization work has been done that show the use of techniques and tools applied to the analysis of educational data [11,12].

1.2 Association Rules

Association rules express patterns of behavior between data in a database. The rules express the combinations of attribute values that occur most frequently [13].

A rule of association can be seen as rules of the form IF α THEN β, where α and β are two sets of disjoint items. The measures to know the quality of an association rule are the coverage (support) and confidence (confidence). Coverage is defined as the number of instances that the rule predicts correctly. Confidence measures the percentage of times the rule is met when it can be applied [13].

A simple and popular association rules learning algorithm is the Apriori, this algorithm is based on the search of the sets of items with certain coverage, in the first place the sets formed by only one item that exceed the minimum coverage are constructed. This set of sets is used to construct the set of item sets, and so on until a size is reached in which there is no set of items with the required coverage [13].

2 Method and Tools

The scope of this research is descriptive because they will specify properties and characteristics describing trends, the design is non-experimental transectional [14]. The method used in this work is that of the knowledge extraction process that is composed of the following phases: data collection, pre-processing, data mining and interpretation of results [15].

The purpose of this paper is to show the existing associations between academic performance and EDAOM variables: perceived effectiveness, internal contingency, perceived autonomy and orientation to external approval including academic and socio-educational variables. The population is made up of new students belonging to the cohorts 2010, 2011 and 2012 of the Bachelor Degrees: Aquaculture Engineering, Food Engineering, Bachelor of Administration, Bachelor of Law, Bachelor of Nursing and Bachelor of Computer Science of the Multidisciplinary Academic Division of the Rivers (DAMR) of the Universidad Juárez Autónoma de Tabasco (UJAT), the sample was non-probabilistic, directed and for convenience; and is composed of 297 students.

2.1 Inventory of Learning Strategies and Motivational Orientation (EDAOM)

The instrument used was the EDAOM questionnaire, the result of this questionnaire is a self-assessment that the student makes about their learning strategies and motivational orientation to the study [3]. The self-report section measures self-assessments of students on: a) frequency, b) ease or difficulty, and c) the results of using a wide variety of learning strategies.

The EDAOM is composed of four scales and 13 subscales; the Self-regulation, Metacognitive and Metamotivational scale consists of three components: those of the person, those of the learning task and learning materials; Table 1 shows the structure of the EDAOM, the first column specifies the scales and the second the subscales that make it up.

In this work the subscales of the EDAOM are considered perceived efficiency, internal contingency and perceived autonomy. The perceived effectiveness refers to the student's evaluation of the strategies used to study and learn, the contingency refers to the student and recognizes the demands of the task What is required in class, autonomy is how dependent or independent is the student perceived to achieve their learning [16].

Table 1. EDAOM scales and subescales.

Scales	Subscales
Acquisition of information	Selective Generative
Recover of information	Subjected tasks During exams
Processing information	Convergent Divergent
Self-regulation, Metacognitive and Metamotivational	Person: Efficiency perceived Internal Contingency Perceived autonomy Orientation to external approval Learning Task: Orientation to the task itself Orientation to the achievement of goal Materials

2.2 Software Used for Data Analysis

A very important element to implement data analysis techniques are the tools, in this case the R language was used.

R is a free software of statistical computing and graphics. Compile and run and compile on a variety of UNIX, Windows and MacOs platforms. It is widely used for the development of statistical software and data analysis, provides a wide variety of statistical models: linear and non-linear models, classical statistical tests, analysis of time series, classification, grouping among others) and graphic techniques. R is an environment in which statistical techniques are implemented and that is extensible through the packages, there are around eight packages that are supplied with the distribution of R [17].

3 Results

The data set is composed of 297 records of new entrants of the 2010-2011 and 2011 generational cohorts, that of the EDAOM taken into consideration for this work is the self-regulation, metacognitive and metamotivational person dimension that make up the subscales: effectiveness , contingency, autonomy and approval, the criteria used for the interpretation of the results of the EDAOM evaluation are shown in Table 2.

The set of data for the generation of the rules is composed of 297 records of new entrants of the degrees of: Aquaculture Engineering, Food Engineering, Administration, Administrative Computing, Law and Nursing; by the nature of the type of study and the algorithm to implement the data were discretized; Table 3 shows the correspondence between the attribute, the type of data and its corresponding value, the attributes that identify groups of students are: career, sex, age, high school graduates;

EDAOM variables: effectiveness, contingency, autonomy and approval; variables of academic trajectory such as: enrolled subjects, approved subjects, average of marks in ordinary exams and academic performance in addition to the variables of expectations of maximum degree of studies to reach, studies of the mother and studies of the father.

Table 2. EDAOM evaluation performance criteria.

Score	Interpretation
100 – 76	Indicates that the students has a good development of learning strategies
75 – 56	Indicates a regular result, so the corresponding subscales have to be reinforced.
55 – 0	Indicates an insufficient result, which is why you have to train the learning strategies

Table 3. Set of attributes used to generate rules.

Atribute	Measurement scale	Values
Bachelor's degree	Ordinal	IAC/LA/ AL/LD/LE/LIA
EST_ALC	Ordinal	TSU/LICENCIATURA/POSGRADO
ESC_MAD	Ordinal	NOESTUDIO/PRIMARIA/SECUNDARIA/BACHILLERATO/CARRERATECNICA/LICENCIATURA/POSGRADO
ESC_PAD		NOESTUDIO/NOLOSE/PRIMARIA/SECUNDARIA/BACHILLERATO/CARRERATECNICA/LICENCIATURA/POSGRADO
SEXO	Nominal	F/M
EDAD	Ordinal	A/B/C/D/E/F/G/H/I
PROM_BACH	Ordinal	EXCELENTE/MUYBIEN/BIEN/REGULAR/SUFICIENTE
MAT_INS	Ordinal	SEIS/SIETE/OCHO/DIEZ/ONCE/DOCE/TRECE/DIECISEIS/DIECISIETE
MAT_ACRE	Ordinal	CERO/UNA/DOS/TRES/CUATRO/CINCO/SEIS/SIETE/OCHO/NUEVE/DIEZ/ONCE/DOCE/TRECE/CATORCE/QUINCE/DIECISEIS
PROM_ORD	Ordinal	BAJO/REGULAR/ALTO
DES_ACA	Ordinal	BAJO/REGULAR/ALTO
EFICACIA	Ordinal	BAJO/REGULAR/ALTO
CONTINGENCIA	Ordinal	BAJO/REGULAR/ALTO
AUTONOMIA	Ordinal	BAJO/REGULAR/ALTO

The Apriori algorithm was applied, using the arules package [17] of the R language, for the generation of the rules a support (s) of at least 0.2 was specified and a confidence (c) of at least 0.9 obtaining 85 rules in total, below are six rules considered most relevant:

Rule 1:
DES_ACAM=BAJO,EFICACIA=BAJO} = => {AUTONOMIA=BAJO} s=0.3468013 c=0.9626168 co=103

This rule specifies that if the student has a low academic performance and in the self-assessment of the strategies used they were effective, it is equal to low then it has a low performance.

Rule 2:
{PROM_A_ORD=BAJO,AUTONOMIA=BIEN} => {DES_ACAM=BAJO} s=0.2828283 c=1.000000 co=84

This rule indicates that students who have a low average in ordinary exams and are perceived as independent in their learning then obtain low academic performance.

Rule 3:
{EST_ALCA=POSGRADO,SEXO=M } => {AUTONOMIA=BIEN} s=0.2558923 c=0.9047619 co=76

This rule indicates that if the students have the expectation of studying a postgraduate course and they are male, they are considered independent in their learning.

Rule 4:
{EST_ALCA=POSGRADO, PROM_A_ORD=ALTO} => {DES_ACAM=ALTO} s= 0.3131313 c=0.9300000 co= 93

This rule indicates that students who have the expectation of studying a graduate degree and have an average of high ordinary then their academic performance is high.

Rule 5:
{SEXO=F, PROM_A_ORD=ALTO, AUTONOMIA=BIEN} => {DES_ACAM=ALTO} s=0.2121212 c=0.9843750 co= 63

This rule indicates that students who have a high average of ordinary exams and that are considered independent in their learning obtain a high academic performance

Rule 6:
{PROM_A_ORD=ALTO, CONTINGENCIA=BIEN,AUTONOMIA=BIEN} => {DES_ACAM=ALTO} s=0.2020202 c=0.9090909 co=60

This rule indicates that students with high average of ordinary exams, which recognizes the demands that tasks require and is considered independent in their learning have a high academic performance.

4 Conclusions

In this work the objective of identifying the subscales of metamotivational orientation that influence the academic performance of university students was achieved, these are self-nomination, effectiveness and contingency, likewise, it is distinguished that if the student has a high motivation to improve by having as expectations to study a postgraduate this is a variable that affects academic performance.

It is observed that students who have a good result in autonomy is to say that they consider themselves independent in their learning that does not depend on others and that they have a good self-assessment in the effectiveness of the learning strategies used and that they know how to recognize the demands The task demands they have a high academic performance.

References

1. Romo-López, A.: La tutoría: una estrategia innovadora en el marco de los programas de atención a estudiantes Asociación Nacional de Universidades e Instituciones de Educación Superior, Dirección de Medios Editoriales (2011)
2. Universidad Juárez Autónoma de Tabasco: Programa Institucional de Tutorías (2003)
3. Castañeda, S.: Educación, aprendizaje y cognición. Teoría en la práctica. México. El Manual Moderno S.A. de C.V. (2004)
4. Reyes-Nava, A., Flores-Fuentes, A., Alejo, R., Lara, E.R.: Minería de datos aplicada para la identificación de factores de riesgo en alumnos. Research in Computing Science 139, pp. 177–189 (2017)
5. Rodríguez-Maya, N., Lara-Álvarez, C., May-Tzuc, O., Suárez-Carranza, B.A.: Modeling Student Dropout in Mexican Universities. Research in Computing Science 139, pp. 163–175 (2017)
6. González-Marrón, D., Enciso-Gonzalez, A., Hernandez-Gonzalez, A.K., Gutierrez-Franco, D., Guizar-Barrera, B., Marquez-Callejas, A.: Evaluación de parámetros de encuesta de ingreso del CENEVAL para alumnos candidatos a ingresar al nivel superior, caso de estudio ITP. Research in Computing Science 139, pp. 135–147 (2017)
7. Márquez-Vera, C., Romero-Morales, C., Ventura-Soto, S, Ventura, S.: Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. IEEE-RITA 7(3) (2012)
8. Pérez, M.: Minería de Datos a través de ejemplos. México. Alfaomega Grupo Editor S.A. de C.V (2014)
9. Ledesma, R., Molina, J., Forrest, W., Valero-Mora, P.: La visualización múltiple en el análisis de datos: una aplicación en ViSta para el análisis de componentes principales. ISNN 01214-9915. Psicothema 19(3), 497–505. http://www.uv.es/molina/journal_publications/2007_vis_multiple.p (2007)
10. Sáez, M.: Diseño e implementación de una aplicación en Processing para la representación de datos multidimensionales. (2015)
11. Lopez-Garrido, M.A., Hernández-Aguilar, J.A., Ochoa-Ortiz Zezzatti, C.A., Morales-Mateos, E.Y., González-Constantino, C.: Comparative Study of Learning Strategies of Bachelor Students in Nursing. Research in Computing Science 122, pp. 153–162 (2016)
12. Morales-Mateos, E.Y., Hernández-Aguilar, J.A., Ochoa-Ortíz Zezzatti, C.A., López Garrido M.A.: A Comparison Represented in the Form of Radar of University Student Engagement in Degrees in Technologies. Research in Computing Science 122, pp. 141–151 (2016)
13. Hernández-Orallo, J., Ramírez-Quintana, M.J., Ferri-Ramírez, C.: Introducción a la Minería de Datos. Pearson Educación S.A. (2008)
14. Hernández, R., Fernández, C., Baptista, M.: Metodología de la investigación (5a ed.). México, D.F., México: McGraw-Hill Interamericana (2010)
15. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine 17(3), p. 37 (1999)
16. Niño, M.: La Relación Estilos de Aprendizaje y Rendimiento Académico en Alumnos de una Facultad de la UANL. Tesis de doctorado. Universidad Autónoma de Tamaulipas.
17. R-Project: R-Project. https://www.r-project.org/about.html (2017)

ISSN 1870-4069

Predicting Airline Customer Satisfaction using *k*-nn Ensemble Regression Models

V. García[1], R. Florencia-Juárez[1], J. P. Sánchez-Solís[1], G. Rivera-Zarate[1],
R. Contreras-Masse[2]

[1] Universidad Autónoma de Ciudad Juárez, División Multidisciplinaria en Ciudad Universitaria
Ciudad Juárez, Chihuahua, Mexico
{vicente.jimenez,rogelio.florencia,julia.sanchez,gilberto.rivera}@uacj.mx
[2] Doctorado en Tecnología
Universidad Autónoma de Ciudad Juárez
Ciudad Juárez, Chihuahua, Mexico
rcontreras@itcj.edu.mx

Abstract. Customer satisfaction questionnaires are a rich and strong source of information for companies to seek loyalty, customer and client retention, optimize resources, and repurchase products. Several advanced machine learning and statistical models have been employed to estimate the customer satisfaction score; however, there is not a single model that can yield the best result in all situations. Ensembles of regression techniques have demonstrated their effectiveness for various applications, where the success of these models lies in the construction of a set of single models. We perform an experimental study using a real database of 129,890 samples from airline companies, in order verify the benefits of ensemble models for predicting customer satisfaction. Accordingly, the present paper evaluates the BAGGING ensemble model using the well-renowned *k*-nn algorithm as the base learner. The obtained results indicate that the BAGGING ensemble performs better than the single classifier in terms of RMSE and MAE.

Keywords: regression, customer satisfaction, ensemble, *k*-NN, BAGGING.

1 Introduction

Companies utilize customer satisfaction surveys and questionnaires to find out what a client think about a product, service, brand or the company. Therefore, the customers' satisfaction databases play a crucial role in the productive decision process. Exploiting datasets to get valuable information allows to make an overall marketing strategy that helps the companies retain existing customers and add new customers. Consequently, an effective customer satisfaction data analysis represents a challenge and provide opportunities in several areas as machine learning, data mining, and marketing [1,2,3].

Machine learning methods can predict the customer satisfaction when a service is provided. The research in this problem has been addressed from a classification and regression point of view. Park et al. [2] used four machine learning algorithms (naïve

Bayes, decision tree, logistic regression, and support vector machine) on a dataset acquired from a contact center. The prediction task was addressed as a classification problem, where the best model obtained an accuracy of 66%. Roy et al. [4] employed naïve Bayes, multiclass classifier, k-star, and IBK (Instance-Based learning with parameter K) as classifiers models for predicting customer satisfaction from a database constructed from a customer survey conducted by the San Francisco International Airport. Aktepe et al. [5] showed that using classifier algorithms combined with programming software and structural equation modeling is able to analyze the level of customer satisfaction and loyalty. Farhadloo et al. [6] analyzed the satisfaction of customers using reviews from different states parks in California. These reviews were left by real visitors on TripAdvisor.com. Grigoroudis and Politis [7] suggested that the customer satisfaction problem can be seemed as a multicriteria evaluation problem. Thus, they proposed the MUlticriteria Satisfaction Analysis system (MUSA) that uses ordinal regression techniques. Bockhorst et al. [8] developed a hybrid customer satisfaction system by integrating a linear ranking sub-model and a non-linear isotonic regression. The system was trained on a database constructed by using phone calls from five surveys. Experimental results revealed that the proposed model is better than standard regression techniques. Other algorithms that have also been employed in the customer satisfaction analysis are the CART (Classification And Regression Tree) algorithm [9], artificial neural network approaches [10,11,12], the principal component analysis [13], and the support vector machine algorithm [14].

Although previous studies conclude that data mining and machine learning techniques can be successfully used for prediction of customer satisfaction, there is not an overall best algorithm for dealing with customer satisfaction problems. Consequently, ensembles models have emerged to exploit the different behavior of individual techniques and reduce prediction errors. Several practical investigations have demonstrated that ensemble models perform better than single prediction methods in classification [15,16,17,18,19] and regression problems [20,21,22 23].

The focus of this paper is primarily on exploring the use of ensembles of regression models in the scope of customer satisfaction. In particular, we analyze the performance of the k-nearest neighbor (*k*-nn) model for regression as base classifier in the BAGGING (Bootstrap AGGregatING) ensemble model. This approach is the most common ensemble learning algorithm, where the diversity is achieved by the manipulation of the training samples [24]. BAGGING has showed a good performance on various real problems and provide practical algorithms for constructing ensemble models [16]. Therefore, the paper shows how *k*-nn ensemble regression models can result useful to estimate a customer satisfaction score from a real-database constructed from 129,889 surveys supplied by several airline companies.

Henceforth, the rest of the paper is organized as follows. Section 2 introduces the bases of the ensemble model and regression technique that will be explored in this study. Section 3 provides the experimental set-up and the description of the database used in our experiments. Next, the results are reported and discussed in Section 4. Finally, Section 5 summarizes, the main conclusions and points out some directions for future research.

2 Regression Ensemble Models

An ensemble of regressors consists of a set of individually trained models (the base learners) whose decision is integrated in some way when predicting the output of new examples. By combining individual regression models, the ensemble approaches aiming to minimize the error on problems where the output variable is continuous. The construction process of ensembles is performed by following three steps [25]: 1) generation, 2) pruning, and 3) integration. In the former a set of base learners is constructed; if the base learners are the same, then the construction is homogenous, on the otherwise is heterogeneous. Some redundant models can be generated; therefore, a pruning process can be performed. Finally, the prediction of the models is combined in different ways by fusion or selection.

The set of models is trained by using different subsamples of the training data. A popular resampling method is the bootstrap that, given a set D containing n examples, generates training sets by drawing n examples at random with replacement from D.

2.1 *k*-NN Regression Model

The k-nearest neighbor (k-NN) model is a non-parametric technique that works under the assumption that new samples share similar properties with the set of stored samples. In brief, given a set of n labeled examples (training set), say $D = \{(x_1, a_1), (x_2, a_2), \cdots, (x_n, a_n)\}$ and identically distributed (i.i.d.) random pairs (x_i, a_i), where $x_i \in \mathbb{R}^d$ and a_i denotes the target value associated it, the k-nn classifier consists of assigning a new input sample y to the class most frequently represented among the k closest instances in the training set, according to a certain similarity measure (generally, the Euclidean distance) in the d-dimensional feature space $\mathbb{R}^d$.

The straightforward implementation of the k-nearest neighbor regression model applied to a test sample y first calculates $\delta_i = d(x_i, y)$, for $i = 1, \cdots, n$. Then we get the indices $\{p_1, \cdots, p_k\}$ of the k smallest values such that $\delta_{p_b} \leq \delta_j$, $\forall j \notin \{p_1, \cdots, p_k\}$, $b = 1, \cdots, k$, and $\delta_{p_1} \leq \cdots \leq \delta_{p_k}$. We define x_{p_b} as the p_b $-$nearest neighbor sample and a_{p_b} is the corresponding target value.

Regression aims to learn a function $f\colon y \to a$ to predict the a value for a new sample $y = [y_1, y_2, \cdots, y_d]$. The k-nn regression model estimates the target value $f(y)$ of a new input sample y by averaging the estimated target values of its k-nearest neighbors [26]:

$$f(y) = \tfrac{1}{k} \textstyle\sum_{b=1}^{k} a_{p_b}, \tag{1}$$

where a_{p_b} denotes the target value of the b-th nearest neighbor.

An illustrative example of the k-nn regression model is displayed in Fig. 1. Using $k = 3$, the output of a new example y is estimated computing the mean among the responses of the 3 nearest neighbors (instances A, B, and C). When $k = 5$, the output of instances A, B, C, D and E are averaged. If $k = 1$, then the k-nn regression model assigns to $f(y)$ the value a_i from x_i, that is the training sample nearest to y.

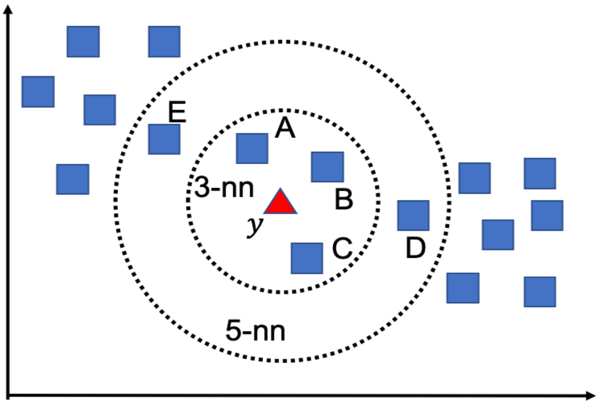


Fig. 1. An example of k-NN regression model with k = 3 and 5.

2.2 BAGGING Ensemble Model

Breiman [26] proposed the BAGGING algorithm that creates multiple models trained from different bootstrap subsamples $D_1, D_2, \cdots, D_M$, each one consisting of n samples drawn at random with replacement from the original D training set. Fig. 2 shows an example of a regression ensemble model based on BAGGING, where in the first step M bootstrap replicates of the training set D are generated. Afterwards, each base regressor R_i is trained on the bootstrap D_i. Thus, the ensemble is conformed by different models, where each one is not exposed to the same set of samples. This creates the diversity necessary to cover a wide range of situations [16]. By this way, the output of new observations will be predicted by taking the average of the ensemble R^* built from $R_1, R_2, \cdots, R_M$. The procedure of training and testing of BAGGING can be resumed in nine steps:

Training step

1. $R^* = \emptyset$, the ensemble
2. M number of regression models to train
3. $For\ i = 1, \cdots, M$
4. Create a boostrap sample of size n, D_i from D
5. Train a regression model R_i using D_i as the training set
6. Add the trained model to the current ensemble, $R^* = R^* \cup\ R_i$

Testing step

7. Run $R_i, \cdots, R_M$ on the new input y and compute $f_{R_i}(y)$ using Eq. (1)
8. Average the outputs of $R_i, \cdots, R_M$ using

$$f_{R^*}(y) = \frac{1}{M}\sum_{1=1}^{M} f_{R_i}, \tag{2}$$

9. f_{R^*} is the predicted target value.

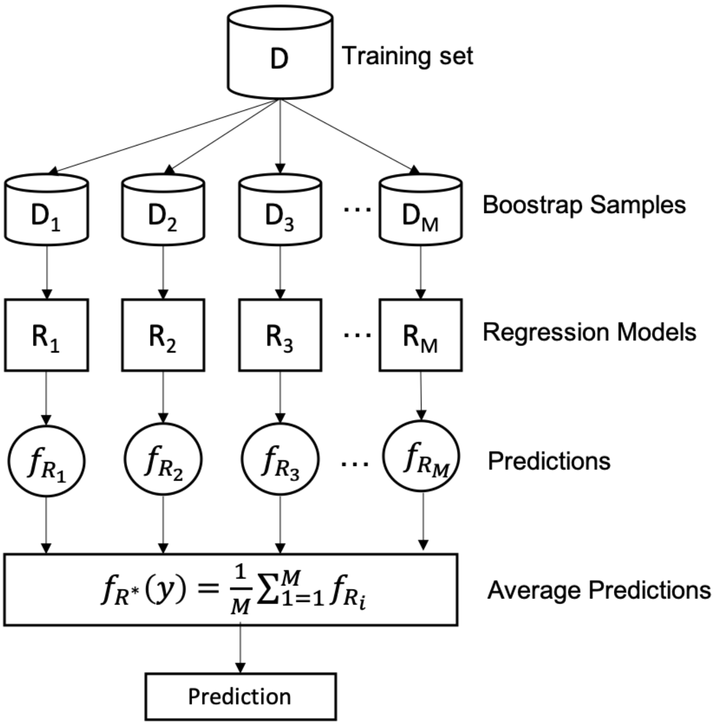


Fig. 2. An ensemble regression model based on BAGGING.

3 Database and Experimental Set-up

As already stated, this study aims to evaluate the performance of regression ensembles for airline customer satisfaction prediction. Thus, we conducted a pool of experiments on a real dataset taken from airline companies in the USA. This is a subset of samples collected on 2014 from customer satisfaction questionnaires. It consists of 129,889 instances, where each one is represented by 24 input variables: airline status, age, gender, price sensitivity, year of first flight, no. of flights, percentage of flights with other airlines, type of travel, no. of other loyalty cards, shopping amount at the airport, eating and drinking at airport, class, day of month, airline code, airline name, origin city, destination, schedule departure hour, departure delay in minutes, arrival delay in minutes, flight cancelled, flight time in minutes, flight distance, and arrival delay greater 5 minutes. The output variable is satisfaction that is a five-point score measurement (i.e., from 1 to 5). Since some variables were categorical, they were first converted into numeric value, and then these values were normalized into the range [0,1].

Table 1 sums up the main characteristics of the database used in the empirical analysis: the attribute number, the attribute description and some statistics, such as the minimum and maximum values of the attribute, the mean and the standard deviation.

We focused our study on the BAGGING ensemble model and the simple *k*-nn algorithm used as the base classifier. The *k*-value used for the regression algorithm was set to 3. Besides, we aim to analyze the performance of the BAGGING ensemble when varying the number of bootstrap subsamples; therefore, six different bootstrap values were tested: 5, 10, 15, 20, 25, and 30. All models were taken from the WEKA toolkit [28].

To prevent inaccurate performance estimates and following the standard strategy used in other works, we evaluate the performance of regression models by a 5-fold cross-validation model [29,30,31]. The original dataset was randomly divided into five disjoint parts. Each fold is used as a test set and the remaining four folds are used for training the regression models. Thus, we can get 5 different test set performances and therefore 5 different trained models. Note that the bootstrap samples are generated for each training set.

To assess the model performance in regression problems we used two of the most popular performance measures that estimate how much the prediction $(e, \cdots, e_n)$ deviate from the actual target values $(a_1, \cdots, a_n)$ [32]. These metrics are the Root Mean Square Error (RMSE) [33,35,36],

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i - a_i)^2}, \tag{3}$$

and the Mean Absolute Error (MAE) [34, 35],

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i - a_i|. \tag{4}$$

Both metrics are minimized when the predicted value for each test sample agrees with their true value. For each fold, we record the RMSE$^{(i)}$ and the MAE$^{(i)}$ (i=1,2, …,5) and compute the final estimate as the mean of all folds:

$$RMSE_{avg} = \frac{1}{5}\sum_{i=1}^{5} RMSE^{i}, \tag{5}$$

$$MAE_{avg} = \frac{1}{5}\sum_{i=1}^{5} MAE^{i}. \tag{6}$$

Table 1. Characteristics of the airline customer satisfaction dataset used in the experiments.

No.	Description	Minimum	Maximum	Mean	Std. Dev.
	Input variables				
1	Airline status	1	4	1.747	1.205
2	Age	15	85	46.196	17.321
3	Gender	1	2	1.435	0.496
4	Price sensitivity	0	5	1.276	0.546
5	Year of first flight	2003	2012	2007.209	2.977
6	No. of flights	0	100	20.091	14.362
7	Percentage of flights with other airlines	1	110	9.314	8.761
8	Type of travel	1	3	1.696	0.911
9	No. of other loyalty cards	0	12	0.884	1.142
10	Shopping amount at the airport	0	879	26.553	53.081
11	Eating and drinking at the airport	0	895	68.242	52.21
12	Class	1	3	2.024	0.431
13	Day of month	1	31	15.723	8.659
14	Airline code	1	14	8.134	4.441
15	Airline name	1	14	7.061	4.341
16	Origin city	1	295	128.852	83.059
17	Destination city	1	296	123.289	82.153
18	Scheduled departure hour	1	23	12.896	4.623
19	Departure delay in minutes	0	1592	14.713	38.07
20	Arrival delay in minutes	0	1584	15.045	38.415
21	Flight canceled	1	2	1.018	0.135
22	Flight time in minutes	0	669	109.164	72.8
23	Flight distance	31	4983	793.804	592.125
24	Arrival delay greater than 5 minutes	1	2	1.343	0.475
	Output variable				
25	Satisfaction	1	5	3.379	0.965

4 Results and Discussions

Fig. 3 and 4 display the RMSE and the MAE averaged across the five runs. For each prediction model, we have plotted the results for the base learner (3-nn) and also the results of the 3-nn ensemble regression model when varying the number of bootstrap subsamples from D. Note that the line parallel to X-axis corresponds to the case of base learner, which indicates that the results were achieved by learning directly from the training set D.

From Fig. 3 and 4 as expected, the individual regression model achieves the highest error values ($RMSE \approx 0.7985$, $MAE \approx 0.6365$), whereas, the ensemble regression models appear as the learning model with the lowest overall error ($RMSE \approx 0.7650$ and $MAE \approx 0.5664$ when number or regression models is set to 30). This is a difference between the two models ≈ 5% when $RMSE$ is used and ≈ 11% in the case of MAE.

On the other hand, when varying the number of bootstraps from D, both the RMSE and the MAE rates decrease as the number of bootstrap subsamples increases, although the configuration of the base classifier does not vary along the different bootstrap subsamples.

From the results here reported, it appears that the customer satisfaction prediction problem can be handled better using ensembles approaches than single models. Likewise, when the number of base models is increased, the error decrease.

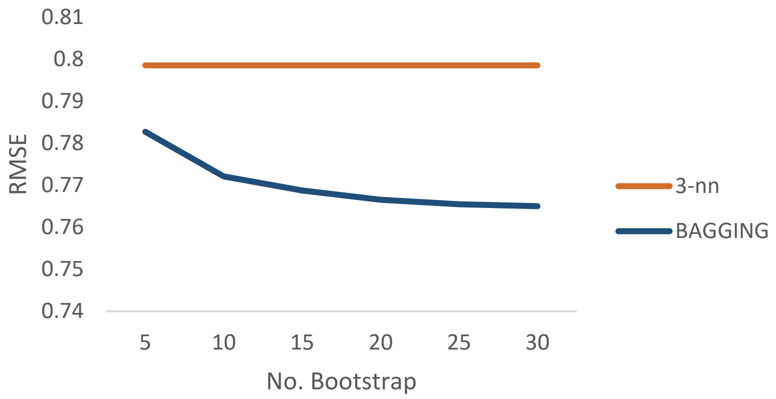


Fig. 3. Average RMSE for the *3*-nn classifier (single learner) and the BAGGING when varying the number of bootstrap subsamples.

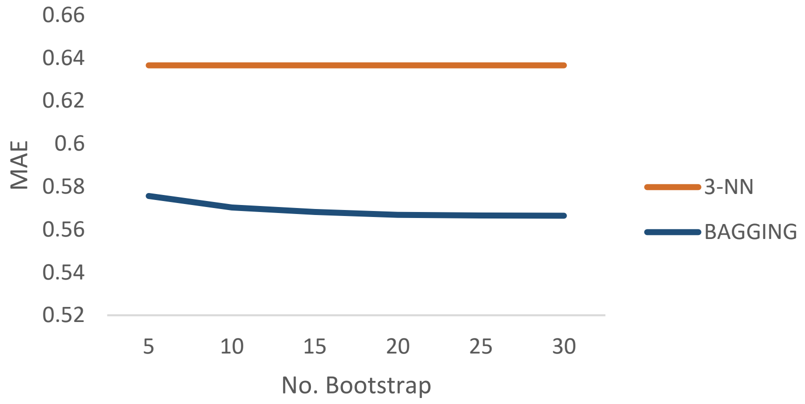


Fig. 4. Average MAE for the *3*-nn classifier (single learner) and the BAGGING when varying the number of bootstrap subsamples.

5 Conclusions and Future Work

The present paper has analyzed the performance of ensemble regression models for an airline customer satisfaction prediction problem. In particular, BAGGING has been taken as representative of ensemble models, whereas the *k*-nn has been employed as the base learner. To this end, a real-life airline customer satisfaction dataset was used to build all the models. From the experiments carried out, we have observed that regarding both RMSE and MAE the ensemble regression models have produced the best results. A final indication of the experiments is that using a significant number of bootstrap subsamples the error may decrease.

Several directions for further research have emerged from this study: (i) to extend the present analysis to other ensemble approaches; (ii) to incorporate a feature selection phase to remove any attribute that might be considered noisy or irrelevant; (iii) to use some prototype selection method to reduce the complexity of data (border and redundant points).

Acknowledgement. The authors would like to acknowledge the financial supports from the Mexican PRODEP [UACJ-PTC-373, UACJ-PTC-423].

References

1. 10 key marketing trends for 2017, https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN, last accessed 2018/06/25
2. Park, Y., Gates, S.: Towards real-time measurement of customer satisfaction using automatically generated call transcripts. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 1387–296. ACM (2009)
3. Gurau, C., Ranchhod, A.J.: Measuring customer satisfaction: A platform for calculating, predicting and increasing customer profitability. Journal of Targeting, Measurement and Analysis for Marketing 10(3), 203–209 (2002)
4. Roy, S.S., Kaul, D., Barna, C., Mehta, S., Misra, A.: Prediction of customer satisfaction using naïve Bayes, multiclass classifier, k-star, and IBK. In: Balas V., Jain, L., Balas, M. (eds.) Soft computing, SOFA 2016, Advances in Intelligent Systems and Computing, vol 634, 153–161. Springer, cham, Arad, Romania (2016)
5. Aktepe, A., Ersoz, S., Toklu, B.: Customer satisfaction and loyalty analysis with classification algorithms and structural equation modeling. Computers & Industrial Engineering 86, 95–106 (2015)
6. Farhadloo, M., Patternson, R.A., Rolland, E.: Modeling customer satisfaction form unstructured data using a Bayesian approach. Decision Support Systems 90, 1–11 (2016)
7. Grigoroudis, E., Politis, Y.: Multiple criteria approaches for customer satisfaction measurement. In: Matsatsinis, N., Grigoroudis, E. (eds) Preference Disaggregation in Multiple Criteria Decision Analysis. Multiple Criteria Decision Making, pp. 95–123. Springer, cham (2018)
8. Bockhorst, J., Yu, S., Polania, L., Fung, G.: Prediction self-reported customer satisfaction of interactions with a corporate call center. In: Altun Y. et al. (eds) Machine Learning and Knowledge Discovery in Databases, ECML PKDD 2017, Lectures Notes in Computer Science, vol. 10536, pp. 179-190. Springer, Skopje, Macedonia (2017)

9. Dobrota, M., Bulajic, M., Radojicic, Z.: Data mining models for prediction of customer satisfaction: The CART analysis. In: Jaksic, M.L., Rakocevic, S.B. (eds) Innovative Management and Firm Performance, pp. 401-421. Palgrave (2014)
10. Mashinchi, R., Selamat, A., Ibrahim, S., Krejcar, O., Penhaker, M.: Evaluating customer satisfaction: Linguistic Reasoning by Fuzzy Artificial Neural Network. In: Barbucha, D., Nguyen, N., Bartubara, J. (eds), New Trends in Intelligent Information and Database Systems, Studies in Computational Intelligence, vol. 598, 91–100, Springer cham (2015)
11. Yau, H.K., Tang, H.Y.H.: Analyzing customer satisfaction in self-service technology adopted in airports. Journal of Marketing Analytics 6(1), 6–18, (2018)
12. Segura, C., Balcells, D., Umbert, M., Arias, J., Luque, J.: Automatic speech feature learning for continuous prediction of customer satisfaction in contact center. In: Abad, A. et al. (eds), Advances in speech and language technologies for Iberian languages, IberSPEECH 2016. Lecture Notes in Computer Science, vol. 10077, pp. 255–265, Springer, Cham (2016)
13. Hu, B.: Application of data mining in power customer satisfaction evaluation. In: Qian, Z., Cao, L., Su, W., Wang, T., Yang, H. (eds), Recent advances in computer science and information engineering, Lecture Notes in Electrical Engineering, vol. 124, 37-44, Springer, Berling (2015)
14. Jiang, Z., Zan, W., Liu, X.: Customer satisfaction analysis based on SVM. In: Zu, Q., Hu, B. (eds), Human centered computing, HCC 2016. Lecture Notes in Computer Science, vol. 9567, 683–688. Springer, Cham (2016)
15. Rodriguez, J.J., Kuncheva, L. I., Alonso, C.J.: Rotation forest: a new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(10), 1619–1630 (2006)
16. Kuncheva, L. I.: Combining pattern classifiers: Methods and algorithms. 2nd edn. Wiley & Sons, Hoboken (2014)
17. Osareh, A., Shadgar, B.: An efficient ensemble learning method for gene microarray classification. BioMed Research International 210, 1–10 (2013)
18. Rokach, L.: Ensemble-based classifiers. Artificial Intelligence Review 33(1-2), 1–39 (2010)
19. Cabrera-Hernández, L., Morales-Hernández, A., Casas-Cardoso, G.M.: Medidas de diversidad para la construcción de sistemas-multiclasificadores usando algoritmos genéticos. Computación y Sistemas 20(4), 729–747 (2016)
20. Mendes-Moreira, J., Soares, C., Jorge, A.M., De Sousa, J.F.: Ensembles approaches for regression: A Survey. ACM Computing Surveys 45(1), 10:1–10:40 (2012)
21. Frayman, Y., Rolfe, B.F., Webb, G.I.: Solving regression problems using competitive ensemble models. In: LNAI vol. 2557, 511–522 (2002)
22. Sun, Q., Pfahringer, B.: Bagging ensemble selection for regression. In: Thielscher, M., Zhang, D. (eds), AI 2012: Advances in Artificial Intelligence. Lecture Notes in Computer Science, vol 7691, 695–706, Springer, Berlin (2012)
23. Trejo, K., Angulo, C.: Single-camera automatic landmarking for people recognition with an ensemble of regression trees. Computación y Sistemas 20(1), 19–28 (2016)
24. Bonte, I., Rodríguez, A., García, M. M, Grau, R.: Combinación de clasificadores para bioinformática. Computación y Sistemas 16(2), 191–201 (2012)
25. Ren, Y., Zhang, L., Suganthan, P.N.: Ensemble classification and regression – Recent developments, applications and future directions. IEEE Computational Intelligence 11(1), 41–53 (2016)
26. Breiman, L.: Bagging Predictors. Machine Learning 24(2), 62–-635 (1996)
27. Guyader, A., Hengartner, N.: On the mutual nearest neighbors estimate in regression. Journal of Machine Learning Research 14, 2361–2376 (2014)

28. Witten, I., Frank, E., Hall, M., Pal, C.: Practical machine learning tools and techniques. 4th edn., Morgan Kaufmann (2016)
29. Molinaro, A. M., Simon, R., Pfeiffer, R.M.: Prediction error estimation: a comparison of resampling methods. Bioinformatics 21(15), 3301–3307 (2005)
30. Reader, T., Forman, G., Chawla, N. V.: Learning from imbalanced data: evaluation matters. In: Holmes, D.E., Jain, L.C. (eds), Data Mining: Foundations and Intelligent Paradigms. Intelligent System Reference Library, vol. 23., 315–331, Springer, Berling
31. Gacto, M.J., Galende, M., Alcará, R., Herrera, F.: METSK-HD^e: A multiobjective evolutionary algorithm to learn accurate TSK-fuzzy systems in high-dimensional and large-scale regression problems. Information Sciences 276, 63–79 (2014)
32. Caruana, R., Niculescu-Mizil, A.: Data mining in metric space: an empirical analysis of supervised learning performance criteria. In: 10th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 69–78, New York (2004)
33. Madrigal Espinoza, S. G.: Modelos de regression para el pronóstico de series temporales con estacionalidad creciente. Computación y Sistemas 18(4), 821–831 (2014)
34. Luna Sandoval, M. R. O, Ruiz Ascencio, J.: MUREM: Un método multiplicativo de regresion para estimar el esfuerzo de desarrollo de software. Computación y Sistemas 20(4), 763–787 (2016)
35. Béjar Chacón, W. E., Valeriano Valdez, K. Y., Ilachoque Umasi, J. L., Sulla Torres, J.: Predicción de caudales medios diarios en la cuenca del Amazonas aplicando redes neuronales artificiales y el modelo neurodifuso ANFIS. Research in Computing Science 113, 23–25 (2016)
36. Pinzón Pineda, S. A., Hernández Aguilar, J. A., Arroyo-Figueroa, G.: Aplicación de modelos auto regresivos para la predicción de generación de energía eléctrica a partir de datos eólicos. Research in Computing Science 139, 59–70 (2017)

ISSN 1870-4069

Predictive Model as a Tool for Acquiring a Certification for Client Companies and Certifying Entities with Machine Learning

Edgar Gonzalo Cossio Franco[1], Jorge Alberto Delgado Cazarez[2], Daniel Noel Torres Godoy[3]

[1] Instituto de Información Estadística y Geográfica de Jalisco, Mexico
[2] Universidad de Guadalajara, Mexico
[3] Universidad Enrique Díaz de León, Mexico
kofrran@gmail.com, guero10delgado@gmail.com, dt_godoy@hotmail.com

Abstract. Companies that seek to certify their processes do so with the aim of guaranteeing quality, although few seek it and least of all do so. CMMI is a model that certifies the maturity of the development of products and services. The present work has two proposals: the first is a tool in Java that determines when a client company is apt or not to a CMMI certification and the second is an intelligent analysis model based on machine learning that determines, from predictions, scenarios for decision making.

Keywords: CMMI, machine learning, predictive model, java.

1 Introduction

The quality of a company can be measured by the maturity of its processes [1]; In the case of software development companies, certification of capability maturity model integration (CMMI) is sought in order to guarantee, in addition to quality, productivity, customer satisfaction, performance and business model [2].

In Mexico, for the year 2016, there was a registry of 781 certified development centers with some quality model, as shown in Figure 1, 62% have MOPROSOFT certification while 35% have CMMI certification and 3% PACE.

In Figure 2 they have the certifications that have the development results that have 2 and 3, and those that have CMMI, that have more than one constellation. Thus, a representative of the Development Centers that have verification and certification in two quality models with 31, b Represents Development Centers that have verification, certification and evaluation in three quality models with 2 and c Represents Development Centers that have certification in more than one CMMI constellation with 11.

The importance of companies starting to see in certifications a door of opportunity to ensure maturity in construction processes that translate into quality is increasing as without certification they will be building software of poor quality.

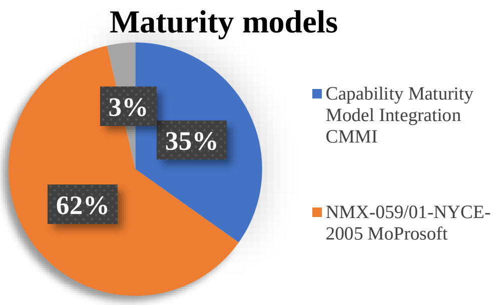


Fig. 1. Maturity models [3].

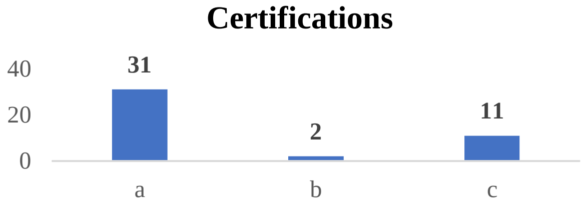


Fig. 2. Certifications center [3].

A. CMMI

The Capability Maturity Model Integrated (CMMI) is a reference framework for improving the processes of developing products and services of companies. It consists of five levels, three constellations and nineteen process areas [4]. Figure 3 shows CMMI maturity levels.

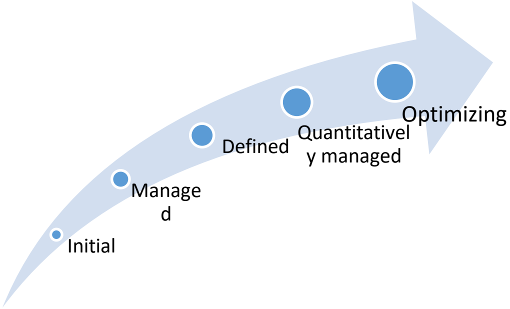


Fig. 3. CMMI Process [5].

The initial level shows a lack of control and chaos in the process; there are no maturity, control or documentation mechanisms. Success depends on superhuman effort.

In level two, the improvement actions are after something happened. No improvement scenarios are anticipated. There is knowledge of project management. At level three there are defined processes and there are metrics.

At level four, the collected data is used to manage and improve processes.

At level five, the processes continuously improve in a natural, sustained and constant manner.

B. Machine Learning

Machine learning is defined as the process by which computers learn automatically [6]. It is based on algorithms such as Bayesian classifiers for probabilities, in nearest neighbor classifiers for similarities, in artificial neural networks, in decision trees, in genetic algorithms, clustering, fuzzy logic and bioinspired algorithms to build global and local searches [7,8].

In 1943, Warren McCullock and Walter Pitts, in an attempt to explain the biological brain and apply it to the design of artificial intelligence, introduced the first concept of a simplified brain cell [9].

Later, in 1957, “Frank Rosenbelt proposed an algorithm that would automatically learn the optimal weight coefficients that are then multiplied with the input features in order to make the decision of whether a neuron fires or not” [9].

The classification is a supervised task belonging to Machine Learning; this identification the properties of a batch of data and, taking into account parameters or labels, previously configured, assign this data to one of the labels; this refers to a supervised action.

One of the algorithms dedicated to the classification is j48; its function is based on generating a decision tree with the variables it has, through partitions made recursively according to the search strategy first in depth [10]. The grouping is an unsupervised task of Machine Learning, which, automatically, divides the data into groups to help us obtain the groups [11]. This refers to an unsupervised task.

The k-means algorithm achieved the grouping principle in a fast and efficient way. Make a random selection of observations, such as clustered groups, then, later, assigned to the nearest point. This division of patched space is known as a Voronoi tasseling [12].

Some of the strengths of the k-means algorithm:

- It uses simple principles to identify clusters that can be expanded in non-statistical terms.
- It is highly flexible and can be adapted to address almost all its deficiencies with simple adjustments.
- It is quite efficient and performs well by dividing the data into useful groups. [11]. Unlike classification algorithms, association rules are used for the discovery of unsupervised knowledge in large databases. In the case of association rules, it is not necessary for the data to be tagged ahead of time, the program simply triggers a set of data, which are expected to find interesting.
- [11] The analysis of association rules is used to find a good connection between a large number of variables. The most common example is the analysis of the market basket, however, it can also be used for topics such as:
  - Search for interesting and frequent patterns in DNA and protein sequences in a cancer data analysis.
  - Find patterns of purchases or medical claims that occur in combination with fraudulent credit cards or insurance use.

  - Identify combinations of behavior that cause customers to abandon their cell phone service or update their cable television package.

According to Lantz Brett, you can say that this type of task, if it can be done by a person, you need to have a person who has an expert level in the area in which you work and that, based on your great experience, can define some pattern or algorithm. However, it is the case in which the database is very large, it is an impossible job for a single person; look for a needle in a haystack [11].

2 Problem

Currently, companies that seek to certify their maturity processes in the construction of products and services do not succeed because, more than beyond the methodologies, standards, documentation and good practices, it has to do with the change in the way of being of capital intellectual, that is, of the people who are involved in the areas. According to [13] the problem lies in fears, ignorance, resistance to change, not wanting to invest money, little training and lack of time, as shown in Figure 4.

time
fear
training
Issues
change
money

Fig. 4. Issues implementing CMMI [9].

3 Proposal

In the present work a predictive model is presented to implement CMMI by means of machine learning which has two objectives: the first objective is to support the client companies (CliC) to know their company or is candidate to aspire to certification through a developed program in JAVA that determines the score based on the answers provided by the personnel involved in the project. The second objective is to support the certification company (CertC) in the analysis of data for decision making and thus identify the client companies (CliC) that are candidates for certification and those that do not, by filtering, classifying, grouping and association.

4 Methodology

This section shows the proposed methodology which is distinguished in the sections, the first explains the tool and the second the intelligent model. In Figure 5 the proposed methodology is shown.

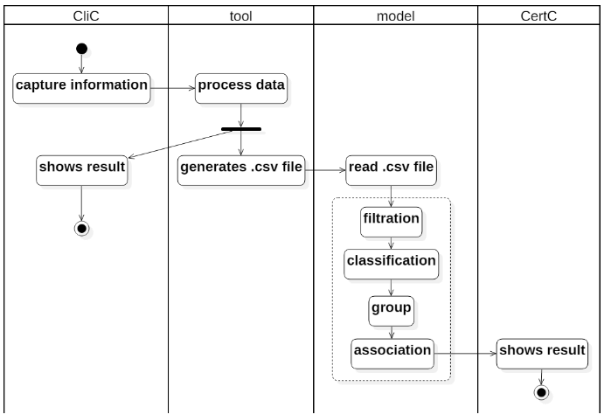


Fig. 5. Methodology.

On CliC Street the client company captures the information in the system. This is where the CMMI questionnaire is answered.

The data is sent to the street tool for processing in the JAVA software. In parallel, the results are sent to CliC and a .csv file is created with the captured information.

The .csv file is read by model and starts the analysis process with machine learning which consists of filtering, classifying, grouping and associating the information of the file.

Finally, the results are shown in CertC.

CliC

It answers a questionnaire of one hundred and twenty questions in order to know its status regarding the good practices of the CMMI level two and three independently of the constellation. The answers given to the questionnaire can be given by more than one person.

The dimensions considered for the questionnaire are: institution: thirteen questions, project management: fifty-four questions, organization: twenty-two questions, engineering: thirty-one questions to give a total of one hundred and twenty questions.

It is expected that the answers will be provided with ample and solid knowledge of what there is and what the company does. The answers must be dichotomous, that is, yes or no. Each response has a same weighting that will determine the level of maturity in each dimension and will be compared with the maximum of each dimension. Table 1 shows the process areas with the weights.

Tool

The proposed tool is based on the JAVA programming language. In this tool the survey is developed. In the Figure 6 the process is shown.

The tool processes the data provided by the CliC represented by the hundred and twenty dichotomous questions. A weight is assigned to each item of the survey. There is a question to request more answers; if there are more, the data is received and reprocessed until there are no more answers.

Table 1. Process areas.

Process area	Top Score
PP	16
PMC	6
MA	4
SAM	7
CM	6
REQM	6
PPQA	3
RD	9
TS	6
GG	13
PI	7
VER	6
VAL	5
OPD	7
OPF	3
OT	6
RSKM	5
DAR	1
IPM	6
TOTAL	122 points

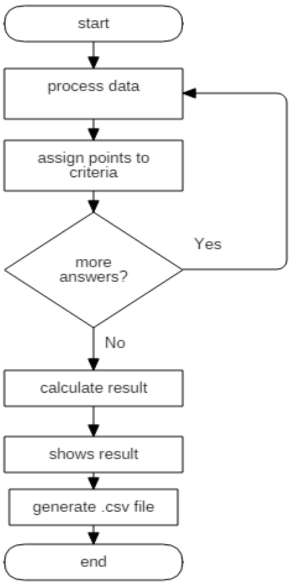


Fig. 6. Process of the tool.

When there is no more, the result is calculated and displayed at the same time that a .csv file is generated that will go to the model process.

The tool has 4 sections, one for each dimension. Figure 7 shows the prototype where it is possible to respond to the questions of each dimension by means of radio button controls to make the response to the reagents more usable and simple.

When the capture of a section is finished, the next button is clicked so that by means of this element it is moved to another section of the form. When you get to the engineering section there is a button to calculate yourself that will perform two actions; the first is to take all the answers and match them with the weighting criteria for each one of them; the second action is to create a .csv file with the results of the survey to send it to the model.

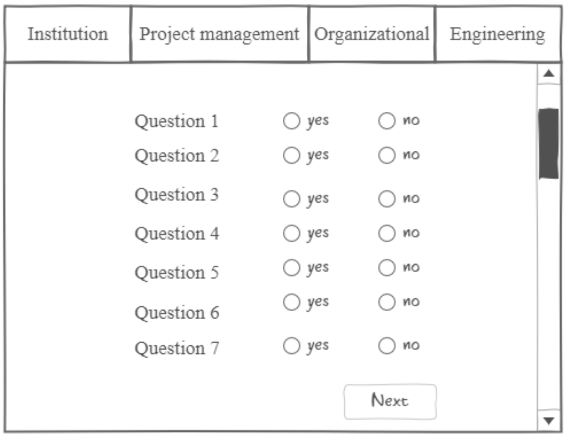
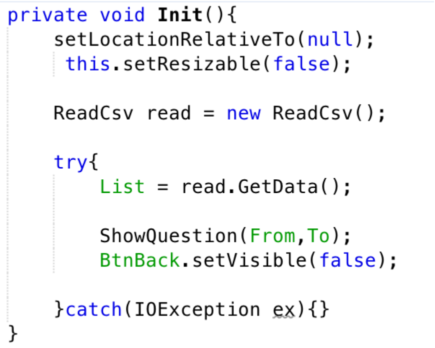


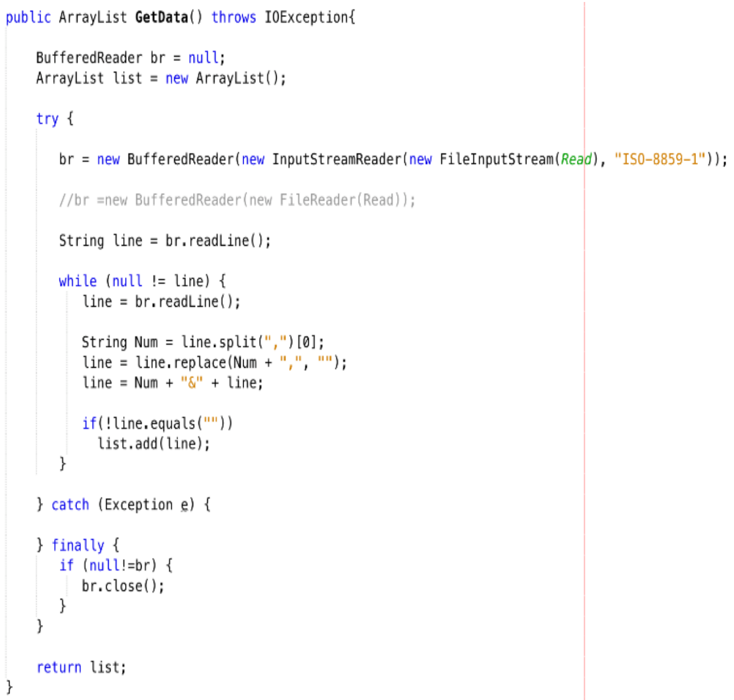
Fig. 7. Prototype.

The general process consists of five methods: Init, GetData, ShowQuestion, CreateRadioButton, CreateQuestions; and two buttons: Next and Back. The Init method positions the window and reads the file where the questions are, to do it in a dynamic way and to be able to read other files with more questions if required.



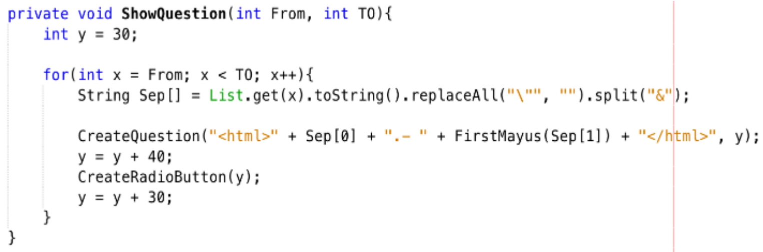
```
private void Init(){
    setLocationRelativeTo(null);
     this.setResizable(false);

    ReadCsv read = new ReadCsv();

    try{
        List = read.GetData();

        ShowQuestion(From,To);
        BtnBack.setVisible(false);

    }catch(IOException ex){}
}
```

The GetData method reads the file in the ISO-8859-1 format (coding of the Latin alphabet, as accented letters). This already read file is saved in an ArrayList to be able to handle it.



```
public ArrayList GetData() throws IOException{

    BufferedReader br = null;
    ArrayList list = new ArrayList();

    try {

        br = new BufferedReader(new InputStreamReader(new FileInputStream(Read), "ISO-8859-1"));

        //br =new BufferedReader(new FileReader(Read));

        String line = br.readLine();

        while (null != line) {
            line = br.readLine();

            String Num = line.split(",")[0];
            line = line.replace(Num + ",", "");
            line = Num + "&" + line;

            if(!line.equals(""))
              list.add(line);
        }

    } catch (Exception e) {

    } finally {
        if (null!=br) {
            br.close();
        }
    }

    return list;
}
```

The ShowQuestion method shows and generates the questions already read from the file with their respective selectable responses in a dynamic way.



```
private void ShowQuestion(int From, int TO){
    int y = 30;

    for(int x = From; x < TO; x++){
        String Sep[] = List.get(x).toString().replaceAll("\"", "").split("&");

        CreateQuestion("<html>" + Sep[0] + ".- " + FirstMayus(Sep[1]) + "</html>", y);
        y = y + 40;
        CreateRadioButton(y);
        y = y + 30;
    }
}
```

Model

The objective of the model is to build a predictive scheme, based on machine learning, for which the process is shown in Figure 8 where it is possible to identify the sections through which the information captured in the tool is transformed.

The process consists of providing the .csv file generated by the tool to the WEKA software in which it is necessary to create a .arff file from the .csv where the structure of attributes and data must be specified in the order of numeric, real, categorical, dichotomous in such a way that if the arrangement of options is shown inside the file.

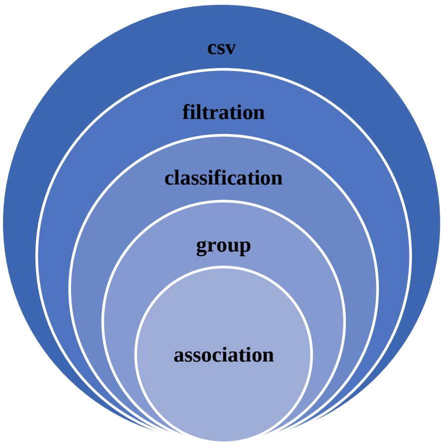


Fig. 8. Model process.

To continue with the filter, WEKA is informed of the location of the .csv file and it is opened. To work in the proper format you need to convert the .csv to .arff. The format that is used to work with the reagents is to assign them a name with q1, q2, q3 up to q120 for the reagents of the CMMI diagnosis for level two and three all as numeric types. The .arff file must contain, in addition to the .csv information, a special structure; @relation, @attribute and @data, as shown below.

```
@relation results

@attribute question {r1,r2,r3,r4,r5}

@attribute q1 numeric
@attribute q2 numeric
@attribute q3 numeric
@attribute q4 numeric
@attribute q5 numeric
@attribute q6 numeric
@attribute q7 numeric
@attribute q8 numeric
@attribute q9 numeric
@attribute q10 numeric

@datar1,0,0,1,0,1,1,0,1,1,1
```

@relation establishes the name of the file, @attribute shows each one of the questions of the questionnaire (for illustrative purposes, only the first ten records of the hundred and twenty are shown), @data contains the information that accompanies each attribute, that is, they are the data that make up the file result .csv.

In WEKA to perform the transformation from .csv to .arff the Arffviewer tool must be used, there the .csv is opened and saved as .arff.

The .arff is opened by the Explorer application and the file is indicated. Once opened, it proceeds with classification, grouping and association.

CertC

The certifying company observes the results of WEKA processing to determine or predict future scenarios. The results obtained are shown in the following section.

5 Results

The results obtained are based on the choice of one of the one hundred and twenty questions of the questionnaire; by convention, the one question that was chosen by the team is the most representative for the purpose of the certification; the question is: 1. Do you have an organizational policy that dictates the discipline of process monitoring?

The classification showed that only 5 of the 5 studied did not have it. It is presumed at this point that more than half do have an organizational policy. The grouping built two groups. The association shows five rules:

1. q73=1 ==> q16=1
2. q86=0 ==> q16=1
3. q96=1 ==> q16=1
4. q112=0 ==> q16=1
5. q86=0 ==> q73=1

Derived from the rules, the following findings are obtained:

1: Who do I establish operational scenarios that help me describe the product specification (flowcharts, presentations, prototypes, demos, sequence diagrams, etc.) ALSO document the life cycle that describes the phases through which these projects go unwinding
2: Those who do NOT document the architecture of our solutions and their design DO document the life cycle that describes the phases through which these projects are evolving.
3: Who do I run tests in customer environments to ensure that I meet their expectations. This may also involve UAT tests or guarantee periods of my solution ALSO we document the life cycle that describes the phases through which these projects are developing.
4: Those who do NOT document the standard work environment for our projects (materials, tools, licenses, corporate software, etc.) DO document the life cycle that describes the phases through which these projects are evolving.
5: Those of us who do NOT check that the technical description of interfaces is complete and well defined IF I establish operational scenarios that help me describe the product specification (flowcharts, presentations, prototypes, demos, sequence diagrams, etc.).

Therefore:
Those who document the life cycle that describes the phases through which these projects are evolving
As well
* I do establish operational scenarios that help me describe the product specification (flowcharts, presentations, prototypes, demos, sequence diagrams, etc.)

But no
* We check that the technical description of interfaces is complete and well defined
* Yes, I run tests in client environments to ensure that I meet your expectations. This may also involve UAT tests or guarantee periods of my solution
But no
* We document the architecture of our solutions and their design
* We document the standard work environment for our projects (materials, tools, licenses, corporate software, etc.)

6 Conclusions

Based on the results, the conclusion of this paper is that:

It is possible that companies can aspire to a certification under a trust scheme because they count on the present proposal with a tool that allows them to know their status.

That certification companies can, based on an analysis scheme, predict the behavior of client companies seeking certification.

Save resources (time, money and effort) by using this proposal.

7 Future Work

As part of the future work it is expected to build the tool an Android so that in this way it is more practical to apply the surveys as well as a possibility to store the historical data in a MySQL database.

It is also expected to have the application in a web environment in the same way with connection to MySQL in order to have the records of all the surveyed companies stored and be able to apply these results in another Big Data scheme.

References

1. Noyel, M., Thomas, P., Charpentier, P., Thomas, A., Brault, T.: Implantation of an on-line quality process monitoring. In Proceedings of 2013 International Conference on Industrial Engineering and Systems Management (IESM), pp. 1–6 (2013)
2. CMMI Institute - Maturity Profile ending 31 December 2017. (s. F.). Retrieved June 2, 2018, from https://cmmiinstitute.com/resource-files/public/quality/maturity-profiles/maturity-profile-ending-31-december-2017
3. PADRON_CENTRO DEVELOPMENT CURRENT_2016_abr-18.pdf. (s. f.) Retrieved June 3, 2018, from https://prosoft.economia.gob.mx/doc/PADRON_CENTRO%20DE%20DESARROLLO%20VIGENTE_2016_abr-18.pdf
4. Chrissis, M., Konrad, M., Shrum, S.: CMMI for Development: Guidelines for Process Integration and Product Improvement, Third Edition. Addison-Wesley Professional (2011)

5. Menezes, W.: Capability Maturity Model Integrated. Encyclopedia of Software Engineering, Volume 1, 1112–120 (2002)
6. Huang, K., Yang, h., King, I., Lyu, M.: Machine Learning. Modeling Data Locally and Globally. Springer (2008)
7. Kubat, M.: An Introduction to Machine Learning. Springer (2015)
8. Mohamed, K.: Machine Learning for Model Order Reduction. Springer (2018)
9. RaschKa, S.: Python Machine Learning. UK, Birmingham: Packt (2015)
10. Eckert, K.B., Suénaga, R.: Analysis of Attrition-Retention of College Students Using Classification Technique in Data Minig. Form. Univ. 8(5), (2015)
11. Lantz, B.: Machine Learning with R. UK, Birmingham: Packt (2013)
12. Rohwer, R., Wynne-Jones, M., Wysotzki, F.: Neural Networks. In: Michie, D., Spiegelhalter, D.J., Taylor, C.C. (eds): Machine Learning, Neural and Statistical Classification, pp. 84–106. Inglaterra, Birmingham (1994)
13. Palacios, H., Porcell, N.: Obstacles when implanting the CMMI model. Bogotá (2008)

ISSN 1870-4069

Realtime Recoloring Objects using Artificial Neural Networks through a Cellphone

Martín Montes Rivera[1], Alejandro Padilla[2], Juana Canul[3], Julio Ponce[2], Alberto Ochoa Zezzatti[3]

[1] Universidad Politécnica de Aguascalientes, Mexico
[2] Universidad Autónoma de Aguascalientes, Mexico
[3] Universidad Juárez Atónoma de Tabasco, Mexico
martin.montes@upa.edu.mx, apadilla2004@hotmail.com, juana.canul@ujat.mx,julk.ponce@gmail.com, alberto.ochoa@uacj.mx.

Abstract. Recoloring it is a technique for changing the color an image resulting in a different new one. Recoloration is a common photo edition operation since digital images are around every media resource and several algorithms are used for editing these pictures, nevertheless, recent digital cameras have increased enormously the quantity of pixels for producing them. This increase in the size of digital images makes difficult the recoloring operation. In order to solve the recoloring problem, there had been applied several algorithms, some algorithms directly detect the color by performing transformations on color representations to different spaces where color is easily separated but this transformation require several no linear operations. On the other hand, numerical parameters on CNNs make than this approach cannot be trained or implemented on a mobile device, more over the time required for computing an input image will made that the processed pictures be delayed continually. Considering this limitation is proposed a specific short architecture for detecting a specific color in general objects using a feedforward neural network trained with gradient descent backpropagation with variable learning rate.

Keywords: realtime recoloring images, artificial neural networks, gradient descent backpropagation.

1 Introduction

Recoloring it is a technique for changing the color or the theme of an image resulting in a different new one with almost imperceptible changes for the human eye. Recoloration is a common photo edition operation since digital images are around every media resource and several algorithms are used for editing these pictures. Nevertheless, recent digital cameras have a great quality and definition, increasing enormously the quantity of pixels for producing them. This increase in the size of digital images makes difficult the recoloring operation, especially when is required to be performed in real time, like is desired in an augmented reality device (Yan, Ren, & Cao, 2018).

Augmented Reality (AR) is an emerging technology where real elements are added with artificial elements, allowing overlapping virtual and real environments, the use of

pp. 229–238; rec. 2018-09-13; acc. 2018-10-07

AR is not limited to visual content, but it is its most common representation using displays, cameras and sensors for the experience. This is been applied to several areas, like games, industry, education, communication, tourism, marketing, among others (Tussyadiah, Jung, & tom Dieck, 2017).

In order to solve the recoloring problem there had been applied several algorithms, some algorithms directly detect the color by performing transformations on color representations to different spaces where color is easily separated and noise or other kind of variations like luminance are suppressed, some of this transformations include change to CIE L*a*b, HSV, HSM, among others, but the required operations for completely transform the image demands high computational power due to no linearities in the transformations and the required operations (Montes Rivera, Padilla Díaz, & Ponce Gallegos, Comparative between RGB and HSV color representations for color segmentation when it is applied with artificial neural networks and evolutionary algorithms, 2016).

Despite of the time required for changing a color representation in an image it is possible to determine its color only by using labeled input data with images of the color to detect like is described in (Montes Rivera, Padilla, Canul, Ponce, & Ochoa, 2018), where several colors can be classified without using any color transformation.

The detected colors are classified with specific linear equations that describe the color. Coefficients in the equations are updated depending from the color to classify these numerical parameters which are obtained using comparatively Particle Swarm Optimization (PSO) and Genetic Algorithms (GA), despite of the results in (Montes Rivera, Padilla, Canul, Ponce, & Ochoa, 2018) working with linear equations implies that colors are linearly separable but in several cases this will limit the algorithm.

Another technique that has exhibit great results recoloring pictures in Artificial Intelligence (AI) are Artificial Neural Networks, where recently its limitation as linear classifiers has been improved with several hidden layers, Graphics Processing Unit (GPU) optimization of training algorithms, regularization in training algorithms, the increasing access to huge labeled data sets and the improvement of processing hardware (Goodfellow, Bengio, & Courville, 2016).

The Work in (Levinshtein, et al., 2017) performs Realtime recoloring of hair using a convolutional neural network (CNN) for detecting the hair and then performs a color transformation in the detected hair, the performance reported is of 300 ms per photogram for mobile devices but the required time for training the large amount of data in a CNN makes than this approach cannot be trained on a mobil device, more over the time required for computing an input image will made that the processed pictures be delayed continually.

On the other hand, Artificial Neural Networks with few hidden layers would produce good results if labeled input data is correctly introduced and algorithms for training are well optimized like shown in (Goodfellow, Bengio, & Courville, 2016).

Considering this alternative is proposed a specific short architecture for detecting a specific color in general objects using a feedforward neural network trained with input data of the pixels of the objects that must be transformed, transformation is reached in real time decreasing the time for computing per photogram and maintaining quality during recoloring.

This paper is organized as follows, section 1 describes the problem and general ideas in the state of the art, section 2 describes concepts required for implementing the methodology in section 3, results are shown in section 4 and section 5 expose the conclusions of this work based on the obtained results.

2 Theoretical Framework

ANNs are mathematical models that was inspired by the human nervous system, they can solve complex tasks by learning numerical values that represent its synapsis and have been used since the early 50s (Engelbrecht, 2007). The very first architectures were highly limited until backpropagation allow to train hidden layers allowing to separate no linearly separable classes in a supervised training algorithm which is one of the most popular algorithms used today in commercial applications (Goodfellow, Bengio, & Courville, 2016).

Biological neurons have inputs called dendrites and a single output called axon with on or off state depending from the dendrites and its soma which is the body of the cell that classify those inputs. Synapse is the number of dendrites linked to a specific path. ANNs first architecture was proposed by McCuloch and Pits in 1943 has inputs x_i and output y and its synapse is reproduced with w_i weights, like is shown in the figure bellow and its activation function representing the soma is the shown in equation (1) (Nguyen, Prasad, Walker, & Walker, 2003):

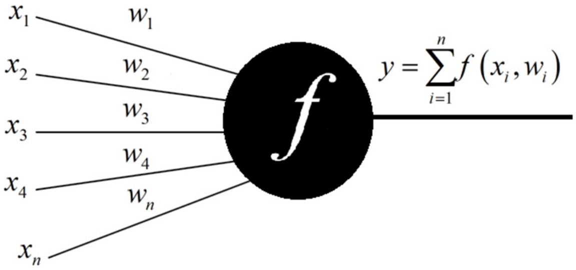


Fig. 1. McCulloch and Pits ANN.

$$F(z) = \begin{cases} 1 & z \geq \theta \\ 0 & z < \theta \end{cases} \tag{1}$$

z is the total input to the neuron multiplied by w_i weights, like in equation (2):

$$z = \sum_{i=1}^{n} x_i w_i \tag{2}$$

Despite of McCuloch and Pits neuron was the first introduced, was hardly limited because in it is a binary neuron that could not have several layers like occurred with the first ANNs and it could separate only linear classes (Goodfellow, Bengio, & Courville, 2016).

Backpropagation Gradient Descent is the most popular learning rule that allows to train ANNs with hidden layers, and it was initially used a with a sigmoid activation function with outputs in range $0 \leq y \leq 1$, In this work is used a hyperbolic sigmoid activation function (3) for hidden layers since this function accelerates the convergence of backpropagation method:

$$\tan sh(\mathrm{z}) = \frac{2}{\left(1+e^{-\mathrm{z}}\right)-1} \tag{3}$$

$$linear(z) = z \tag{4}$$

The proposed architecture is trained using Backpropagation algorithm learning rule updating weights with equations (5) and (6):

$$w_{ij} = w_{ij} + \alpha x_i \delta_j \tag{5}$$

$$w_{jk} = w_{jk} + \alpha y_j \delta_k \tag{6}$$

With $\delta_k = \frac{\partial y_k}{\partial x_k} e_k$ and $\delta_j = \frac{\partial y_j}{\partial x_j} \sum_{k=1}^{n} w_{jk} \delta_k$ with $e_k = d_k - y_k$ where d_k is the desired output in the y_k output and k is the index of the number of outputs, i is the index for the number of connection in the j layer of the neuron.

3 Methodology

The color selected for classification must be photographed in the object, then all pixels in the image M are marked as white or equal to 1, in a modified image M^* in bright of pixels, i.e. pixels take value of 1 if they belong to the color C or reduce its value for a correct classification if they do not belong, like is shown in equation (7):

$$M^*(\mathrm{i},\mathrm{j},\mathrm{c}) = \begin{cases} 1 & M(\mathrm{i},\mathrm{j},\mathrm{c}) \in \mathrm{C} \\ 0.9M(\mathrm{i},\mathrm{j},\mathrm{c}) & M(\mathrm{i},\mathrm{j},\mathrm{c}) \notin \mathrm{C} \end{cases} \tag{7}$$

Output vector for training is obtaining by scanning values of M^* if they are 1 then output is equal to 1 if they do not then output is equal to 0:

$$O = \begin{cases} 1 & M(\text{i},\text{j},\text{c}) = 1 \\ 0 & M(\text{i},\text{j},\text{c}) < 1 \end{cases} \tag{8}$$

The proposed architecture (Fig. 2) then is trained with the input data M and the desired output data O, using those transfer functions suggested in section 2.

ANN output is maintained as line for having a scalar value indicating a percentage of security of activation when 1 will represent 100%, since desired output is 0 or 1 depending from M^*.

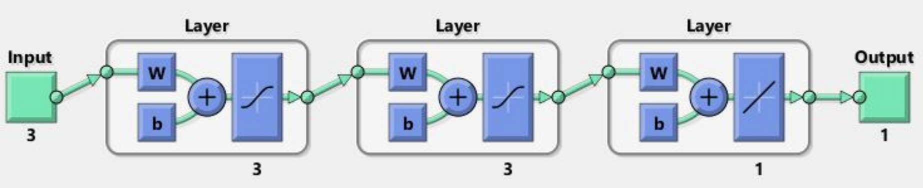


Fig. 2. Proposed Architecture for recoloring objects of a specific color.

After that colors of pixels detected by neural network are transformed by following and equation depending from its RGB values equation (7):

$$\begin{bmatrix} NR \\ NG \\ NB \end{bmatrix} = \begin{bmatrix} \text{H}(R,G,B) \\ \text{I}(R,G,B) \\ \text{J}(R,G,B) \end{bmatrix} \tag{9}$$

4 Results

Some of images used as input data are shown below, the images labeled by the user (M^*) are shown in the first row of Fig. 3 and the original images (M) are in the second row Fig. 3 as described in equation (7)

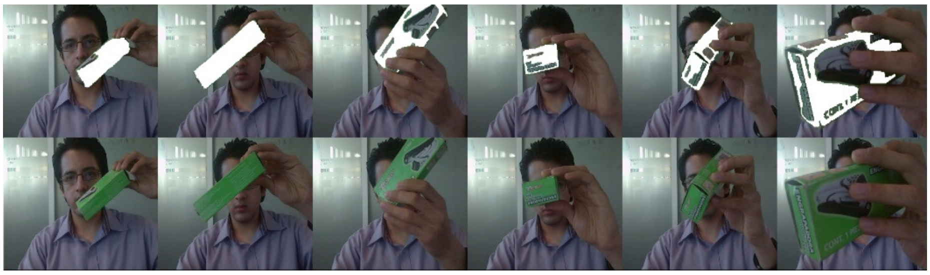


Fig. 3. Training input example for color transformation.

Then the ANN with architecture in Fig. 2 is trained using the input M in and the output O obtained from M^* using equation (8).

The training response for green color using the proposed method is shown in Fig. 4.

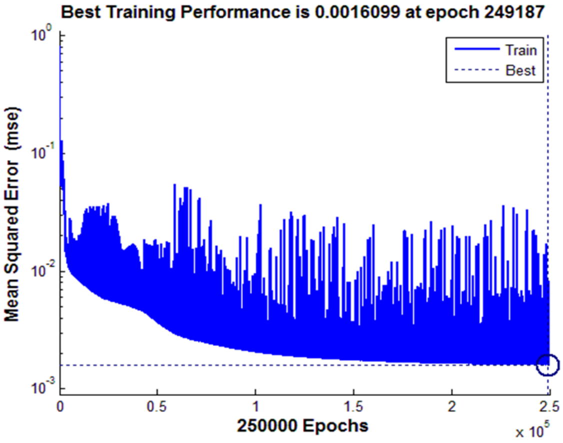


Fig. 4. Training response for ANN proposed.

The processed images presented as results are generated with Matlab^{TM} using a PC with Core I7 6700 3.4 Ghz processor and parallel computing toolbox using NVIDIA GeForce GTX 970, the parallelizing allows to process every pixel in the images with 960x720 resolution reaching 8.1483 Frames Per Second (FPS), the images are acquired using a Samsung^{TM} S Note 8 cellphone wireless connected to the computer using DroidCam^{TM} app, for sending the images to the computer.

Figures from 5 to 9 shows a comparison between processed image and original transforming its color using $R = 0.7G + R \quad B = B$, $G = 0$ as functions described in equation (9), maintaining an equation structure depending from color makes possible to retain bright of colors and shadows.

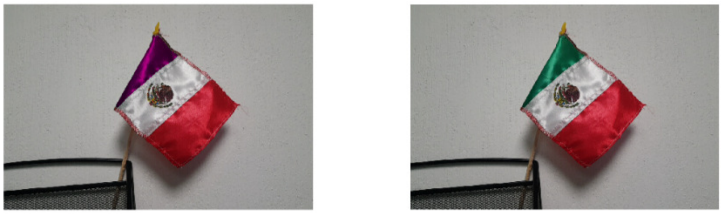


Fig. 5. Result 1 recoloring image (right processed image, left original image).

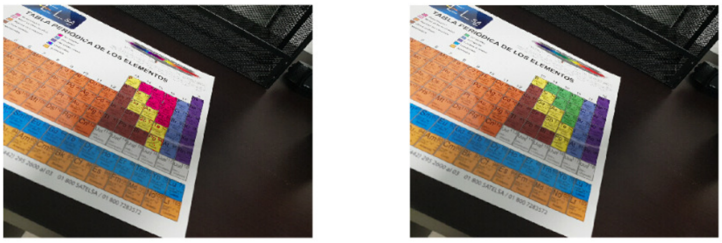


Fig. 6. Result 2 recoloring image (right processed image, left original image).

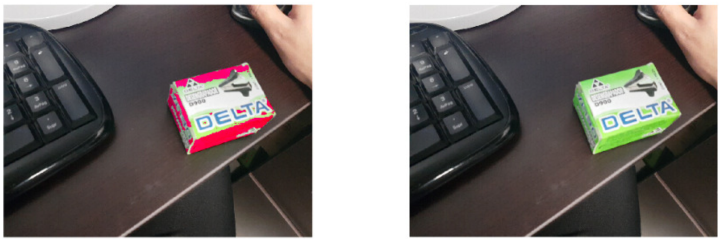


Fig. 7. Result 3 recoloring image (right processed image, left original image).

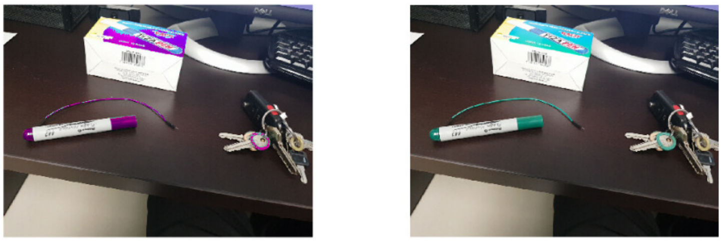


Fig. 8. Result 4 recoloring image (right processed image, left original image).

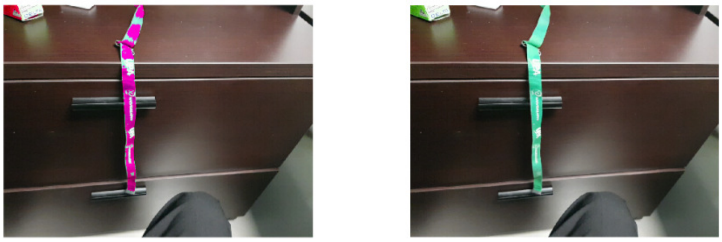


Fig. 9. Result 5 recoloring image (right processed image, left original image).

Table 1. Comparative of number of layers for recoloring images.

Work described	Type of ANN	Layers	Accuracy
(Baldassarre, Gnzalez Moríın, & Rodés-Guirao, 2017)	CNN	14	80%
(Zhang, y otros, 2017)	CNN	27	Uses predicting gray for every pixel (PSNR 22.8)
(Ci, Ma, Wang, Li, & Luo, 2018)	CNN	12	Uses Mean opinion score (MOS)
This work	Feedforward ANN	2	Only training measure 1.6e-3 Mean Square Error (MSE)

This work uses a feedforward ANN with the architecture described in Fig 2. For recoloring a specific color and like is described in table 1 its comparison with other architectures is minimal making that this ANN become a good model specially when computational power is reduced and real-time is required, nevertheless some of the other architectures work with several colors in the same ANN, but require deep Convolutional Neural Networks (CNNs).

5 Conclusions

In this paper is presented a recoloring technique for fast computation that runs real time recoloration with pictures acquired from a cellphone and processed using a feedforward neural network with only 2 layers, the labeled input colors are selected by the user for its change and then are trained in the neural network using gradient descent backpropagation with variable learning rate.

The recoloring task is performed with a very short neural network compared with those in the state of art detailed in table 1.

Despite of the reduced ANN architecture there are not important errors on the detection of the trained color like is shown in figures 5 to 9, however, there must be evaluated the efficiency of the algorithm comparing pictures received against pictures desired in order to measure the possible error.

The proposed technique allows the user to train a pattern color recognition by only selecting the color that must be detected, the training does not require great computational power since only 2 layers are needed, moreover, with the proposed equation (9) is possible to transform the object only in color but maintaining original bright and shadows.

Since the processing speed remains only in 8.1483 FPS for images with 960x720 resolution, a different platform from Matlab^{TM} with lower level laguage must be tested, nevertheless, getting this behavior without deep neural networks like in other works, allows to perform real-time recoloration, that could be used in augmented reality devices, and even in devices with less computational power than a computer like cellphones.

5.1 Future Work

Development of application in C++ interface in order to directly developing a CUDA^{TM} kernel and perform its adaptation to the android system using openGL^{TM} for parallelizing the colors detection and adapting the algorithm for processing and acquiring completely in the cellphone.

Perform a cross validation where desired output images not used during the training be taken for cross validation and efficiency evaluation.

Apply the recoloring technique here proposed in a real time application using a cellphone for recoloring objects that are difficult to perceive for colorblind people and assist them by modifying colors in Ishihara plates test, helping them to pass it or testing this situation in real environments where is required the modifying of a color for perceiving it.

References

1. Baldassarre, F., Gnzalez Morı́ın, D., & Rodés-Guirao, L. (2017). Deep Koalarization: Image Colorization using CNNs and Inception-Resnet-v2. arXiv preprint arXiv:1712.03400.
2. Ci, Y., Ma, X., Wang, Z., Li, H., & Luo, Z. (2018). User-Guided Deep Anime Line Art Colorization with Conditional Adversarial Networks. arXiv preprint arXiv:1808.03240.
3. Demuth, H., & Beale, M. (1998). Neural Network Toolbox for Use with MATLAB. Massachusetts: MathWorks Inc.
4. Engelbrecht, A. P. (2007). Computacional Intelligence. Sudafrica: Wiley.
5. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. http://www.deeplearningbook.org: MIT Press.
6. Levinshtein, A., Chang, C., Phung, E., Kezele, I., Guo, W., & Aarabi, P. (2017). Real-time deep hair matting on mobile devices. arXiv:1712.07168, 1-7.
7. Montes Rivera, M., Padilla Díaz, A., & Ponce Gallegos, J. (2016). Comparative between RGB and HSV color representations for color segmentation when it is applied with artificial neural networks and evolutionary algorithms. In D. A. M. en C. Ma. de Lourdes Sánchez Guerrero, Avances en las Tecnologías de la Información (pp. 611-629). Ciudad de México: ALFA-OMEGA.
8. Montes Rivera, M., Padilla, A., Canul, J., Ponce, J., & Ochoa, A. (2018). Comparative of Effectiveness When Classifying Colors Using RGB Image Representation with PSO with Time Decreasing Inertial Coefficient and GA Algorithms as Classifiers. In O. Castillo, P. Melin, & J. Kacprzyk, Fuzzy Logic Augmentation of Neural and Optimization Algorithms. Studies in Computational Intelligence 749. (pp. 527-546). Springer.
9. Nguyen, H. T., Prasad, N. R., Walker, C. L., & Walker, E. A. (2003). A First Course in Fuzzy and Neural Control. United States of America: Chapman and Hall.
10. Singh, U. P., & Jain, S. (2018). Optimization of neural network for nonlinear discrete time system using modified quaternion firefly algorithm: case study of Indian currency exchange rate prediction. Soft Computing, 8(2667–2681).
11. Tussyadiah, I. P., Jung, T. H., & tom Dieck, M. C. (2017). Embodiment of Wearable Augmented Reality Technology in Tourism Experiences. Journal of Travel Research, 1-15.
12. Valeriy Dubrovin, S. S. (2000). Neural Network Method in Plant Spectral Recognition. En R. S. Muttiah, From Laboratory Spectroscopy to Remotely Sensed Spectra of Terrestrial of Terrestrial Ecosystems (págs. 147-159). Kluwer Academic Publishers .
13. Vogl, T., Mangis, J., Rigler, A., Zink, W., & Alkon, D. (1988). Accelerating the convergence of the backpropagation method. Biological Cybernetics, 257-263.
14. Yan, Y., Ren, W., & Cao, X. (2018). Recolored Image Detection via a Deep Discriminative Model. IEEE Transactions on Information Forensics and Security, 14(1), 5 - 17.

15. Zhang, R., Zhu, J.-Y., Isola, P., Geng, X., Lin, A. S., Yu, T., & Efros, A. A. (2017). Real-time user-guided image colorization with learned deep priors. arXiv preprint arXiv:1705.02999.

ISSN 1870-4069

Recognition of Colors through Use of a Humanoid Nao Robot in Therapies for Children with Down Syndrome in a Smart City

Martha Jiménez[1], Alberto Ochoa[1,2], Daniela Escobedo[1], Ricardo Estrada[1],
Erwin Martinez[1], Rocío Maciel[2], Víctor Larios[2]

[1] Universidad Autónoma de Ciudad Juárez, Mexico
[2] Centro de Innovación en Ciudades Inteligentes, CUCEA-UdG, Mexico
al148887@alumos.uacj.mx, alberto.ochoa@uacj.mx

Abstract. Down syndrome is the most frequent cause of mental disability, presenting similar characteristics among people with this syndrome, among them the scarce short-term memory capacity, fatiguing attention, language delay, among others. In Mexico only 3% of Down children receive education, so in this section we propose the use of a Humanoid robot for future application in therapies for children with Down syndrome. We propose how to improve your ability to work with colors and shapes in a group of children with Down syndrome.

Keywords: color identification, down syndrome, therapies with children with disabilities, NAO humanoid robot.

1 Introduction

Today the number of related genetic abnormalities is quite high; to mention that there are more than 12,000 described genetic syndromes, on of them is Down Syndrome. In 1866 it was John Langdon Down who described for the first this syndrome, and on that year, it was attributed to a delay in the normal development. Subsequently it was stated that Down Syndrome was a consequence of infectious processes, alcoholism, between others. It was until 1958 when Jerome Lejeune and Pat Jacobs, discovered the presence of a third chromosome in the 21 pair of all cells, since then it's considered a genetic syndrome. Being the first genetic abnormality described on human beings.

Down Syndrome creates deficiencies in physical and intellectual development among individuals. Children with Down Syndrome usually show physical characteristics, neuropsychological, sensorial, motor and cognitive quite similar. The last are described by Fernandez Sampedro and others (1993):

- Attention is unstable, scattered and fatigable; to make a progress in the child's learning, it must involve different activities to keep it.
- Short term memory lacks of sensorial & hearing information processing which usually helps to improve with visual efforts. Regarding long term memory, there's a challenge to store and recover information. The child retains memory

pp. 239–252; rec. 2018-09-15; acc. 2018-10-07

by observation because of habit, but lacks of memory, which hardens language and vocabulary learning.

- There's a considerable retardation language wise in comparison to other development fields. There's a gap between comprehension and expression levels.
- Vocabulary delay is present, possibly because there's a lack of comprehension in the relation between objects, people, facts, and words that represent them, there's no retention of this relationship or there's not a space-time frame acknowledgement.
- Slowness in reaction timing. Children with Down Syndrome also show impulsivity, low tolerance to frustration, lack of persistence on tasks, low motivation innate and the need of external supervision to finish a task.

Although all the deficits are marked, some authors state that, when therapy programs integrate motivational aspects, there's an improvement of their intellectual execution. Therefore, it is stated that they possess some learning skills, but, in comparison to their No-Down counterparts, they show instability in knowledge attainment, and this last one is slower.

1.1 Education Problems Faced by Down Syndrome Kids

Down Syndrome is the most frequent cause of psychic disability, the estimated relation in cases is one in every 1000 - 1100 births, according to the World's Health Organization. In Mexico, the Health Department, through the Gender Equality Center and Reproductive Health, in their Technical Guidelines for the Integral Attention for people with Down syndrome, estimates that there's a case in every 650 first borns.

In Mexico there are many institutions that support children with Down syndrome, generally there are specialized institutes for motor development and others in intellectual development. However, most of the public schools since kindergarten won't take in kids with different intellectual skills due to lack of prepared personnel to provide proper education to kids with DS, which leads to, mostly, kids with DS to not receive education. In Mexico only 3% of kids with DS receive education. The main reason is the number of educators and equipment in specialized institutions are not enough to cover the number of children with this syndrome and as a secondary reason, Mexico is a country with low resources that's why it's complicated to take a kid to a specialized institution, and there might be the case that this institution is a few hours away from home.

1.2 NAO Humanoid Robot

The Robotics is a modern technology, which includes the study, design and operation of robots, as well as researches for their further development. The definition given to the robot by the Robot Institute of America is: "A reprogrammable, multifunctional manipulator designed to move material, parts, tools, or specialized devices through various programmed motions for the performance of a variety of tasks". Throughout the last years, there has had a great growth on the interest in robots which are able to complete the task of a human assistance. They have been developed to be an aid in homes, for sanitary and therapeutic assistance. The use of the Nao humanoid robot is proposed to be implemented in the education for kids with DS, altogether with parents and professors who may teach them daily.

The robot has two cameras, four microphones, nine tactile sensors, two ultrasonic sensors, eight pressure sensors to perceive the environment where it interacts. It has twenty-five grades of liberty which allow the robot has greater movement range making it ideal for any environment. Also, it has a voice synthesizer and two speakers that permit the communication between the robot and the user as it is described on Figure 1.

The main objective of this article is to develop certain cognitive skills through the use of the Nao humanoid robot, essentially color recognition and motor coordination. Besides, it will try to motivate the Down infant to speak and collaborate, since these are common social abilities that every infant should learn during his/her development.

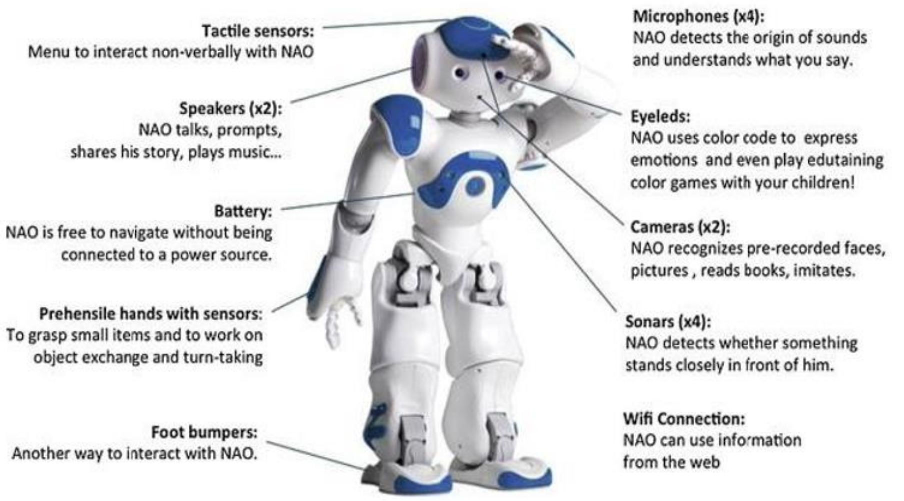

Tactile sensors:
Menu to interact non-verbally with NAO
Speakers (x2):
NAO talks, prompts, shares his story, plays music...
Battery:
NAO is free to navigate without being connected to a power source.
Prehensile hands with sensors:
To grasp small items and to work on object exchange and turn-taking
Foot bumpers:
Another way to interact with NAO.
Microphones (x4):
NAO detects the origin of sounds and understands what you say.
Eyeleds:
NAO uses color code to express emotions and even play edutaining color games with your children!
Cameras (x2):
NAO recognizes pre-recorded faces, pictures , reads books, imitates.
Sonars (x4):
NAO detects whether something stands closely in front of him.
Wifi Connection:
NAO can use information from the web

Fig. 1. Nao Sensors and Joints.

Source: Admin; Features of Nao Robot, *Gigabotics. Robotics Development and Research.* gigabotics.com/robotics/features-of-nao-robot/. (2015). Last accessed 2018/03/25.

There are works that demonstrate the viability of the interaction human or child, specifically, with a robot. One of those was capable to reduce intransitive gesture in preschoolers with autism spectrum disorder, based on a therapy with a social robot.

It has been achieved by the illustration process to recognize all types of colors through the Nao camera, which will mention the color of toys every time the user shows a figure to the humanoid. Moreover, it will be placed each piece inside a container at the beginning or at the end of a therapy session in hand with the humanoid robot to storage the used objects as part of a game.

2 Applications of NAO to Help Children with Down Syndrome

2.1 Color Recognition

In order to make the recognition, the humanoid was placed in front of a white table which measures 15x41x25 cm (WxLxH). As it has been said, the Nao humanoid owns two HD cameras, in this case, the making decision was to use the superior camera due to the position of the robot at the beginning of the action. But any of both may be taken. Regarding the objects, four figures of color red, blue, purple and orange were utilized.

For the image processing, Numpy module was used since it allows matrix working in Python and OpenCV, a computerized visionary library. Those were imported to Python 2.7 Software. The first step was to make the connection of the Nao camera through the computer using the following code:

Algorithm 1. Code to inicialize NAO's superior camera.

```
Dispositivo_Video = ALProxy ('ALVideoDevice', "IP", 9559)

AL_kTopCamera = 0

AL_kQVGA = 1

AL_kBGRColorSpace = 13

captureDevice = Dispositivo_Video.subscribeCamera("Prueba",
AL_kTopCamera, AL_kQVGA,

AL_kBGRColorSpace, 10)

Ancho = 320

Altura = 240

Imagen = np.zeros((Altura, Ancho, 3), np.uint8)
```

The above code specifies the use of the superior camera, space of RGB color and a resolution of 320x240. Each pixel is saved on a new matrix of the same resolution. Once the matrix is obtained, the space of BGR color has to be change to HSV color. The model HSV is based on cylindrical coordinates and it is derived from RGB model. In addition, it represents colors combining three values which are: hue, saturation and value. The first one let color distinguishes each other, taking in account the length of the wave that goes from 0º to 360º.

Nevertheless, the saturation refers to a sensation that goes from high to low color intensity. To finalize, the value refers to the amount of white that a color may contain. Previous description is shown in Figure 2 below.

The next step has the purpose to construct color masks. In this case, orange, red, blue and purple masks were created as the example in algorithm 2. Segmentation is a key concept in digital image processing. The segmentation of an image involves the detection, by means of deterministic or stochastic labeling procedures, of the contours or regions of the image, based on intensity information and / or spatial information.

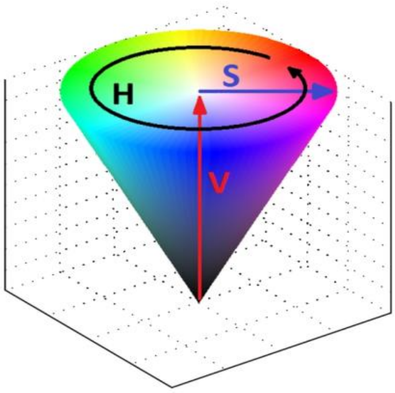


Fig. 2. HSV Model as an Inverted Pyramid.

Source: HSV Color Model, *Wikipedia.* https://es.wikipedia.org/wiki/Modelo_de_color_HSV. Last accessed 2018/03/25.

Algorithm 2. This code shows how to make the purple color mask.

```
mascara_morado = cv2.inRange(hsv, morado_bajos, morado_altos)
```

Then, each mask experienced open and closed morphological transformations. The first transformation consists of converting to zero all pixels from the illustration that do not contain completely the structured element in its surrounding (erosion). That means, getting matrix B and matrix A, as it is demonstrated on Figure 3, they may overlap each other. If all pixels from matrix B intersects with matrix A, they will be storage as ones on a new matrix, on the contrary, they will be storage as zeros.

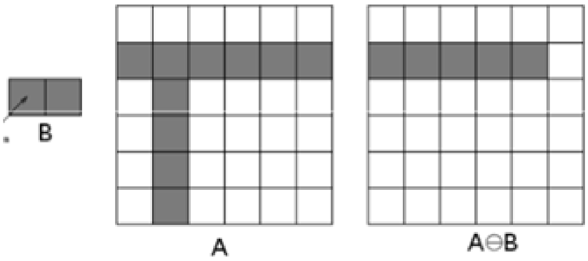


Fig. 3. Erosion example.

Source: Grupo de Topología Computacional y Matemática Aplicada, http://alojamientos.us.es/gtocoma/pid/tema5-1.pdf. Last accessed 2018/03/25.

The second step of the open transformation is dilatation. It consists in overlapping matrix B and matrix A, in difference of erosion, on this phase the first one pixel from matrix B that intersects with matrix A, will storage all ones' pixels from matrix B on a new matrix as it is shown on Figure 4. The closed transformation applies, first, dilatation on hand of the erosion. But, it should be clear that these are not contrary operations.

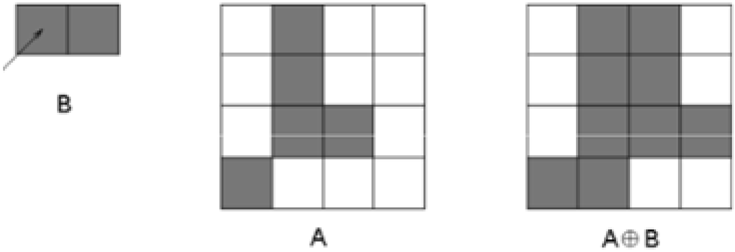


Fig. 4. Dilatation example.
Source: Grupo de Topología Computacional y Matemática Aplicada, http://alojamientos.us.es/gtocoma/pid/tema5-1.pdf. Last accessed 2018/03/25.

The combination of open and closed operations could use them for the filter and segmentation of the image. Also, the opening and closing by reconstruction try to avoid the creation of new information. Next, OpenCV function “countNonZero” is used to follow a counting of all ones from the matrix of each color, which are storage on a variable to compare them among masks. However, the variable with higher value will indicate the color that is on observation. Ultimately, “ALTextToSpeech” was used to program the robot with the objective of saying the next phrase: “El objeto es -color-”:tts = ALProxy('ALTextToSpeech', "IP", 9559) tts.say("El objeto es azul").

2.2 Movement

For the movement the easiest way to make it is using the software Choregraphe at first you will need to save the movements performed and latter these can be used in conjunction with another program like the color recognition one to make the NAO act in a certain way depending of the needs of the problem. For example:

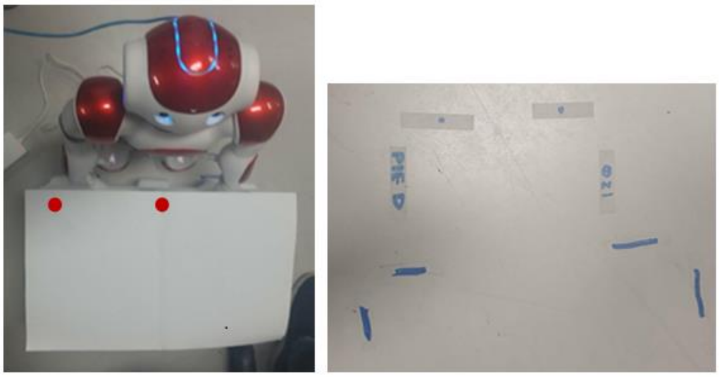


Fig. 5. Nao Location, figure marks and bottle marked with red.

In order to accomplish the movement of toy pieces and storage them in a container, Choregraphe 2.1.4.14 Software was managed. To achieve this goal, the same white table was used to locate the humanoid from a distance of 5 cm between its feet and the table.

On the left, the piece is located and on the right the bottle is placed, a cylinder bottle of 14 cm of height and 7.5 cm of diameter was used and a mark, as shown in Figure 5, was painted to point the place where pieces are located for the humanoid achieves to pick each piece and place it inside the container that also has a specific place to locate it.

3 Results

The illustration process worked since figures of different colors were divided into sections and noticed them with a small error percentage, depending of the illumination, as it is shown below on Figure 6.

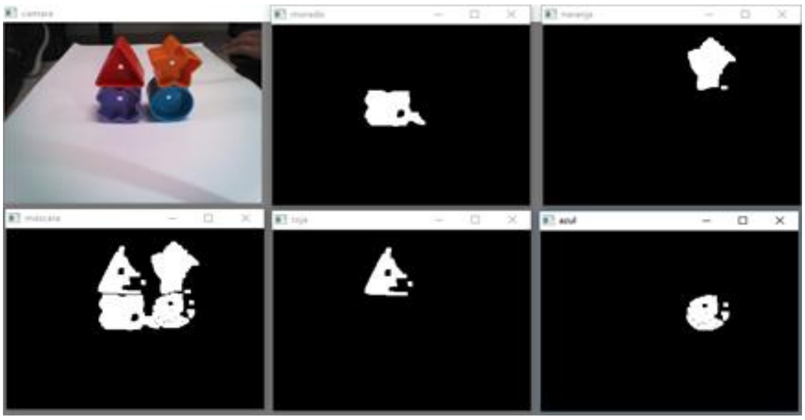


Fig. 6. Color Masks taking with Nao camera. A) Original image, B) Purple Color Mask, C) Orange Color Mask, D) Masks of four colors, E) Red Color Mask and F) Blue Color Mask.

So next we are going to show the code used to differentiate the colors using the NAO the result may vary a little bit this code is the one used for the results above:

Algorithm 3. Code to add the libraries and define the hsv values for each color.

```
import sys
import cv2
import numpy as np
from naoqi import ALProxy
naranja_bajos = np.array([5,150,50],dtype=np.uint8)
naranja_altos = np.array([15, 255, 255], dtype=np.uint8)

morado_bajos = np.array([110,50,50], dtype=np.uint8)
morado_altos = np.array([150, 255, 255], dtype=np.uint8)
```

```
rojo_bajos = np.array([400,100,100], dtype=np.uint8)
rojo_altos = np.array([1000, 500, 500], dtype=np.uint8)

azul_bajos = np.array([80, 50, 50], dtype=np.uint8)
azul_altos = np.array([104, 255, 255], dtype=np.uint8)
```

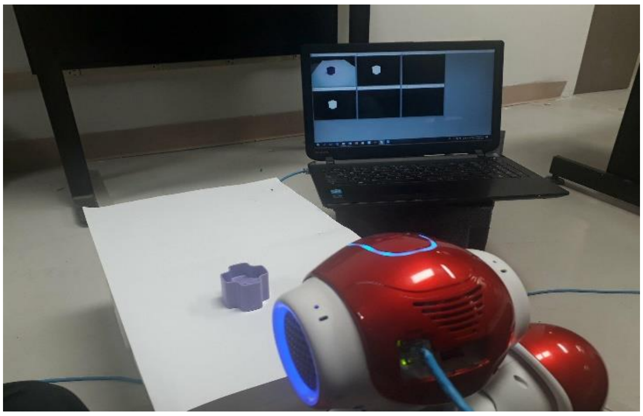


Fig. 7. Image of Nao Humanoid Processing Color Recognition. Once a piece was located on the table, after a pair of seconds, the robot mentioned the color of the figure. Nevertheless, there was not necessity to take screens or restart the software since recognition was executed in real time.

First the first four lines we define the libraries we are going to use for this program. Second we are going to define the arrays of colors to separate each color for this you need to browse in the internet for the colors you want to identify, when you have the spectrum you also need to check for the light in the environment you are at cause this may affect the way colors are seen, so a further adjustment will be needed. Make sure to put the lower spectrum of the color and the higher one.

Algorithm 4. Code to set the artificial vision trough NAO’s superior camera.

```
# Use the Proxy model of NAOqi.169.254.17.106
Dispositivo_Video = ALProxy('ALVideoDevice', "169.254.17.106",
9559) # Connect to NAO.
tts = ALProxy('ALTextToSpeech', "169.254.17.106", 9559)
# Use superior camera.
AL_kTopCamera = 0 # Superior Camera.
AL_kQVGA = 1 # Resolution 320x240.
AL_kBGRColorSpace = 13 # Color space BGR
captureDevice = Dispositivo_Video.subscribeCamera( # Use of the
camera
"Prueba", AL_kTopCamera, AL_kQVGA, AL_kBGRColorSpace, 10) # 10
fps.Ancho = 320
Ancho = 320
Altura = 240
Imagen = np.zeros((Altura, Ancho, 3), np.uint8)
```

Next we are going to connect to the NAO using the proxy in it the one in the code is for the NAO used, so this may vary from NAO to NAO. Be sure to use the one in your robot in order to connect successfully. The line that follows is to active the superior camera and then the resolution of the image is defined and the space for RGB is selected. Then the image is captured in the selected camera. The variables “Ancho” and “Altura” are the ones that define the resolution being “Ancho” equal to 240 and the “Ancho” equal to 320 this is the number of pixels taken for the image so the resolution is 320x240 if higher resolution is wanted you need to adjust these valors. This also make the program slower because of the bigger array and the need to process a bigger image.

Algorithm 5. In this code the matrix of the images taken trough the camera.

```
while True:
    # Obtener imagen.
    Resultado = Dispositivo_Video.getImageRemote(captureDevice);
# Tomar una captura.

    if Resultado == None:
        print 'No se pudo capturar.'
        break
    elif Resultado[6] == None: # En Resultado[6] se guarda la
imagen en formato binario.
        print 'No se obtuvieron datos de imagen.'
        break
    else:

        # Acomodar valores binarios en una matriz para un forma-
to de imagen.
        Valores = map(ord, list(Resultado[6]))
        i = 0
        for y in range(0, Altura):
            for x in range(0, Ancho):
                Imagen.itemset((y, x, 0), Valores[i + 0])
                Imagen.itemset((y, x, 1), Valores[i + 1])
                Imagen.itemset((y, x, 2), Valores[i + 2])
                i = i + 3
```

In this part if there is an error or it doesn’t detect the image in the screen will appear “Cannot capture” or “Not enough data for the image”. If the data was captured correctly the command “else” of the “if” will accommodate the binary values in an array for the format of the image. Starting from 0 giving values for “y” that is “Altura” (height) and x that is “Ancho” (width) this will accommodate de values as previously said.

Algorithm 5. Code to initialize the artificial vision and create the color masks, it is necessary to apply this to each color.

```
#_,imagen = camara.read()
        hsv = cv2.cvtColor(Imagen, cv2.COLOR_BGR2HSV)
        mascara_morado = cv2.inRange(hsv, morado_bajos, mora-
do_altos)
        kernel = np.ones((5,5), np.uint8)
        mascara_morado = cv2.morphologyEx(mascara_morado,
cv2.MORPH_CLOSE, kernel)
        mascara_morado = cv2.morphologyEx(mascara_morado,
cv2.MORPH_OPEN, kernel)
        Morado = cv2.countNonZero(mascara_morado)
```

In this part we use the images and apply the range of colors the lower and higher spectrum defining what is considered to be the color, in a light tone or a dark one. This part will give multiple “mascaras” this are the ones that separate each color and detect it. For this is needed the variable previously used to establish the colors ranges in the spectrum we establish the morphology with this along with the kernel that was defined for the image. The last

Algorithm 6. This code shows the color masks in real time.

```
         cv2.imshow('camara',Imagen)
        cv2.imshow('mascara', mascara)
        cv2.imshow('morado', mascara_morado)
        cv2.imshow('naranja', mascara_naranja)
        cv2.imshow('roja', mascara_roja)
        cv2.imshow('azul', mascara_azul)
```

This code shows 6 images, as sown in figure 6, one for each color used, showing all things that the camera of the NAO is available to detect.

Algorithm 7. In this part of the code the values of every mask are compared.

```
if Morado > Rojo & Morado > Naranja & Morado > Azul:
            print 'Morado'
            tts.say("El objeto es morado")
        elif Morado < Rojo & Rojo > Naranja & Rojo > Azul:
            print 'Rojo'
            tts.say("El objeto es rojo")
        elif Naranja > Rojo & Morado < Naranja & Naranja > Azul:
            print 'naranja'
            tts.say("El objeto es naranja")
        elif Azul > Rojo & Azul > Naranja & Morado < Azul:
            print 'Azul'
            tts.say("El objeto es azul")
```

Finally, the command if is used again in this case to make the NAO said the color detected. In order to achieve this, it’s needed to make the program recognize the highest value of the spectrum for the selected color. In the first line we can see how the color “Morado” (purple) needs to be higher than the others color like “rojo” (red), “azul” (blue) and “naranja” (orange) in order to make NAO able to differentiate from the other colors and said the name of it.

This kind of therapy differs from traditional therapies because of the interaction between the humanoid and the children. As the kids with DS can lose the attention really quick, at the moment that the child is watching something different (NAO), also very attractive because of the color LED's and the movements NAO can hold their attention a little bit longer so the kid can analyze, understand and learn what NAO is saying, increasing the ability to relate the names with the colors not only those of the figures, but the colors that the child sees around him.

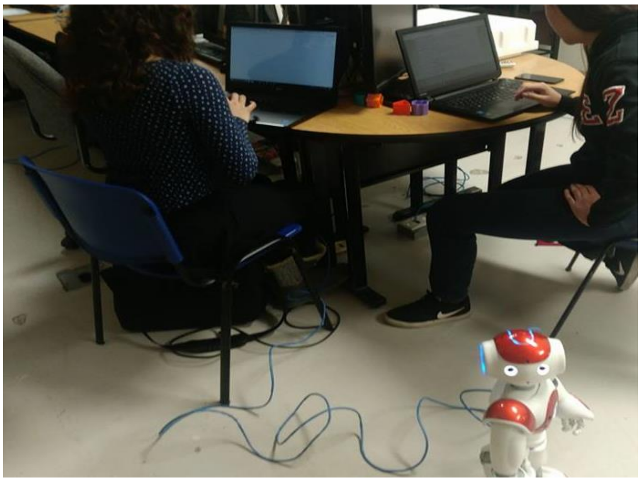


Fig. 8. In this picture it can be observed that NAO, without the necessity of being running a program at the moment turn to see the person that was behind him.

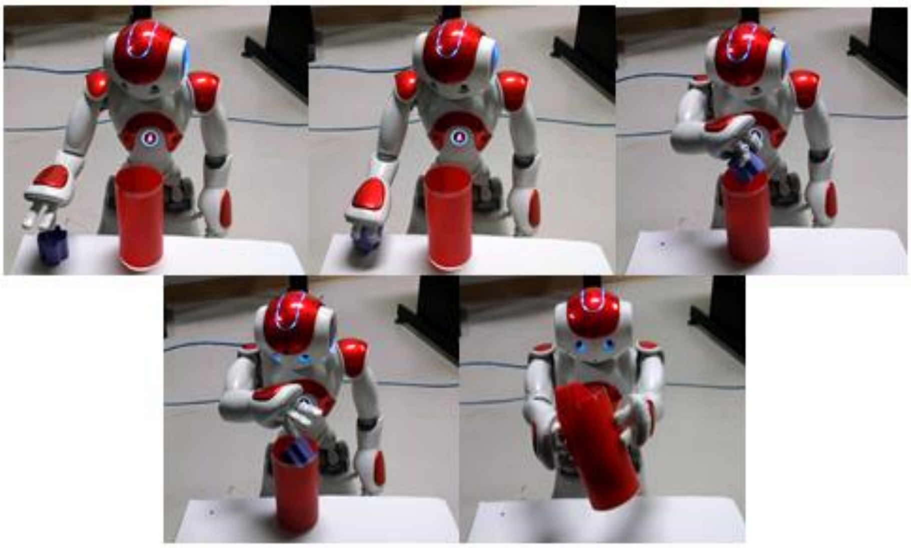


Fig. 9. Main Moments of Figures Movement: 1) Locate hand over the Figure, 2) Pick the Figure, 3) Locate the Figure and Hand on the Bottle, 4) Drop off the figure inside the bottle, 5) Lift the Bottle**.**

Also tried to give NAO autonomy so he can play and interact with the children without the need of a specific program to achieve each action or activity and it does not look so automatized to facilitate the interaction, as is shown in figure 8.

Referring to the movement of figures, the goal is achieved whether they are located on the mark or near, at the same the bottle where figures are placed. This action is done in five main phases: locate hand over the figure, pick the figure, locate the figure and hand on the bottle, drop off the figure inside the bottle and lift the bottle. It is necessary to locate pieces one by one on the pointer place. Once the humanoid realizes the last movement, it synthesizes the phase: “We need to save this”. The moments are shown in figure 9.

4 Conclusion

To conclude, due by the lack of skilled people to educate children with Down syndrome, the ninety-seven percent of those children do not receive the special education they require, resulting very difficult to include them in the society. For that reason, the use of a Nao humanoid would be a great support for the education of kids with Down syndrome. Since the operation of efficient tools as tactile and precision sensors, cameras, microphone and voice synthesizer takes advantage because they attract the whole children attention. That means, an interaction between the humanoid and the user is achieved. Kids, who own Down syndrome, have short and long term deficiencies. So, activities have to be repetitive in order to accomplish the child remembers them. Meanwhile, Nao operates three learning ways: kinesthetic, visual and auditory. This research lists the first steps to support kids with Down syndrome to develop certain abilities. First, the knowledge of the four colors, which were mentioned before, and locating each piece for the robot makes its work. Also, it hopes that motor coordination would be an essential contribution for the child does the same movements from the robot. When being the child who places figures inside the container. The goal is the kid keeps this action to be applied on others objects, taking in account this will be done after any activity where several objects were operated. That is expected on a future, the humanoid programming enables an entire intersection within the infant with Down syndrome

5 Future Research

Due to the lack of personnel specialized in the education of children with Down syndrome, 97% of down children are left without education and, therefore, the inclusion of them is difficult. Therefore, the use of the humanoid Nao would be very useful as support in the education of children with SD taking advantage of their tactile and pressure sensors, cameras, microphone and speech synthesizer can capture the attention of the child's complete child using each of the components of the robot to achieve a humanoid-user interaction. Children with DS have deficiencies in their short and long term memory, so the activities must be repetitive to get the child to retain it and with Nao the three forms of learning are used: kinesthetic, visual, and auditory, as is proposed in figure 9.

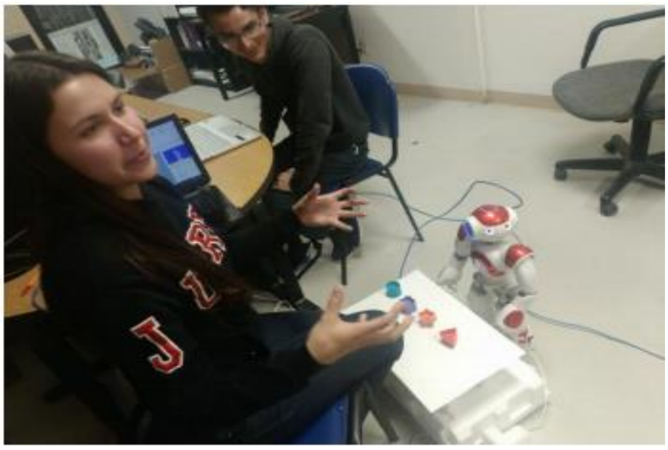


Fig. 10. Organization of objects by color to support to Down Syndrome Children.

In this section, the first steps are taken to help the child develop certain skills. First of all, learning the four colors that were used in this section, followed by the collaboration, placing each piece for the humanoid to do his work. One contribution plus motor coordination, it is expected that the child imitates the movements of the robot and is he, the one who places the pieces inside the boat, looking for this action to be recorded and can apply it in other objects, understanding that it is an action which must be carried out after any activity in which various objects were used. It is expected that in the future the programming of the humanoid will allow a total interaction with the Down child developing a complete therapy.

References

1. Perfil psicológico síndrome de Down, http://asnimo.com/wp-content/uploads/2016/02/PERFIL-PSICOLOGICO-DEL-SD.pdf.
2. Bruni, M., Bethesda, M.D.: Fine Motor skills for Children with Down Syndrome: A Guide for Parents and professionals. Woodbine House (2006)
3. 2.Corretger, J., Serés, A, Casaldaliga, J, Trias, K.: Síndrome de Down. Masson, Barcelona, Spain (2005)
4. Solo el 3% de personas con Síndrome de Down estudian. Available: www.adn40.mx/noticia/mexico/nota/2017-07-07-08-30/solo-el-3--de-personas-con-sindrome-de-down-estudian/ (2017)
5. García, M., Bello, R., Martín, G.: Habilidades cognitivas, conducta y potencial de aprendizaje en preescolares con síndrome de Down. Electrical Journal of Research in Educational Psychology 8(1), 87–110 (2010)
6. Gavin, M.: El síndrome de Down, https://kidshealth.org/es/parents/down-syndrome-esp.html (2012)
7. González, A., Martínez, F., Pernía, A., Elías, F., Castejón, M., Ordieres, J., Vergara, E.: Técnicas y algoritmos básicos de visión artificial. Universidad de la Rioja, España (2006)
8. Heinrich, S., Folleher, P., Springst¨ube, P., Strahl, E., Twiefel, J., Weber, C., Wemter, S.: Object Learning with Natural Language in a Distributed Intelligent System: A Case Study of Human-Robot Interaction. DOI https://doi.org/10.1007/978-3-642-37835-5_70 (2012)

9. Martinez, A.: Síndrome de Down Necesidades educativas y desarrollo del lenguaje. Vitoria-Gasteiz, España (1997)
10. Mohamed, H.: El habla y el lenguaje en niños con síndrome de Down. Propuesta e intervención. Valladolid, España. Available: https://uvadoc.uva.es/bitstream/10324/7755/1/TFG-G%20866.pdf
11. Monjas, M.: Programa de Enseñanza de Habilidades e Interacciones Sociales. Ciudad de México, México: CEPE (2018)
12. Robotrónica. NAO: Los robots del futuro son ya una realidad. AliveRobots.Web. aliverobots.com/nao/
13. Sierra, M., Navarrete, E., Canún, S., Reyes, A., Valdés, J.: Prevalencia del síndrome de Down en México utilizando los certificados de nacimiento vivo y de muerte fetal durante el periodo 2008-2011. Bol Med Hosp Infant Mex. 71(5), 292–297, Elsevier (2014)
14. Valero, A. Principios de color y holopintura. Alicante, España: Editorial Club Universitario (2013)
15. Angulo, J., Serra, J.: Segmentación de Imágenes en Color utilizando Histogramas Bi-Variables en Espacios Color Polares Luminancia/Saturación/Matiz. Computación y Sistemas, 8(4), 303–316 (2005)
16. Mendiola-Santibañez, J., Arias-Estrada, M., Santillán-Méndez, I., Rodríguez-Reséndiz, J., Gallegos-Duarte, M., Gómez-Meléndez, D., Terol-Villalobos, I.: Morphological Filtering Algorithm for Restoring Images Contaminated by Impulse Noise. Computación y Sistemas 19, 243–254 (2015)
17. Chartomatsidis, M., Androulakis, E., Kavallieratou, E.: Training NAO using kinect. Research in computing science. 123, 27–37 (2016)

ISSN 1870-4069

Sign Language Recognition Based on EMG Signals through a Hibrid Intelligent System

Bernabé Rodríguez-Tapia[1,2], Alberto Ochoa-Zezzatti[2], Angel Israel Soto Marrufo[2],
Norma Candolfi Arballo[1], Patricia Avitia Carlos[1]

[1] Universidad Autónoma de Baja California, Escuela de Ciencias de la Ingeniería y Tecnología,
Baja California, Mexico

[2] Universidad Autónoma de Ciudad Juárez,
Departamento de Ingeniería Industrial y Manufactura, Chihuahua, Mexico

rodriguez.bernabe@uabc.edu.mx, alberto.ochoa@uacj.mx,
angel.soto@uacj.mx,ncandolfi@uabc.edu.mx and
patricia_avitia@uabc.edu.mx

Abstract. Non-verbal communication is an important part of everyday interactions and human-computer interaction. Vision techniques and instrumented gloves for sign language recognition are commonly used, but these are often expensive and considered invasive to the user. This research proposes the recognition of words from the American Sign Language (ASL) using the SCEPTRE database acquired by two Myoelectrical bracelets. Computational intelligence techniques were used to optimize the number of attributes using Principal Component Analysis (PCA) and a classifier based on Neural Networks (NN). The results suggest that it is possible to reduce the attributes using PCA without significantly losing the quality in classification. This allows faster processing, a convenient feature for classifiers for real-time SL recognition.

Keywords: sign language recognition, ASL, myoelectric signals, EMG.

1 Introduction

Nonverbal communication is an important part of everyday interactions, Sign Language (SL) is the native language of hearing impaired people and consists of a set of specific gestures [1]. Sign language recognition (SLR) and gesture-based control are two applications for hand gesture technologies[2], in this regard most of the present work on sign language recognition focuses on two methods: the use of a single-lens reflex camera that interprets signs through computer vision and on the other hand the use of glove-based gesture recognition [3].

The vision-based single-lens reflex camera may perform poorly under low-light conditions and captured videos/images may be considered invasive to user's privacy, as well as the detection glove is often expensive [3].

Many studies have suggested the use of surface electromyography (sEMG) signals as a method of interacting with machines. In an EMG-based interaction system, hand

pp. 253–262; rec. 2018-09-15; acc. 2018-10-07

gestures are captured by sEMG sensors measuring the activities of the muscular system [4].

The human-computer communication usually occurs by text or voice; trained interfaces are necessary to recognize gestures on the traditional elements of the user interface.

The aim of the present research is to decode sign language taken by two Myoelectric bracelets using the SCEPTRE database [5, 6] using computational intelligence techniques to optimize the amount of pattern through Principal Component Analysis (PCA) and a classifier based on neural networks (NN).

2 Related Work

According to the sensing technologies used to capture gestures, conventional researches on hand gesture recognition can be categorized into two main groups: data glove-based and computer vision-based techniques [7].

The work in [8] reported a system using two data gloves and three position trackers as input devices and a fuzzy decision tree as a classifier to recognize Chinese Sign Language (CSL) gestures. The average classification rate of 91.6% was achieved over a very impressive 5113-sign vocabulary in CSL. However, glove-based gesture recognition requires the user to wear a cumbersome data glove to capture hand and finger movement

The work [3] developed an impressive real-time system recognizing sentence-level American Sign Language generated by 40 words using HMMs. From a desk-mounted camera, word accuracies achieved 91.9% with a strong grammar and 74.5% without grammar, respectively. The work in [9]employed a spatiotemporal feature extraction

Unlike the approaches mentioned earlier, the accelerometer (ACC) and electromyography (EMG) sensor provide two potential technologies for gesture sensing. Previous studies indicated that the combined sensing approach could improve the performance of hand gesture recognition significantly [10]. [11] have compared the performance of ACC-based and EMG-based techniques in the detection of functional motor activities for rehabilitation and provided evidence that the system based on the combination of EMG and ACC signals can be built successfully.

The work [12] has done a study on combining accelerometer and EMG sensors to recognize sub-word level gestures for Chinese Sign Language and [4] show that combination of multiple sensors helps to increase the recognition accuracy.

Following this path, we add the EMG measurements from eight built-in pods in the Myo device to get information that can be leveraged to detect subtle finger movements and configurations, which are essential for detecting certain signs and distinguishing them from others.

3 Methodology

The global gesture recognition methodology consists of the following steps as shown in Fig. 1.

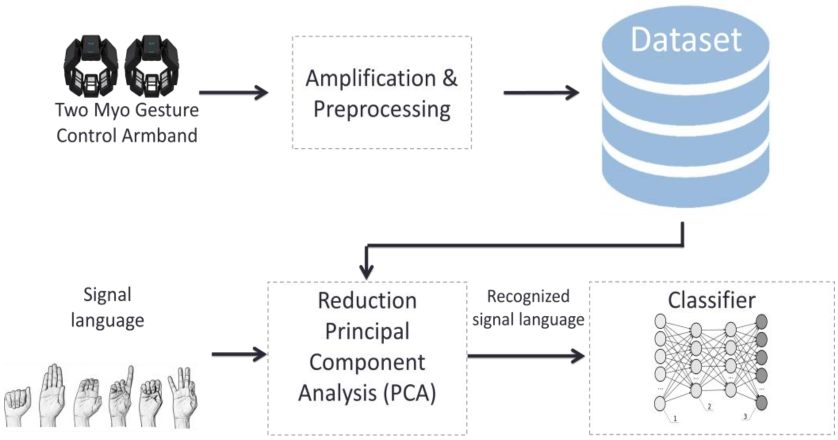

Two Myo Gesture Control Armband
Amplification & Preprocessing
Dataset
Signal language
Reduction Principal Component Analysis (PCA)
Recognized signal language
Classifier
1
2
3

Fig. 1. Classification process of Signal Language recognition with Myo Gesture Control Armband. Adapted from Raez et al /Biological Procedures (2006) 11-35 [13].

Amplification and Preprocessing

The data set represented in Fig. 1, were taken from the database developed by [5], these data were obtained by using two MYO bracelets for some signs of American Sign Language (ASL). The data set represents a subsampling data from a Myo placed on the left hand and the other one on the right hand of each user. The data reading is made by 8 EMG sensors per bracelet at a frequency of 200Hz; it also has a nine axis Inertial Measurement Unit (IMU) including three axis gyroscop (gyp), three axis accelerometer (accl) and three axis magnetometer (o). The Myo bracelet can extract IMU data at a sampling frequency of 50Hz. The range of potentials provided by the Myo bracelet is amplified between -128 and 128 trigger units.

Fig. 2 represents the implementation of a user with wrist devices in each hand making a gesture.

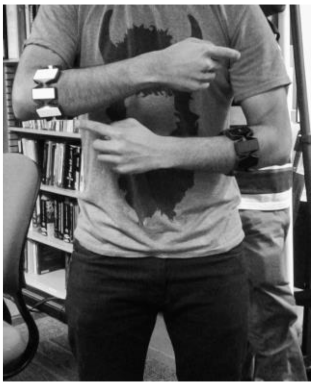
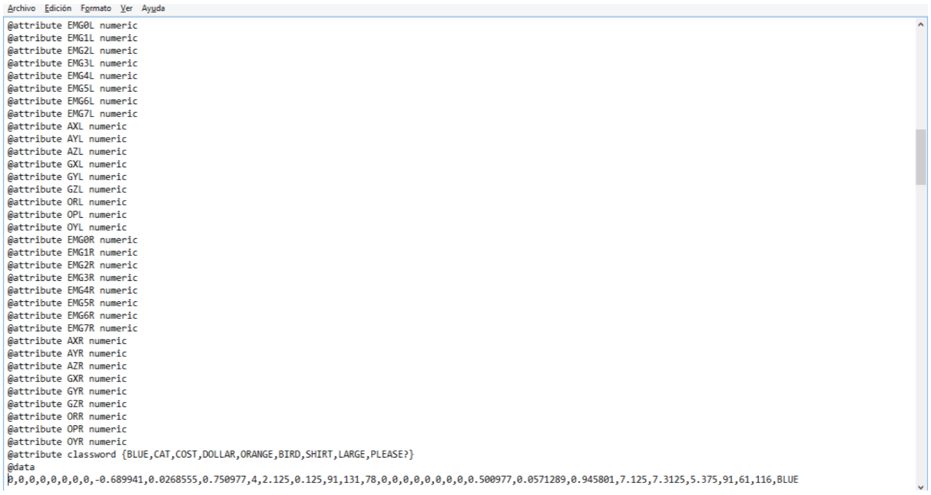


Fig. 2. Deployment: A user with a wrist-band devices on each hand performing a gesture[5].

Data Collection

The database consists of gestures taken from 10 healthy adults between the ages of 22 and 35, each of whom performs a total of 20 ASL gestures. For this research the information of 9 words made by 3 different users was selected, each user repeated the word 5 times. The data set consists of 34 attributes and 1 class. The attributes in this data set are: 16 EMG vectors, 6 accelerometer vectors, 6 gyroscope vectors and 6 position vectors. The number of classes is equivalent to the recognition of the 9 ASL words: (1) BLUE, (2) CAT, (3) COST, (4) DOLLAR, (5) ORANGE, (6) BIRD, (7) SHIRT, (8) LARGE and (9) PLEASE. Fig. 3 presents the description of the attributes and an example of the analyzed data vector.



```
@attribute EMG0L numeric
@attribute EMG1L numeric
@attribute EMG2L numeric
@attribute EMG3L numeric
@attribute EMG4L numeric
@attribute EMG5L numeric
@attribute EMG6L numeric
@attribute EMG7L numeric
@attribute AXL numeric
@attribute AYL numeric
@attribute AZL numeric
@attribute GXL numeric
@attribute GYL numeric
@attribute GZL numeric
@attribute ORL numeric
@attribute OPL numeric
@attribute OYL numeric
@attribute EMG0R numeric
@attribute EMG1R numeric
@attribute EMG2R numeric
@attribute EMG3R numeric
@attribute EMG4R numeric
@attribute EMG5R numeric
@attribute EMG6R numeric
@attribute EMG7R numeric
@attribute AXR numeric
@attribute AYR numeric
@attribute AZR numeric
@attribute GXR numeric
@attribute GYR numeric
@attribute GZR numeric
@attribute ORR numeric
@attribute OPR numeric
@attribute OYR numeric
@attribute classword {BLUE,CAT,COST,DOLLAR,ORANGE,BIRD,SHIRT,LARGE,PLEASE?}
@data
0,0,0,0,0,0,0,0,-0.689941,0.0268555,0.750977,4,2.125,0.125,91,131,78,0,0,0,0,0,0,0,0,0.500977,0.0571289,0.945801,7.125,7.3125,5.375,91,61,116,BLUE
```

Fig. 3. Description of attributes and data vector.

Principal Component Analysis

The PCA [14] method, a common dimensionality reduction technique, aims to find a set of orthonormal vectors in the amount of data, which can maximize the variance of the data and map the data into a lower sub-dimensional space encompassed by those vectors [15]. It is widely used to retrieve important information from a multivariable table data and to retrieve that information as a set of few new variables called main components. These new variables correspond to a linear combination of the original ones. The number of main components is less than or equal to the number of original variables.

Weka@ allows to apply a wide range of filters on the data, allowing to make transformations on them of all kinds. It carries out an analysis of the main components and transformation of the data. Dimensionality reduction is achieved by choosing enough vectors of its own to reflect some percentage of the variance in the original data, by default 0.95 (95%). Based on the code from the attribute selection scheme "Principal Components" by Mark Hall and Gabi Schmidberger [16].

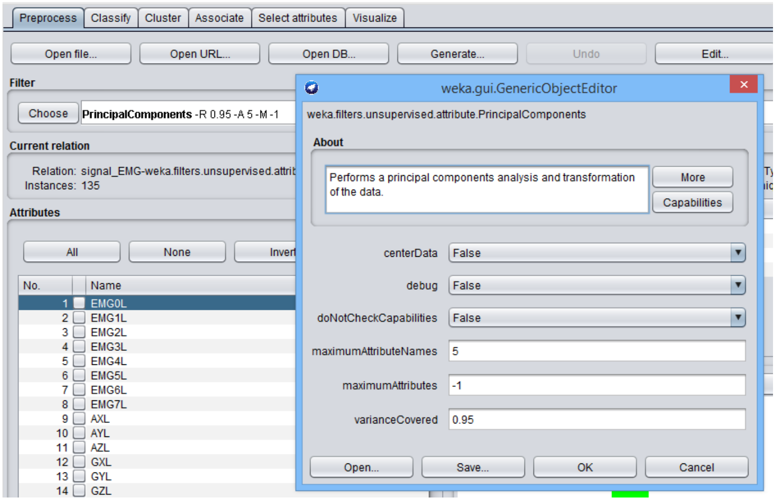
Classification

A multilayer perceptron (MLP) is often used as a classifier for it is the most popular type of neural network architecture. A multi-layer perceptron has any number of inputs, has one or more hidden layers with any number of units, uses generally sigmoid activation functions in the hidden layers, has connections between the input layer and the hidden layer, between the hidden layers and between the last hidden layer and the output layer[17].

The advantage of a neural network is its ability to represent both linear and non-linear relationships, and learn these relationships directly from data being modeled. It also meets real time constraints, which are an important feature in control systems [18]

The multilayer perceptron is a classifier that uses backpropagation to learn a multi-layer perceptron to classify instances.

The network can be built by hand in WEKA@ or be configured using simple heuristics. The network parameters can be monitored and modified during the training period. The nodes in this network are all sigmoid (except when the numerical class, in which case the output nodes become linear units with no threshold) [16].



PrincipalComponents -R 0.95 -A 5 -M -1

Relation: signal_EMG-weka.filters.unsupervised.attrib
Instances: 135

No.		Name
1	□	EMG0L
2	□	EMG1L
3	□	EMG2L
4	□	EMG3L
5	□	EMG4L
6	□	EMG5L
7	□	EMG6L
8	□	EMG7L
9	□	AXL
10	□	AYL
11	□	AZL
12	□	GXL
13	□	GYL
14	□	GZL

weka.filters.unsupervised.attribute.PrincipalComponents

Performs a principal components analysis and transformation of the data.

centerData: False
debug: False
doNotCheckCapabilities: False
maximumAttributeNames: 5
maximumAttributes: -1
varianceCovered: 0.95

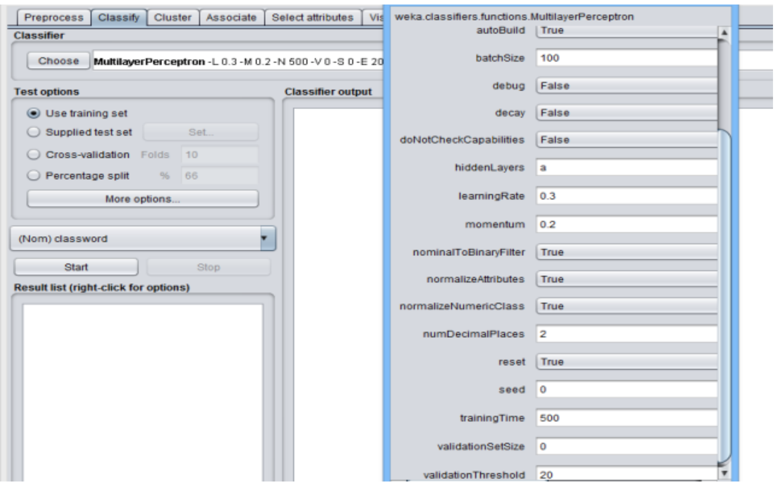
Fig. 4. Configuración WEKA$^{@}$. Multilayer perceptron.

Ranking and Optimization

The classification and optimization were carried out in the WEKA@ software. A data-based of 3 users was used with 5 repetitions each for the 9 selected words.

Neural network training without optimization was performed with the 34 selected sensor attributes and using the multilayer perceptron. The training and testing was performed with the fusion "Use training set"; with this option WEKA$^{@}$ will train with all available data and then will apply it again on them.

For the optimization, we used the filter "principal components" within the "unsupervised" attributes of WEKA @. The configuration of the PCA filter and the MLP neural network are described in Fig. 4 and Fig. 5.



MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) classword

weka.classifiers.functions.MultilayerPerceptron

autoBuild True
batchSize 100
debug False
decay False
doNotCheckCapabilities False
hiddenLayers a
learningRate 0.3
momentum 0.2
nominalToBinaryFilter True
normalizeAttributes True
normalizeNumericClass True
numDecimalPlaces 2
reset True
seed 0
trainingTime 500
validationSetSize 0
validationThreshold 20

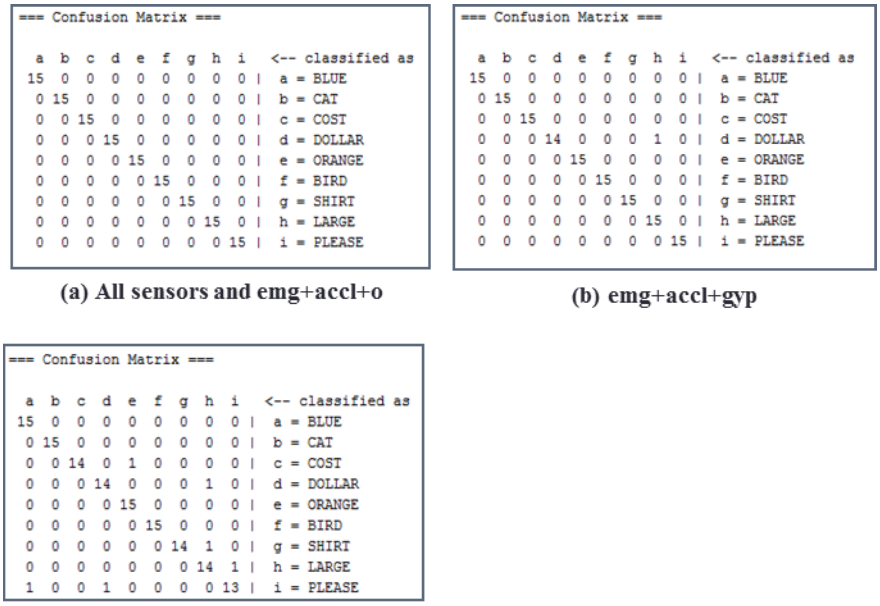
Fig. 5. Configuración WEKA$^{@}$. Principal Components.

4 Evaluation and Results

The experiments were carried out with a 9-word database with 7 sensor sets (attributes): (1) emg+accl+gyp+o (34 attributes) , (2) emg+accl+gyp (29 attributes), (3) emg+accl+o (29 attributes), (4) emg+accl (23 attributes), (5) emg+gyp (23 attributes), (6) emg+o (23 attributes); (7) emg (16 attributes). The results statistics for the 3 subjects are shown in Table 1 and the confusion matrix for the best 4 classification results using only the MLP neural network is shown in Fig. 6.

Table 1. Classification and optimization results

Atributes	*Multilayer Perceptron*	*Multilayer Perceptron with PCA*
(1) All (emg, accl, gyp, o)	100%	97.77%
(2) emg+accl+gyp	99.25%	94.81%
(3) emg+accl+o	100%	99.25%
(4) emg+accl	95.55%	90.62%
(5) emg+gyp	77%	74.07%
(6) emg+o	93.44%	93.33%
(7) emg	47.4%	47.4%



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 15  0  0  0  0  0  0  0  0 |  a = BLUE
  0 15  0  0  0  0  0  0  0 |  b = CAT
  0  0 15  0  0  0  0  0  0 |  c = COST
  0  0  0 15  0  0  0  0  0 |  d = DOLLAR
  0  0  0  0 15  0  0  0  0 |  e = ORANGE
  0  0  0  0  0 15  0  0  0 |  f = BIRD
  0  0  0  0  0  0 15  0  0 |  g = SHIRT
  0  0  0  0  0  0  0 15  0 |  h = LARGE
  0  0  0  0  0  0  0  0 15 |  i = PLEASE
```

(a) All sensors and emg+accl+o

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 15  0  0  0  0  0  0  0  0 |  a = BLUE
  0 15  0  0  0  0  0  0  0 |  b = CAT
  0  0 15  0  0  0  0  0  0 |  c = COST
  0  0  0 14  0  0  0  1  0 |  d = DOLLAR
  0  0  0  0 15  0  0  0  0 |  e = ORANGE
  0  0  0  0  0 15  0  0  0 |  f = BIRD
  0  0  0  0  0  0 15  0  0 |  g = SHIRT
  0  0  0  0  0  0  0 15  0 |  h = LARGE
  0  0  0  0  0  0  0  0 15 |  i = PLEASE
```

(b) emg+accl+gyp

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i   <-- classified as
 15  0  0  0  0  0  0  0  0 |  a = BLUE
  0 15  0  0  0  0  0  0  0 |  b = CAT
  0  0 14  0  1  0  0  0  0 |  c = COST
  0  0  0 14  0  0  0  1  0 |  d = DOLLAR
  0  0  0  0 15  0  0  0  0 |  e = ORANGE
  0  0  0  0  0 15  0  0  0 |  f = BIRD
  0  0  0  0  0  0 14  1  0 |  g = SHIRT
  0  0  0  0  0  0  0 14  1 |  h = LARGE
  1  0  0  1  0  0  0  0 13 |  i = PLEASE
```

(c) emg+accl

Fig. 6. Confusion matrix for better results.

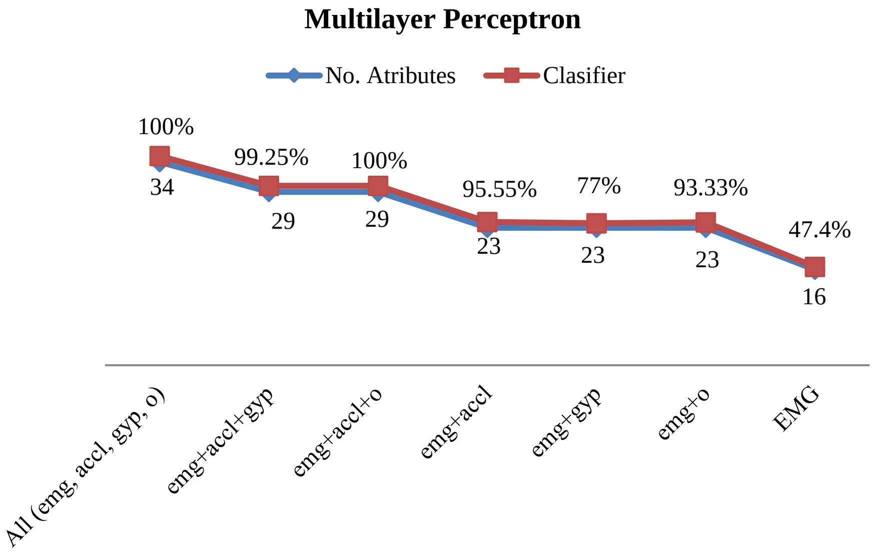


Fig. 7. % classification by no. of attributes.

In addition to using the 34 sensors of both myoelectric bracelets for MLP training and testing, the results indicate that the 2, 3 and 4 sensor sets retain sufficient characteristics to obtain a good rating. In particular, set 2 (emg+accl+o) achieves the same classification as when using all sensors.

The emg+accl data set reflects the best data set, as it decreased from 34 attributes to only 23 without the need for an optimization stage.

The Fig. 7 shows the quality of classification with respect to the number of attributes.

Applying the PCA filter to the total number of attributes showed a significant decrease of these attributes, the results for all data sets are shown in Fig. 8. The most effective option can be identified when working with the emg+accl+o sensors, since when PCA was applied it was possible to reduce from 34 to 20 attributes which achieved a classification of 99.25%, followed by this set, with only the emg+o sensors there is a total of 18 characteristics, almost half of the initials and there is a classification of up to 93.33%.

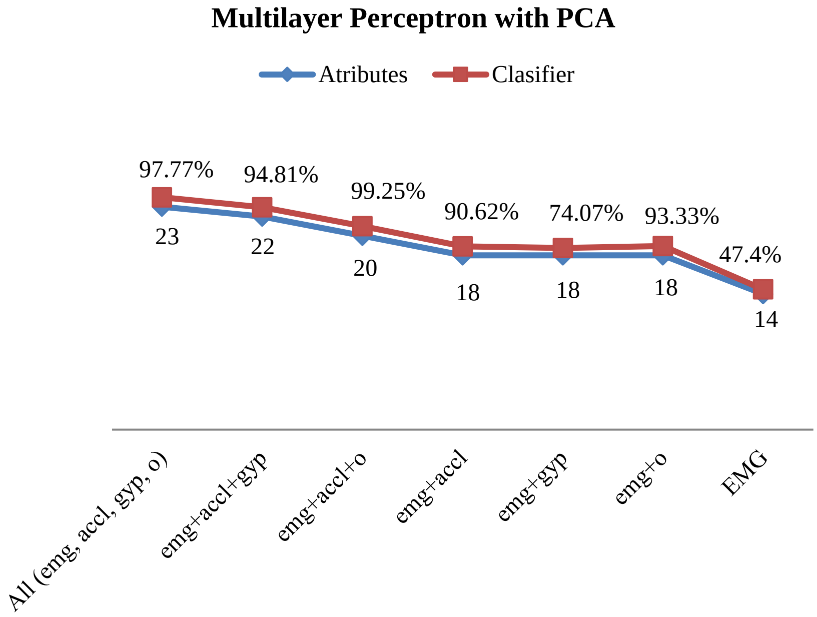


Fig. 8. % classification by no. of attributes with PCA.

5 Conclusion and Future Work

The aim of the present research was to decode the sign language taken by two Myoelectric bracelets using the SCEPTRE database. Using computational intelligence techniques, 34 to 20 attributes were optimized through the Principal Component Analysis (PCA) filter, having a 99.25% classification with a "multilayer perceptron" neural network.

For the further investigation, it is expected to use the PCA optimization method and the MLP classifier in an HMI system with the characteristics of real time operation by means of the bluetooth protocol, for the recognition of SL gestures on the traditional elements of the user interface

References

1. Kosmidou, V.E., Hadjileontiadis, L.J., Panas, S.M.: Evaluation of surface EMG features for the recognition of American Sign Language gestures. En: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 6197-6200. IEEE, New York, NY (2006)
2. Xu Zhang, Xiang Chen, Yun Li, Lantz, V., Kongqiao Wang, Jihai Yang: A Framework for Hand Gesture Recognition Based on Accelerometer and EMG Sensors. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 41, 1064-1076 (2011). doi:10.1109/TSMCA.2011.2116004
3. Wu, J., Tian, Z., Sun, L., Estevez, L., Jafari, R.: Real-time American Sign Language Recognition using wrist-worn motion and surface EMG sensors. En: 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN). pp. 1-6. IEEE, Cambridge, MA, USA (2015)
4. Chen, X., Zhang, X., Zhao, Z.-Y., Yang, J.-H., Lantz, V., Wang, K.-Q.: Hand Gesture Recognition Research Based on Surface EMG Sensors and 2D-accelerometers. En: 2007 11th IEEE International Symposium on Wearable Computers. pp. 1-4. IEEE, Boston, MA, USA (2007)
5. Paudyal, P., Banerjee, A., Gupta, S.K.S.: SCEPTRE: A Pervasive, Non-Invasive, and Programmable Gesture Recognition Technology. En: Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI ’16. pp. 282-293. ACM Press, Sonoma, California, USA (2016)
6. Paudyal, P., Lee, J., Banerjee, A., Gupta, S.K.S.: DyFAV: Dynamic Feature Selection and Voting for Real-time Recognition of Fingerspelled Alphabet using Wearables. En: Proceedings of the 22nd International Conference on Intelligent User Interfaces - IUI ’17. pp. 457-467. ACM Press, Limassol, Cyprus (2017)
7. Mitra, S., Acharya, T.: Gesture Recognition: A Survey. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews). 37, 311-324 (2007). doi:10.1109/TSMCC.2007.893280
8. Fang, G., Gao, W., Zhao, D.: Large Vocabulary Sign Language Recognition Based on Fuzzy Decision Trees. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 34, 305-314 (2004). doi:10.1109/TSMCA.2004.824852
9. Shanableh, T., Assaleh, K., Al-Rousan, M.: Spatio-Temporal Feature-Extraction Techniques for Isolated Gesture Recognition in Arabic Sign Language. IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics). 37, 641-650 (2007). doi:10.1109/TSMCB.2006.889630
10. Brashear, H., Starner, T., Lukowicz, P., Junker, H.: Using multiple sensors for mobile sign language recognition. En: Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings. pp. 45-52. IEEE, White Plains, NY, USA (2003)

11. Sherrill, D.M., Bonato, P., De Luca, C.J.: A neural network approach to monitor motor activities. En: Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology. pp. 52-53. IEEE, Houston, TX, USA (2002)
12. Li, Y., Chen, X., Tian, J., Zhang, X., Wang, K., Yang, J.: Automatic recognition of sign language subwords based on portable accelerometer and EMG sensors. En: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction on - ICMI-MLMI '10. p. 1. ACM Press, Beijing, China (2010)
13. Raez, M.B.I., Hussain, M.S., Mohd-Yasin, F., Reaz, M., Hussain, M.S., Mohd-Yasin, F.: Techniques of EMG signal analysis: detection, processing, classification and applications. Biological procedures online. 8, 11-35 (2006). doi:10.1251/bpo115
14. M. Turk, A. Pentland: Eigenfaces for Recognition. Journal of Cognitive Neuroscience. 3, 71-86 (1991). doi:10.1162/jocn.1991.3.1.71
15. D. Huang, W. Hu, S. Chang: Vision-Based Hand Gesture Recognition Using PCA+Gabor Filters and SVM. En: 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. pp. 1-4 (2009)
16. Witten, I.H., Frank, E., Trigg, L.E., Hall, M.A., Holmes, G., Cunningham, S.J.: Weka: Practical machine learning tools and techniques with Java implementations. (1999)
17. Bu, N.B.N., Okamoto, M., Tsuji, T.: A Hybrid Motion Classification Approach for EMG-Based Human-Robot Interfaces Using Bayesian and Neural Networks. IEEE Transactions on Robotics. 25, 502-511 (2009). doi:10.1109/TRO.2009.2019782
18. Rechy-Ramirez, E.J., Hu, H.: Bio-signal based control in assistive robots: a survey. (2016). doi:10.1016/j.dcan.2015.02.004

ISSN 1870-4069

Visual Association Rules on the Psychological Connection of University Students with their Studies

Erika Yunuen Morales Mateos[1], María Arely López Garrido[1],
José Alberto Hernández Aguilar[2], Carlos Alberto Ochoa Ortiz[3],
Oscar Alberto González González[1], Arturo Corona Ferreira[1]

[1] Universidad Juárez Autónoma de Tabasco Cunduacán, Tabasco, Mexico
[2] Universidad Autónoma del Estado de Morelos, Cuernavaca, Morelos, Mexico
[3] Universidad Autónoma de Ciudad Juárez, Ciudad Juárez, Chihuahua, Mexico
erika.morales@ujat.mx, arely.lopez@ujat.mx,
jose_hernandez@uaem.mx, alberto.ochoa@uacj.mx,
oscar.gonzalez@ujat.mx, arturo.corona@ujat.mx

Abstract. The objective of this work is to express patterns of data behavior related to the psychological connection of the university students with their studies, this through the association rules, likewise these rules are presented in an interactive visual way, which facilitates the review of these to select those that are considered of interest. For this study, students of four careers in technologies from a university in southern Mexico were considered. To measure the psychological engagement with the studies, the Utretch Work Engagement Scale for Students instrument (UWES-S) was used, other elements were included to know opinions related to careers. The learning algorithm of association rules A priori was used. For the generation and interactive visualization of the association rules, the Apriori algorithm was used in the R language. It is concluded that the interactive visual display is an important support for the selection of the rules that are considered relevant.

Keywords: rules of association, visualization, apriori, psychological connection, student engagement.

1 Introduction

Research is making an important advance given that nowadays it is possible to rely on techniques that offer different forms of data analysis. The data storage to subsequently apply different forms of knowledge extraction has presented a great advantage, which can be applied to the development of psychological, educational, and other studies. Many of the psychological studies usually involve applying questionnaires, which can be automated to generate data sets that include these items, for further analysis. Data mining techniques offer several options for discovering knowledge, and it is possible to make use of them in psychology.

One area of psychology where much research is carried out is that of occupational health, which studies syndromes such as burnout that determines to what extent people feel physically and mentally exhausted, or the engagement to determine human potential and psychological capacities in his activities such as work or professional studies. The engagement in the activities is defined as "engagement is a positive psychological state characterized by high levels of energy and vigor, dedication and enthusiasm for work, as well as total absorption and concentration in the work activity" [1]. This concept of engagement has also been taken to the universities, defining it as the psychological engagement with the studies. These data allow knowing new factors that impact on the training of university students, their dedication, absorption and vigor, in relation to the studies. The Utrecht Work Engagement Scale Students (UWES-S) [2] is used to know the levels of the dimensions that involve the psychological connection with the studies.

As mentioned from the data generated from psychological instruments, the application of data mining is possible. Data mining aims to analyze the data to extract knowledge. The knowledge obtained can be in the form of relationships, patterns or rules inferred from the data, or in a more concise description. These relationships or summaries integrate the model of the data analyzed. The models can be represented in different ways and each of them determines the type of technique that can be used to infer them. Descriptive models identify patterns that explain or summarize the data, work to explore the properties of the data examined, but do not predict [3]. Association rules are a descriptive task, similar to correlations that identify non-explicit relationships between categorical attributes. There can be many formulations, the most common is the one of the style: if the tribute X has the value to then the attribute Y has the value b. These rules do not involve a cause-effect relationship, so there may be no cause for the data to be related. The rules are evaluated using the precision and support parameters [3].

The presentation of data by visual techniques such as the rules of visual association is that they help to understand the behavior of data with a glance, in this way data analysts can save time when making judgments about patterns, trends, variability, among others, of the data. One of the advantages of using visual techniques to represent data is that they can show characteristics that would otherwise have been complex or impossible [4].

For the development of type of studies that involve generating graphics such as visual association rules, the R language is suggested, which is an integrated collection of software services for data management, calculation and visualizations. R is free under the terms of the GNU General Public License of the Free Software Foundation [5].

In this work, we propose to know the psychological engagement with the university studies, as well as other variables related to their career, which imply in the decision of stay and continuity of the studies, for this purpose it is proposed to use a data mining task: The Association Rules. These rules let to express patterns of a data set, these patterns allow to know the general behavior of the data and in this way obtain descriptive information that can assist in decision making. Likewise, it is intended to go a step further, presenting the resulting rules in a visual graphic format, which facilitates the analysis and selection of the rules of interest.

There are works in which educational and psychological studies have been carried out, where data mining techniques are used, this is the case of the data mining work applied to the identification of risk factors in students in the state of Mexico , among the techniques used is that of association rules [6], a model was proposed for school desertion in Universities of Mexico [7], on the same subject in terms of school failure prediction [8], another work was developed to evaluate parameters of the CENEVAL entrance survey for students who are candidates to enter the upper level, an ITP study case, for which a classification model was generated [9]. There are also works where visual association rules have been generated for different administrative problems, natural disasters, among others [10,11]. Additional work on this research was developed using data analysis techniques based on graphic representation [12,13].

2 Materials and Proposed Methods

2.1 Description of the Data

The objective of this research is to present the results obtained from the application of the rules of association under the Apriori algorithm, including its interactive visualization, to a set of data containing information about psychological engagement, obtained by applying the UWES-S instrument, and other variables related to his career, which imply in the decision of stay and continuity of the studies. The students belong to a sample population of four computer science degrees in a computer and systems faculty, in a university in southern Mexico. The bachelor's degrees considered for this study were the Computer Systems (LSC), Information Technology (LIA Information Technology (LTI) and Telematics (LT). The sample was non-probabilistic, directed and by convenience [5], a survey was applied to a total of 141 students, who accepted to answer the questionnaire voluntarily, in the period February-August 2015 carried out the data collection.

2.2 Wellness Scale in the Academic Context (Utrecht Work Engagement Scale for Students, UWES-S)

The psychological connection with the studies was obtained with the application of the Utretch Work Engagement Scale for Students instrument (UWES-S). This instrument has its origins in the need to know the psychological connection with the activities carried out at work, that is, the engagement to work. Thus, the engagement to work was defined by Salanova and Schaufeli [1] as: "a positive psychological state characterized by high levels of energy and vigor, dedication and enthusiasm for work, as well as total absorption and concentration in the work activity". To measure this concept Schaufeli and Bakker [2] developed the UWES-S, this instrument was subsequently modified, giving rise to a version that would allow knowing the psychological engagement with the studies or also known as engagement to the studies, because the studies are considered an activity in which they have responsibilities and a goal, just like the work. In this version of the instrument for students the same concepts and instruments of engagement were used in the work, there are some differences in the writing of the elements adapting them to the student activity, giving rise to the questionnaire Utretch Work Engagement Scale for Students (UWES-S) [2]. This UWES-S consists of 17 items, divided into three constructs, vigor, dedication and absorption. The vigor consists of six items, which refer to high levels of energy and resilience, willingness to dedicate efforts, not fatigue easily, and persistence in the face of difficulties. The dedication, composed of five items that refer to the meaning or meaning of the work, to feel excited and proud of their work, and feel inspired and challenged by the work. Absorption, which consists of six items that refer to being happily immersed in your work and present difficulty in leaving it, so that time passes quickly, and you forget everything around you "[2].

In the UWES-S the answers are measured according to a Likert scale, where zero means "never" and six "always", in this way the scores go from zero to six for each dimension that makes up the student engagement. The original interna consistency obtained by the authors for the UWES-S version of 17 items in Dutch students was for strength of 0.63, dedication 0.81 and absorption of 0.72, complying with the criterion of superiority to 0.60 for an instrument of recent development. The specifications of the instrument can be found in the manual of Schaufeli and Bakker [2].

2.3 Analysis of Data

The generation of association rules was carried out based on the Apriori algorithm, to find behavior patterns between the data according to the joint appearance of values of two or more attributes, as well as the interactive visual display of these rules to facilitate the observation and selection of the rules. To generate these results, the R language was used, since it is a powerful and versatile software, especially in graphics development, because it has packages that allow complex analysis and design custom graphics, as well as programming at an advanced level [14].

Association Rules

An association rule is a correspondence between itemsets, an itemset is a set of one or more items. If X and Y are two itemsets, the rule that associates X with Y is written: X → Y, where X is the antecedent of the rule and Y the consequent. There are two measures for evaluating the support of an association rule and the confidence of an association rule [15]:

- The support of an association rule is defined as the proportion of transactions that contain the antecedent and the consequent of the rule.

$$\text{supp}\ (X \rightarrow Y) = \text{supp}\ (X \cup Y).$$

- The confidence of an association rule is defined as the proportion of transactions that are met when the rule can be applied.

$$\text{conf}\ (X \rightarrow Y) = \text{supp}\ (X \cup Y)/\text{supp}\ (X).$$

It is important to note that the confidence of the rule X → Y is an estimate of the conditional probability of Y given X obtained from the relative frequencies of occurrence of X and Y. The results of an association analysis in a particular application must be taken with caution. The fact that a rule X → Y has a high confidence does not directly result in X being the cause of Y. The rule shows the simultaneous occurrence of X and Y, but not that X is the cause and Y the consequence.

The application of the association rules is described: Given a universe U, a set of transactions T and two values minsupp ∈ [0,1], minconf ∈ [0,1] find all the association rules X → Y that comply with the following two elements [15]:

- Minimum support supp (X → Y) ≥ minsupp,
- Confidence minima conf (X → Y) ≥ minconf.

An exhaustive search is carried out where all the association rules that can be obtained from universe U must be generated, calculate the support and trust of each rule taking into account the set of transactions T, store the rules that exceed the minimum thresholds minsupp and minconf

If the set has k elements, the number of rules that can be obtained is 2k -2. The search for rules that meet the minimum support and trust requirements can be developed in two stages [15]:

- Generate frequent sets. Find the sets of items whose support is greater than or equal to minsupp.
- Build the association rules. For each of the sets obtained in the previous stage, find the association rules that have a confidence greater than or equal to minconf.

For the efficient generation of frequent sets it is important to consider that if a set of items is not frequent, neither will the sets that contain it. This is how the search tree is pruned and the efficiency of generating frequent sets is increased.

The measure of statistical independence lift is defined as a relationship between the simultaneous occurrence of X and Y, when the sets of items that make up the antecedent and the consequent of the rule are statistically independent. Experts in the area conclude that knowledge rules discovered through support and trust should be filtered using their lift values, since lift values greater than 1 indicate association between items; and values less than 1 indicate their independence between items and should not be considered for decision making [16]:

$$\text{lift}\ (X \rightarrow Y) = \text{conf}\ (X \rightarrow Y)/\text{supp}\ (Y).$$

The previous measures of support, trust and elevation, are to obtain valid association rules, these measures to evaluate the quality of association rules and the elimination of data. These measures are also interesting calls, In general, they are used as input parameters by algorithms of induction of association rules [17].

Apriori Algorithm

The Apriori algorithm is a simple and popular association rules learning algorithm; its operation is based on the search of the sets of items with certain coverage. For this, first, the sets of formed by only one item that exceed the minimum coverage are built. This set of sets is used to construct the set of sets of two items, and so on until a size is reached in which there are no sets of items with the required coverage. The learning of association rules is commonly divided the phase of extraction of sets of items that meet the required coverage from the data, and the generation of the rules from these sets [3].

The Apriori algorithm is described below, which is based on all the observations presented above. The purpose of this algorithm is to efficiently find the frequent sets of items, for which it is proposed to intelligently apply the property of downward closure of the support that indicates that every subset J of a frequent set I is also frequent, likewise, if an itemset does not It is frequent, nor will the itemset that contain it. The specifications are the following: data: T, transactions; entry: minsupp, minimum support threshold; variable: C_k, k-items candidates to be frequent; variable: F_k, k-frequent items; output: $U^{k}_{i = 1} F_i$, frequent sets of items [15].

```
Algorithm 1: Apriori(transactions: T, minimum support: minsupp)
k = 1;
$F_1$ = {sets of frequent 1-itemsets};
repeat
 Generate $C_{k+1}$ joining pairs of elements of $F_k$;
 Prune the itemsets of $C_{k+1}$ that violate the downward closure;
 Determine $F_{k+1}$;
 k = k + 1;
until ($F_k$ = ∅);
return ($U^{k}_{i=1} F_i$);
```

Algorithm 1. Apriori algorithm [15].

3 Results

A categorization was developed (see Table 1) to qualify the level of psychological engagement with the studies, where values were established that allow the evaluation of the observations of a case or group of cases [18]. There are other methods to calculate these levels through the UWES-S, to consult Schaufeli and Bakker [2].

Table 1. Categorization to qualify the UWES-S.

Category	UWES-S	Vigor	Dedication	Absorption
Very low	score < 2.20	score < 2.80	score < 2.50	score < 2.70
Low	2.20 ≤ score < 3.30	2.80 ≤ score < 3.80	2.50 ≤ score < 3.50	2.70 ≤ score < 3.60
Medium	3.30 ≤ score < 4.70	3.80 ≤ score < 5.20	3.50 ≤ score < 4.50	3.60 ≤ score < 4.70
High	4.70 ≤ score ≤ 6.00	5.20 ≤ score ≤ 6.00	4.50 ≤ score ≤ 6.00	4.70 ≤ score < 6.00

Source: Adapted from Schaufeli and Bakker [7]

The data set with which we worked for this study has 141 records of students from four careers in technologies: LSC, LIA, LTI and LT. The attributes that make up the data set are 16 in total, the values that make up this set of data are categorical, considering the nature of the type of study desired. Table 2 shows the correspondence between the values, as well as the scales to which they belong.

Table 2. Correspondence between attributes and values, as well as their measurement scales. Source: Prepared by the researcher (2018).

Attributes	Variable	Scale	Values
Career	Carrera	Nominal	LSC/LIA,/LT/LTI
Semester	Semestre	Nominal	Tenth/Ninth/Eighth/Seventh/Sixth/Fifth/Fourth/ Third/Second/ First. Decimo/Noveno/Octavo/Septimo/Sexto/Quinto/Cuarto/ Tercero/Segundo/Primero.
Gender	Genero	Nominal	Female/Male Mujer/Hombre
Age	Edad	Ordinal	22yGreater/21yLess22/18yLess21 22yMayores/ 21Menor22/ 18yMenor21
UWES-S	UWES-S	Ordinal	VeryLow/Low /Medium/High MuyBajo /Bajo/Medio/Alto
Vigor	Vigor	Ordinal	VeryLow/Low /Medium/High MuyBajo /Bajo/Medio/Alto
Dedication	Dedicación	Ordinal	VeryLow/Low /Medium/High MuyBajo /Bajo/Medio/Alto
Absorption	Absorción	Ordinal	VeryLow/Low /Medium/High MuyBajo /Bajo/Medio/Alto
Average	Promedio	Ordinal	VeryLow/Low /Medium/High MuyBajo /Bajo/Medio/Alto
Selected career	Carrera.seleccionada	Nominal	Yes/No Si/No
Failed class	Reprobado. materia	Nominal	Yes/No Si/No
Change career	Cambio.carrera	Ordinal	Nothing/VeryLittle/Little/Moderately/Much/Enough Nada/ Muypoco/Poco/Medianamente/Mucho/Bastante
Leaving university	Baja. universidad	Ordinal	Nothing/VeryLittle/Little/Moderately/Much/Enough Nada/ Muypoco/Poco/Medianamente/Mucho/Bastante
Finish successful studies	Finalizar.exito	Ordinal	Nothing/VeryLittle/Little/Moderately/Much/Enough Nada/ Muypoco/Poco/Medianamente/Mucho/Bastante
Financing	Financiamiento	Ordinal	Parents/Work/Other Padres/Trabajo/Otros
Social class	Estrato.social	Ordinal	LowLow/LowHigh/MediumLow/MediumHigh/ HighLow/High High BajaBaja/BajaAlta/MediaBaja/MediaAlta/AltaBaja/AltaAlta

Note. The variables and values of the data set are in Spanish, so its translation is shown in this table.

The Table 2 is used for the tests to be approved. The attributes career, semester, gender, identify the groups of students, for the values that take the dimensions that make up the UWES-S, that is, vigor, dedication, absorption, the categorization presented in Table 1 was considered. We considered the variables to complete the study, which imply in the decision of the stay and the continuity of the studies, as the average, others that can be registered as if they had found in the career they selected at the beginning of their studies, if have failed a class, if they have thought about changing careers or even leaving the university, if what they want is to finish their studies successfully, they will finish the two things about the studies of another school and another about the social stratum in which they find themselves.

3.1 Association Rules

When applying the algorithm A priori, using the arules package [19] of the R language. For the generation of the rules, the most common validation measures were considered: support, trust and lift. Thus, a support (s) of at least 0.25 and a confidence (c) of at least 0.9 was specified, and a lift (l) of at least 1, which resulted in 37 rules in total, as follows: describe the ten rules considered most relevant, accompanied each of these of the support, trust and count of the times the rule is correctly applied (co) and lift (l). In Table 2, the values for each of the attributes of the data set can be verified. The 10 rules of interest presented contribute to the knowledge generated, since the value of lift obtained in each of them is greater than 1, which means that their associations are statistically dependent, there is an association between the elements involved.

Rule 1

{dedication = high, change.career = nothing} => {leaving.university = nothing} s=0.39 c=0.93 co=56, l=1.41

Rule 1 indicates that if a student's dedication is high and he has not thought about changing his career, then he does not think about dropping out of the university at all.

Rule 2

{dedication = high, selected.career = yes, change.carrera = nothing => {baja.universidad = nothing} s=0.34 c=0.92 co=49 l=1.40

Rule 2 indicates that if a student's dedication is high, his career was selected and he has not thought about changing his career, then he does not think about unsubscribing from the university.

Rule 3

{absorption = high} => {uwess = high} s=0.34 c=0.94 co=49 l=2.25

Rule 3 indicates that if the absorption in a student is high then the psychological engagement with the studies is also high.

Rule 4

{uwess = high, financing = parents} => {dedication = high} s=0.29 c=0.95 co=42 l=1.49

Rule 4 indicates that if the psychological connection in a student is high and the financing of their studies is on the part of their parents, then the dedication is high.

Rule 5

{absorption = high, selected.career = yes} => {uwess = high} s=0.26 c=0.95 co=38 l=1.52

Rule 5 indicates that if the absorption in a student is high and his career was selected, then the psychological connection with the studies is high.

Rule 6

{absorption = high, financing = parents => {uwess = high} s=0-26 c=0.94 co=37 l=2.26

Rule 6 indicates that if the absorption in a student is high and the financing of their studies is on the part of their parents, then the psychological connection with the studies is high.

Rule 7

{dedication = high, absorption = high, selected.career = yes} => {uwess = high} s=0.26 c=0.94 co=37 l=2.26

Rule 7 indicates that if the dedication in a student is high and the absorption is also high, as well as that the career was selected, then the psychological engagement with the studies is high.

Rule 8

{change.carrera = nothing, finish.successful.studies = much} => {leaving.university = nothing} s=0.28 c=0.90 co=40 l=1.37

Rule 8 indicates that if a student has not thought about changing careers and thinks that he will finish his studies successfully, he does not think about canceling his university career.

Rule 9

{age = 18yLess21} => {financing = parents} s=0.27 c=0.92 co=39 l=1.30

Rule 9 indicates that if the student's age is over 18 and under 21, the financing of their studies is on the parents' side.

Rule 10

{social class= lowhigh} => {selected career. = yes} s=0.27 c=0.90 co=39 l=1.20

Rule 10 indicates that if the social stratum is low high, the career they studied has been the one selected.

3.2 Interactive Visual Representation of the Association Rules

The rules of association are very useful in practical applications, however the number of rules that are generated can be very large, since hundreds of them can be generated, so the review and analysis of them can be a very complex task , almost impossible. A convenient way to explore the rules is through its interactive graphical visualization, in this way it is possible to identify in a more simple way which rules are considered interesting.

Determining interest in a rule is not a simple task, one rule may be interesting for one analyst but not another, so the interest in a rule is considered subjective, depends on the knowledge and interests of the analyst [20]. Through the generation of visual association rules it is possible to facilitate this rule selection task. For the development of interactive visual association rules, the arulesViz package [21] of R. was used.

Next, the rules generated in a visual manner are presented for this case study. Figure 1 shows the 37 resulting rules in a general way, through this graph it is not possible to interact, which makes it difficult to explore the rules. Figure 2 shows the same graphic with the interactive property, the pink colored nodes, refer to the 37 rules, where the size of these circles depends on the number of occurrences of that rule, the green nodes are shown on the labels, where the attribute and the value it takes are specified.

These nodes when joined by arrows and their directions, allow the identification and reading of the rules. The left part of the rule are the attributes (green nodes) that go in the direction of the rule (pink node), the right side of the rule is indicated by the arrow in the exit direction of the rule (pink node) towards some attribute (green node). The labels shown in these diagrams are in Spanish, since the data set is in this language, so you can consult Table 2 to check the correspondence of the data.

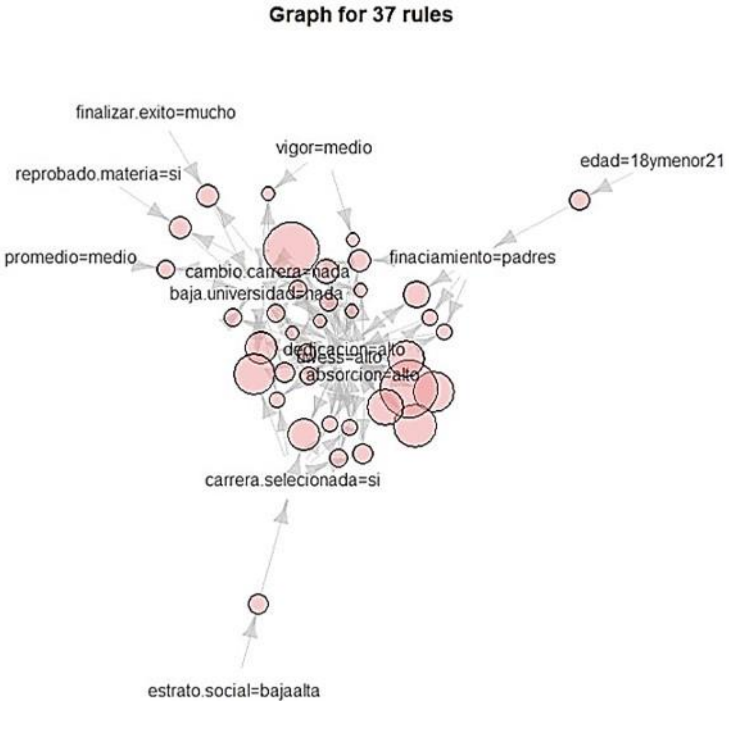


Fig. 1. Association Rules Chart.

In Figure 1 where the general diagram is observed, it is in the lower left, the case of Rule 1, for a better description is presented Figure 3, where this rule is focused (indicated by the pink node) receives as input of two attributes (indicated by green nodes): dedication = high and change.career = nothing, resulting in a red dotted line, directed to the attribute leaving.university = nothing (green node). This Rule 1 indicates that if a student's dedication is high and he has not thought about changing his career, then he does not think about dropping out of the university at all. It is possible to manipulate the visualization of these rules and find the routes that the rule explains when selecting and moving the mouse by the different elements that make up the rules.

Rule 1

{dedication = high, change.carrera = nothing } => { leaving university = nothing }

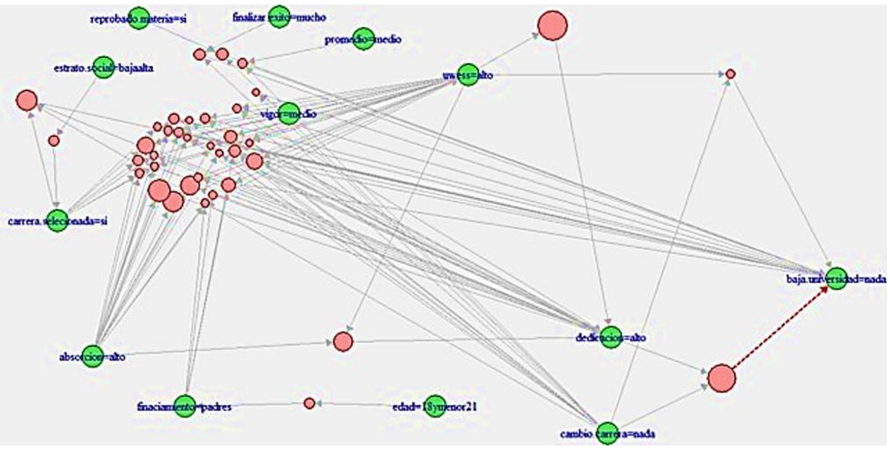


Fig. 2. Interactive Association Rules Chart.

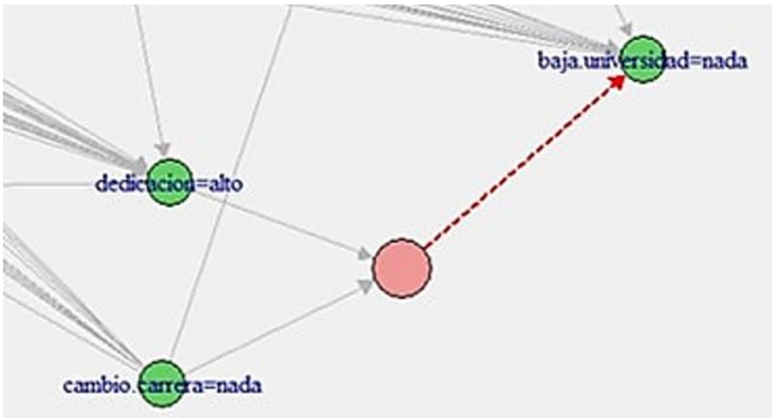


Fig. 3. Rule 1 through Visual Association.

Another rule that is observed is Rule 9 (See Figure 4) where it is indicated that if the student's age is greater than 18 and less than 21, the financing of their studies is on the part of the parents:

Rule 9

{age = 18yLess21} => {financing = parents}

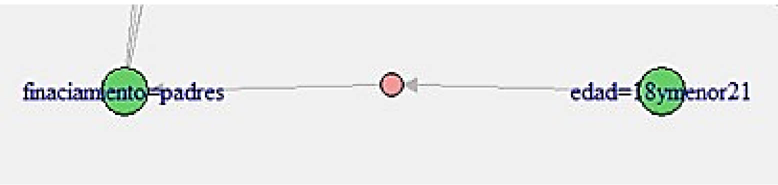


Fig. 4. Rule 9 through Visual Association.

The nodes that represent Rule 1 and Rule 9 (pink nodes) have different sizes (see Figure 3 and Figure 4), the node of Rule 1 (co = 56) is larger than the node of Rule 9 (co = 39), since it depends on the number of occurrences that the rule is fulfilled.

4 Conclusions

The association rules allow finding interesting results as a first approach to the data, since they help to solve data mining tasks of descriptive type. It is considered useful when the nature of the data is categorical, as is the case of the data generated by instruments such as questionnaires, surveys, among others. They allow the reading of the data in a more natural way with respect to other techniques, although when executing this type of algorithms for the generation of association rules, inconveniences such as the number of rules can be presented, since hundreds of algorithms can be generated.

The selection of the rules can be complicated for the analysts, due to the number of rules, besides that the election of the various rules of interest from one analyst to another, up to a certain point is subjective. One way to help a better selection of the rules of association is through its visual representation, and one step further its manipulation, that is, the unfolding of the rules interactively. In this study, this form of knowledge discovery was selected given the nature of the data.

The objective of this study was to find rules of association of interest to determine the relationship of the dimensions of the psychological connection or student engagement with other selected variables that have an impact on the decision of stay and continuity of the students.

This is an approach to this type of studies making use of this technique and it is intended to add more elements that allow knowing more about the psychological connection with the studies. The resulting rules indicate that the dedication variable is the one that has more presence in this connection with the studies to maintain and conclude successfully, that is, the meaning is important for their studies, feeling excited and proud of their career.

References

1. Salanova, M., Schaufeli, W.B.: El Engagement de los empleados un reto emergente para la dirección de recursos humanos. Estudios Financieros 261, 109–138. http://www.wilmar-schaufeli.nl/publications/Schaufeli/221.pdf (2004)
2. Schaufeli, W.B., Bakker, A.: Utrecht Work Engagement Scale (UWES). Escala de Engagement en el trabajo de Utrecht, Ocuppational Health Psychology Unit: Utrech University.http://www.wilmarschaufeli.nl/publications/Schaufeli/Test%20Manuals/Test_manual_UWES_Espanol.pdf (2003)
3. Hernández, J., Ramírez, M.J., Ferri, C.: Introducción a la Minería de datos. Madrid, España: Prentice Hall (2008)
4. Correa, J.C.: Gráficos Estadísticos con R. Universidad Nacional-Sede Medellín (2002)
5. R Development Core Team: A language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org (2016)
6. Reyes-Nava, A. Flores-Fuentes, A., Alejo, R., Rendón-Lara, E.: Minería de datos aplicada para la identificación de factores de riesgo en alumnos. Research in Computing Science 139, 177–189 (2017)
7. Rodríguez-Maya, N., Lara-Álvarez, C., May-Tzuc, O., Suárez-Carranza, B.A.: Modeling Student' Dropout in Mexican Universities. Research in Computing Science 139, 163–175 (2017)
8. Márquez-Vera, C., Romero-Morales, C., Ventura-Soto, S.C., Ventura, S.: Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. IEEE-RITA 7(3) (2012)
9. González-Marrón, D., Enciso-Gonzalez, A., Hernandez-Gonzalez, A.K., Gutierrez-Franco, D., Guizar-Barrera, B., Marquez-Callejas, A.: Evaluación de parámetros de encuesta de ingreso del CENEVAL para alumnos candidatos a ingresar al nivel superior, caso de estudio ITP. Research in Computing Science 139, 135-147 (2017)
10. Ivkovic, S., Yearwood, J., Stranieri, A.: Visualizing Association Rules for feedback within the legal system. ACM 1-58113-747-8. Edinburgh, Scotland, UK (2003)
11. Lee, J., Han, J., Chi, K.: Mining Quantitative Association Rule of Earthquake Data. In: International Conference on Convergence and Hybrid Information Technology. Daejeon, Korea (2009)
12. Morales-Mateos, E.Y., Hernández-Aguilar, J.A., Ochoa-Ortíz Zezzatti, C.A., López Garrido M.A.: A Comparison Represented in the Form of Radar of University Student Engagement in Degrees in Technologies. Research in Computing Science 122, 141–151 (2016)
13. Lopez-Garrido, M.A., Hernández-Aguilar, J.A., Ochoa-Ortiz Zezzatti, C.A., Morales-Mateos, E.Y., González-Constantino, C.: Comparative Study of Learning Strategies of Bachelor Students in Nursing. Research in Computing Science 122, 153–162 (2016)
14. Guisande, C., Vaamonde, A.: Gráficos estadísticos y mapas con R. España: Ediciones Díaz de Santos (2013)
15. Exposito, C., Exposito, A., López, I., Melián, B., Moreno, J.: Minería de patrones de asociación. Departamento de Ingeniería Informática y de Sistemas. Universidad de La Laguna, https://campusvirtual.ull.es (2018)
16. Haro, V., Péres, W., Saquicela, V.: Bibliomining para descubrir reglas de asociación en el Centro de Documentación Regional "Juan Bautista Vázquez". Departamento de Ciencias de la Computación, Universidad de Cuenca, http://dspace.ucuenca.edu.ec/bitstream/123456789/26350/1/TICEC_2016_20.pdf (2016)
17. Pincho, J.: Métodos de clasificación basados en asociación aplicados a sistemas de recomendación. Universidad de Salamanca, https://gredos.usal.es/jspui/bitstream/10366/83342/1/DIA_PinhoLucasJ_Métodosdeclasificación.pdf (2010)

18. Cuñer, N.: Escala de Inteligencia para Niños de Wechsler, WISC-IV. Diccionario de Psicome-tría. Montevideo, Uruguay, http://psicologos.org.uy/documentos13/20130926_Diccionario%20de%20psicometria.pdf (2013)
19. Hahsler, C., Buchta, C., Gruen, B., Hornik., K., Johnson, I., Borgelt, C.: Package 'arules'. https://cran.r-project.org/web/packages/arules/arules.pdf (2018)
20. Liu, B., Hsu, W., Wang, K., Chen, S.: Visually Aided Exploration of Interesting Association Rules. In: Zhong N., Zhou L. (eds), Methodologies for Knowledge Discovery and Data Mining. PAKDD 1999. Lecture Notes in Computer Science, vol. 1574. Springer, Berlin, Heidelberg (1999)
21. Hahsler, C., Tyler, G., Chelluboina, S.: Package 'arulesViz'. https://cran.r-project.org/web/packages/arulesViz/arulesViz.pdf (2018)

ISSN 1870-4069

A Framework for the Construction of a Historical Dictionary for Arabic

Rim Laatar, Chafik Aloulou, Lamia Hadrich Belguith

MIRACL Laboratory, ANLP Research Group,
Faculty of Economics and Management, University of Sfax, Sfax, Tunisia
rimlaatar@yahoo.fr
{chafik.aloulou,l.belguith}@fsegs.rnu.tn

Abstract. Arabic is one of the oldest Semitic languages in the world. But despite its rich historical heritage, Arabic is still bereft of a historical dictionary which traces the first use of its words and the evolution of their meanings and structures. Therefore, creating such a dictionary is of a great importance for the Arab world as it bridges the gap between its present and its past. This task should undergo several stages and requires a lot of effort. In this paper, we present our framework to help the linguists create a historical dictionary for Arabic. For this aim, we propose a platform which helps to trace the evolution of the meanings of a given word throughout time. The developed system allows the user to extract the meaning of an Arabic word according to the historical period in which it appeared. It also provides information about the oldest date of use of the word with a textual example in which it first appeared, and the first place where the word was used.

Keywords: Arabic historical dictionary, word sense disambiguation, word embedding, old Arabic, classical Arabic, modern standard Arabic.

1 Introduction

The human language is subject to a large number of factors and influences. The latter contribute to its development and to the evolution of its vocabulary and constructions. It may also lead to its erosion and fragmentation, or to its total extinction. It develops and renews itself when it finds the conditions that guarantee its development and may fade away when it is neglected by people (lack of use, forgetfulness, etc.).

Thus, the historical events and the political conditions that humanity has experienced have had a decisive impact on the subdivision of the languages of the ancient worls. In fact, each language can be divided into species and groups themselves, each giving rise to several languages linked together by historical and geographical bonds. In order to safeguard their languages, nations have resorted to the rooting of the language and the establishment of its history by means of historical dictionaries. According to Al-Said [2], a historical dictionary is a general dictionary of language which draws its importance from the human heritage gathered from sciences, arts and letters from the different ages and places. It studies the evolution of the construction of words and their meanings through the chronological stages the language has undergone. The historical dictionaries of a language are thus considered to be as the language body which helps to understand the entire human heritage.

However, despite its richness, the Arabic language does not yet have a historical dictionary which helps to monitor the semantic development of the Arabic language throughout history, and to understand its knowledge and scientific heritage correctly. Indeed, words in Arabic have gone through a historical process of growth marked by significance and expectation. Accordingly, certain words have changed in terms of vocations over time and others have completely disappeared from literature.

As a matter of fact, the evolution of the Arabic language from antiquity to the present day has given birth to several linguistic registers. According to Al-Said [3], the Arabic language can be divided into three periods: *(1)* Old Arabic, which is not used currently. It is found in ancient literary works (mainly poems).*(2)* Classical Arabic or literary Arabic, which is the language of the Quran. It is spread through Islamic conquests.*(3)* Modern Standard Arabic (MSA) is the official language of all the Arab countries [16].

In view of that, language evolves and changes by time in terms of its sounds, rules, and especially its meanings. Accordingly, the meanings of words is not fixed, but in a constant change and evolution from one age to another.

In this work, we propose to develop a framework for the construction of a historical dictionary for Arabic. In fact, creating such a dictionary has to go through several stages and requires a lot of effort. One of these stages is the extraction of the appropriate sense of a given word according to its appearance in the document. Recently, several studies have focused on disambiguating words in Modern Standard Arabic, but there seems to be no work concerned with disambiguating Arabic words according to their historical period in which they appeared. The principal objective is to disambiguate words appearing in Old and Classical Arabic in order to study the semantic evolution of each word of the language through its historical ages. Therefore, the main contributions of this paper are as follows:

- Propose a method which helps to extract automatically the meaning of a given word according to its historical period. This method aims not only to identify what a word means in a given context, but also to disambiguate it according to the historical period in which it appeared.
- Suggest a method that helps to give clear and precise information about the oldest use of a given word, its first date of appearance, its users, its first places of appearance, and the date of its sense evolution.
- Help the linguists build a historical dictionary for Arabic by proposing a tool which enables to automatically extract the meaning of a given word and describe its evolution historically and geographically.

The rest of this paper is divided as follows. Section 2 reviews some works on existing historical dictionaries in different languages as well as some previous attempts to build a historical dictionary for Arabic. It also gives an overview of the works which focus on Arabic word sense disambiguation. Section 3 explains our method to help the linguists to construct such a dictionary. Section 4 presents the developed system. Experimental results are tackled in section 5. In section 6 we draw a conclusion.

2 State of the Art

2.1 Historical Dictionaries Background

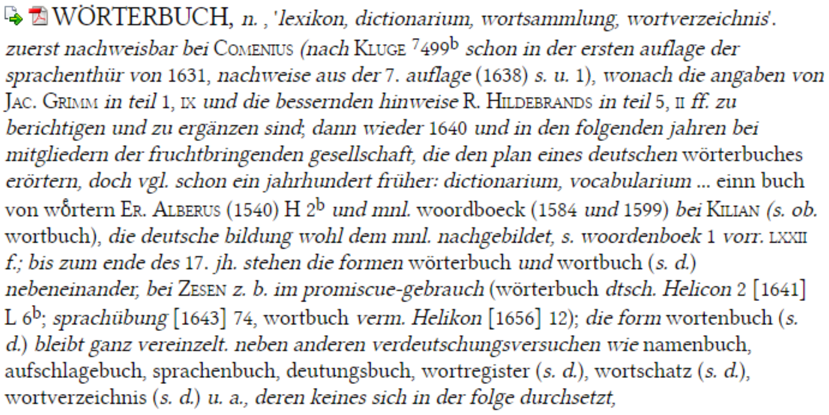
The idea of creating historical dictionaries appeared during the second half of the 19^{th} century following the appearance of the method of historical analysis [2]. The primordial objective of creating historical dictionaries was to gather information about the words of the language by studying their evolution over time in terms of phonetics, structure, form and meaning. Several international projects have been launched in different countries whose purpose was to develop a historical dictionary. The first attempt was with the German Historical dictionary in 1838. Then some other endeavors were with the Dutch Historical dictionary in 1849 and with the English Dictionary in 1849.

The German Historical Dictionary. The German historical dictionary *Deutsches WörterBuch* (DWB) is the most important German historical dictionary since the 16^{th} century. It is also called the Grimm dictionary, referring to the names of its creators the Grimm brothers (Jacob and Wilhelm Grimm), who began working on it in 1838 with more than 80 collaborators. It is a historical dictionary that traces the history of each word using many quotes. Indeed, the purpose of this dictionary is to analyze and explain exhaustively the origin and use of each German word [2]. Figure 1 shows an excerpt[1] from the dictionary for the word ' WÖRTERBUCH/Dictionary'.

According to this excerpt, we note that the search for a given word in the German electronic history dictionary makes it possible to present the various synonyms of the word in focus as well as its inflected forms. The articles which contain this word are classified by their date of appearance, their links, and their search links in other dictionaries. This work was manually and did not depend on NLP tools.

The Dutch Historical Dictionary WNT. The dictionary of the Dutch language (WNT) “Woordenboek der Nederlandsche Taal” was announced in 1849. The WNT contains about 95 000 main entries and about 1.7 million citations. This dictionary indicates for each word the grammatical characteristics, the origin, the meaning(s), their use in compounds, sentences and proverbs, and derivations [2].

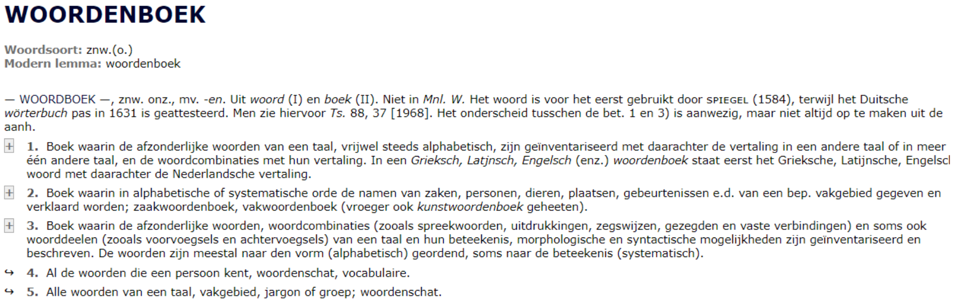
[1] `http://woerterbuchnetz.de/cgi-bin/WBNetz/wbgui_py?sigle=DWB&mode=Vernetzung&lemid`



WÖRTERBUCH, *n. , 'lexikon, dictionarium, wortsammlung, wortverzeichnis'. zuerst nachweisbar bei* COMENIUS *(nach* KLUGE $^{7}499^{b}$ *schon in der ersten auflage der sprachenthür von* 1631, *nachweise aus der* 7. *auflage* (1638) *s. u.* 1), *wonach die angaben von* JAC. GRIMM *in teil* 1, IX *und die bessernden hinweise* R. HILDEBRANDS *in teil* 5, II *ff. zu berichtigen und zu ergänzen sind; dann wieder* 1640 *und in den folgenden jahren bei mitgliedern der fruchtbringenden gesellschaft, die den plan eines deutschen* wörterbuches *erörtern, doch vgl. schon ein jahrhundert früher: dictionarium, vocabularium* ... einn buch von wo̊rtern ER. ALBERUS (1540) H 2^{b} *und mnl.* woordboeck (1584 *und* 1599) *bei* KILIAN *(s. ob.* wortbuch), *die deutsche bildung wohl dem mnl. nachgebildet, s. woordenboek* 1 *vorr.* LXXII *f.; bis zum ende des* 17. *jh. stehen die formen* wörterbuch *und* wortbuch (*s. d.*) *nebeneinander, bei* ZESEN *z. b. im promiscue-gebrauch* (wörterbuch *dtsch. Helicon* 2 [1641] L 6^{b}; *sprachübung* [1643] 74, wortbuch *verm. Helikon* [1656] 12); *die form* wortenbuch (*s. d.*) *bleibt ganz vereinzelt. neben anderen verdeutschungsversuchen wie* namenbuch, aufschlagebuch, sprachenbuch, deutungsbuch, wortregister (*s. d.*), wortschatz (*s. d.*), wortverzeichnis (*s. d.*) *u. a., deren keines sich in der folge durchsetzt,*

Fig. 1. Excerpt from the DWB dictionary.

Figure 2 shows an excerpt[2] from the dictionary for the word ' WOORDENBOEK /Dictionary'. This figure indicates that the WNT dictionary provides for each word the part of speech, the lemma, the meaning, the date of appearance, the inflected forms, the quotations and the origin of this word. Once the word is composed, the dictionary displays the previous information for each word that composes it. This work was manually and did not depend on NLP tools.



WOORDENBOEK

Woordsoort: znw.(o.)
Modern lemma: woordenboek

— WOORDBOEK —, znw. onz., mv. *-en*. Uit *woord* (I) en *boek* (II). Niet in *Mnl. W.* Het woord is voor het eerst gebruikt door SPIEGEL (1584), terwijl het Duitsche *wörterbuch* pas in 1631 is geattesteerd. Men zie hiervoor *Ts.* 88, 37 [1968]. Het onderscheid tusschen de bet. 1 en 3) is aanwezig, maar niet altijd op te maken uit de aanh.

⊞ **1.** Boek waarin de afzonderlijke woorden van een taal, vrijwel steeds alphabetisch, zijn geïnventariseerd met daarachter de vertaling in een andere taal of in meer één andere taal, en de woordcombinaties met hun vertaling. In een *Grieksch, Latjnsch, Engelsch* (enz.) *woordenboek* staat eerst het Grieksche, Latijnsche, Engelscl woord met daarachter de Nederlandsche vertaling.

⊞ **2.** Boek waarin in alphabetische of systematische orde de namen van zaken, personen, dieren, plaatsen, gebeurtenissen e.d. van een bep. vakgebied gegeven en verklaard worden; zaakwoordenboek, vakwoordenboek (vroeger ook *kunstwoordenboek* geheeten).

⊞ **3.** Boek waarin de afzonderlijke woorden, woordcombinaties (zooals spreekwoorden, uitdrukkingen, zegswijzen, gezegden en vaste verbindingen) en soms ook woorddeelen (zooals voorvoegsels en achtervoegsels) van een taal en hun beteekenis, morphologische en syntactische mogelijkheden zijn geïnventariseerd en beschreven. De woorden zijn meestal naar den vorm (alphabetisch) geordend, soms naar de beteekenis (systematisch).

↪ **4.** Al de woorden die een persoon kent, woordenschat, vocabulaire.

↪ **5.** Alle woorden van een taal, vakgebied, jargon of groep; woordenschat.

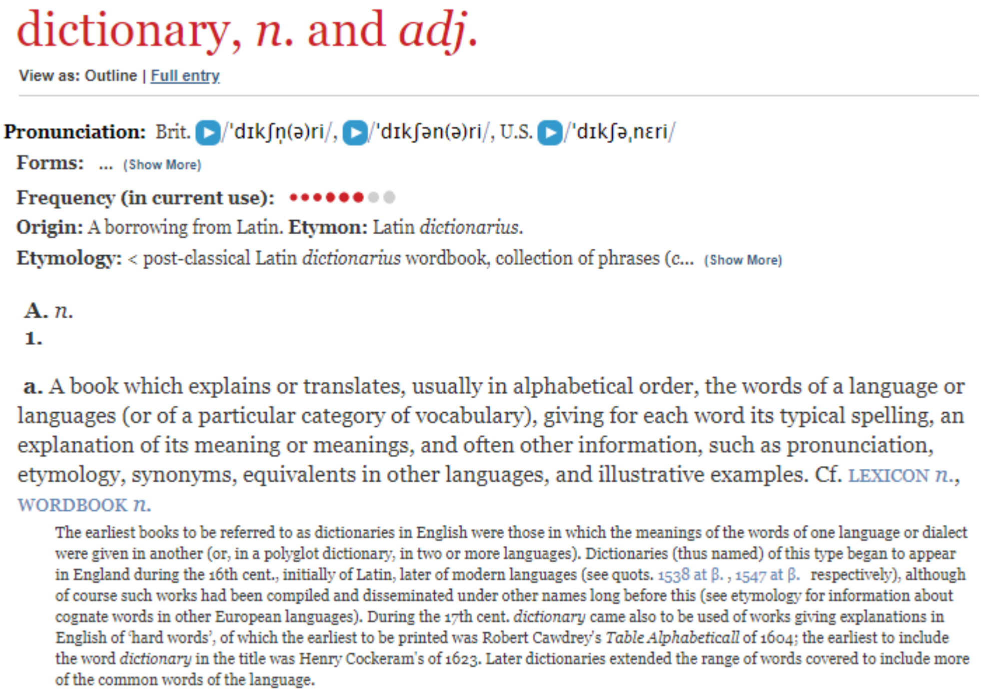
Fig. 2. Excerpt from the WNT dictionary.

The English Historical Dictionary OED. The Oxford English Dictionary (OED) is a reference dictionary for the English language. It is published by The Oxford University Press. The essential task of a dictionary would be to trace the history and the trajectory of each word, illustrating with quotations the nuances of meaning and uses that have emerged over time [2]. Figure 3 shows an excerpt[3] from the dictionary for the word 'Dictionary'. This excerpt shows

[2] http://gtb.inl.nl/iWDB/search?actie=article&wdb=WNT&id=M087164

[3] http://www.oed.com/view/Entry/52325

how the search for a word in the OED online dictionary gives information on the different pronunciation of this word, its etymology and its different meanings.



dictionary, *n.* and *adj.*

View as: Outline | Full entry

Pronunciation: Brit. ▶/ˈdɪkʃn̩(ə)ri/, ▶/ˈdɪkʃən(ə)ri/, U.S. ▶/ˈdɪkʃəˌnɛri/

Forms: ... (Show More)

Frequency (in current use): ●●●●●●●●

Origin: A borrowing from Latin. **Etymon:** Latin *dictionarius.*

Etymology: < post-classical Latin *dictionarius* wordbook, collection of phrases (*c*... (Show More)

A. *n.*

1.

a. A book which explains or translates, usually in alphabetical order, the words of a language or languages (or of a particular category of vocabulary), giving for each word its typical spelling, an explanation of its meaning or meanings, and often other information, such as pronunciation, etymology, synonyms, equivalents in other languages, and illustrative examples. Cf. LEXICON *n.*, WORDBOOK *n.*

The earliest books to be referred to as dictionaries in English were those in which the meanings of the words of one language or dialect were given in another (or, in a polyglot dictionary, in two or more languages). Dictionaries (thus named) of this type began to appear in England during the 16th cent., initially of Latin, later of modern languages (see quots. 1538 at β. , 1547 at β. respectively), although of course such works had been compiled and disseminated under other names long before this (see etymology for information about cognate words in other European languages). During the 17th cent. *dictionary* came also to be used of works giving explanations in English of 'hard words', of which the earliest to be printed was Robert Cawdrey's *Table Alphabeticall* of 1604; the earliest to include the word *dictionary* in the title was Henry Cockeram's of 1623. Later dictionaries extended the range of words covered to include more of the common words of the language.

Fig. 3. Excerpt from the OED dictionary.

According to the study of the state of the art on the creation of historical dictionaries of other languages, we can see that the objectives of these dictionaries revolve around the study of the evolution of the meanings of the words since their first appearance.

Arabic Historical Dictionary. As for the historical dictionary for Arabic language, a first attempt of the historical dictionary was launched and directed by August Fischer in Egypt in 1935.

In April 2004, the Historical Dictionary of the Arabic Language Committee was founded by a decision of the Arab Language and Science Counselling Union in Cairo (Egypt). This project aims to create a historical dictionary of the Arabic words and their uses in order to indicate the change of their meanings through time and space [1]. It is still under the studying stage of the corpus.

A third attempt of the project was by the Arab Center for Research in Doha in 2013. The initial steps have been to prepare a reference bibliography of the sources of the linguistic corpus of the dictionary.

As a best of our knowledge, the creation of historical dictionaries for other languages does not use the natural language processing (NLP) tools, whereas for the attempts that have been made to create the historical dictionary for Arabic, they were not successful.

As for the attempt at the Arab Center for Research in Doha, the team pointed out the need to rely on NLP tools in the different steps of building the desired dictionary. In this context, Khalfallah et al.[12], proposed a platform of Automatic Natural Language Processing (ANLP) tools which permit the automatic indexing and research from Arabic texts corpus. They developed a system which allows to extract contexts from the entered corpus and to assign meaning by the user [13]. Another primordial step in the creation of a historical dictionary is by determining the correct meaning of a word in a given context, also known as WSD.

2.2 Word Sense Disambiguation: State of the Art

WSD Approaches. WSD is a natural language processing task of identifying the particular word senses of polysemous words used in a sentence [28]. WSD has become a prominent research area in the field of NLP. There are three main methods to WSD: knowledge-based method, supervised and unsupervised method [24].

Knowledge based methods rely on dictionaries, thesaurus, and knowledge to extract the definition of the ambiguous word. Unsupervised methods are based on training sets and do not use any structured resource while supervised methods are based on manually sense-annotated data sets [22]. Recently, the great advance in distributed semantics paved the way for the appearance of word embedding. Word embedding enables the computation of semantically related words. It can also be used to represent other linguistic units, such as phrases and short texts. Yet the problem of WSD has been approached from various perspectives in the context of word embeddings [25].

Word embeddings are of a major importance as they exhibit certain algebraic relations and can, therefore, be used for meaningful semantic operations. The latter include computing word similarity and capturing lexical relationships [23].

Recently some works have focused on word representation in vector space for the Arabic language such as[29,27,9].

However, most of the works used word embedding or any other techniques related to WSD were applied to Latin languages like English and French. But in the last decade, some attempts were applied to the Arabic language.

Arabic WSD Approaches. To the best of our knowledge, there seems to be no work concerned with disambiguating Arabic words according to their historical period in which they appeared. Hence, the idea of disambiguating words appearing in Old and Classical Arabic in order to create a historical dictionary is original.

In fact, all the works that focus on Arabic Word Sense Disambiguation are concerned with identifying the meaning of words in MSA.

In this Section, we are going to review the existing works related to the disambiguation of words in MSA. The work proposed by Bouhriz et al.[8] takes into consideration the local context and the global context defined by the full text during the disambiguation process. They have represented local context, global context and each sense of the ambiguous word with the help of vectors. Then, the appropriate sense of the target word is the sense that has the closest semantic proximity to its local and global context.

Alian et al.[5], relied on Arabic Wikipedia to extract the different meanings of the ambiguous word. They have applied Vector Space Model as a mathematical representation for documents. Vector Space Model serves to represent each retrieved texts from Wikipedia as a vector. Then, each text represented with the help of vectors is compared with the context of the word using cosine distance. The appropriate sense of the ambiguous word is therefore the concept of having the most cosine similarity.

Another method was proposed by Menai[17]. They have used genetic algorithms to solve word sense disambiguation problem. They tested their approach using a sample text in Arabic then they compared with naïve Bayes classifier [6].

Zouaghi et al.[30] have proposed an approach based on information retrieval measures. They have generated the contexts of use for each sense of the ambiguous word using its glosses. Then, the most probable sense is chosen by measuring the similarity between the different contexts generated and the current context of the ambiguous word.

The method proposed by Zouaghi et al.[31] is a hybrid method that combines unsupervised and knowledge based methods. They have used a context Matching algorithm that measures the similarity between the contexts of use corresponding to the glosses of the target word and the original sentence [6].

The most recent work proposed by Alkhatlan et al.[7] aims to disambiguate Arabic words using Arabic Wordnet and word embeddings. The main idea of this work is to represent each sense of the ambiguous word by a vector based on word2vec and Glove. The system proposed by Alkhatlan et al.[7] lists all the synsets which represent the ambiguous word along with their similarity to the context. It also chooses the synset that has the maximum similarity among synsets.

Table 1 presents a comparative study in the field of word sense disambiguation for Arabic. This comparison is performed using these criteria:

– The used method for WSD,
– The resources used to WSD,
– The testing data (number of ambiguous word used),
– The rate of precision.

Thanks to this study, we note that most of the works used a knowledge based approach for AWSD because these approaches provide a higher precision than the unsupervised approach.

[4] `https://sites.google.com/site/mouradabbas9/corpora`

Table 1. Comparative study of some AWSD approaches.

Author	WSD method	Used resources	Testing data	Precision
Bouhriz et al. [8]	Knowledge based AWSD	- Arabic Wordnet	A sample text in Arabic	74%
Alian et al. [5]	Knowledge based AWSD	- Arabic Wikipedia - Arabic Wordnet	7 ambiguous words	-
Menai [17]	Knowledge based approach	- Arabic WordNet - A sense annotated corpus	A sample text in Arabic	79%
Zouaghi et al. [30]	Knowledge based approach	- Arabic dictionary Alwassit - A collected corpus of 1500 Arabic texts	- 50 ambiguous words - 130 contexts of use for every word	73%
Zouaghi et al. [31]	Hybrid AWSD	- Arabic dictionary Alwassit - A collected corpus of 1500 Arabic texts	- 10 ambiguous words - 130 contexts of use for every word	79%
Alkhatlan et al. [7]	Knowledge based approach	- Arabic WordNet - Watan and Khaleej corpora[4]	- 10 ambiguous words - A collected corpus of 240 training samples	79%

3 Proposed Method

Our research consists in developing a framework to help linguists create a historical dictionary for Arabic. Thus, we have been inspired from the different outputs of the historical dictionaries already done and based on a description submitted by Doha site[5] of their aims behind the draft of the historical lexicon of Doha for the Arabic language. We have noticed that building the desired dictionary requires fundamental steps. One of them is by tracing the evolution of the meanings of a given word throughout time in addition to its historical information such as the date and the location of its first appearance. So we began by presenting the methodology of determining the meaning of a word in context, also known as Word Sense Disambiguation and then we detail the process of extraction of its historical information.

3.1 Methodology of Word Sense Disambiguation in Arabic

We propose here a method which permits to determine the meaning of an ambiguous word. This method aims not only to identify what a word means in

[5] https://www.dohadictionary.org/AR/Lexical_Services/Pages/Bibliography.aspx

a given context, but also to disambiguate it according to the historical period in which it appeared. Recently, word embedding has become a mainstay of natural language processing thanks to their ability to solve many NLP problems such as machine translation, sentiment analysis and even word sense disambiguation(WSD).

Word embeddings consist in building word representations in vector space based on the distributional hypothesis [11]: words that occur in the same contexts tend to have similar meanings [10].

In the spirit of representing words as vectors in a highly dimensional space, our method also benefits from word embedding to disambiguate Arabic words. This method is made up of two stages. The first one is about building an Arabic word embedding model using the skip gram technique [19]. The second one consists in calculating the similarities between the context of the ambiguous word and its definitions after representing, with the help of vectors, the context of the word to be disambiguated and its different glosses.

Arabic Word Embedding Model. The Word2vec tool [18] remains a popular choice benefited from its fast training and good results. In this work, we explore skip gram architecture to build neural word embeddings for Arabic. In fact, to build the word embedding model, we have used the Historical Arabic Dictionary Corpus (HADC) [4], which is originally designed to build a historical dictionary. It contains texts in Old Arabic, Classical Arabic and Modern Standard Arabic. A preprocessing step has been done before training word2vec model. First of all, we removed punctuation and non Arabic words. Then, we removed the stop words from the corpus based on a predifined list of stop words.

Similarity Calculation. To attribute for each ambiguous word its appropriate sense, we choose the sense with the closest semantic similarity to its local context. To measure the similarity between the context of use and each sense definition, three methods have been used. In what follows, we explain how to compute the semantic similarity among the context of use of the ambiguous word and its sense definition.

No weighting method. The simplest strategy to compare the context and the sense definition of the ambiguous word is by computing the sum of their words vectors. The similarity is subsequently measured for each meaning of the ambiguous word by using a cosine distance metric.

IDF weighting method. The core principle of this methodology is to assign a weight to each word in the context of the ambiguous word and its sense definition. These weights are based on the Inverse Document Frequency. The idea behind this is the word needed to determine most of the sentence's semantics usually have higher idf values [20]. The context vector (respectively sense vector) is represented by the sum of each word vector multiplied with its idf score.

To create a historical dictionary, we should take into consideration that certain words disappeared from the language and some new words appeared, and therefore the idf is calculated according to the periods in which the words appeared. Indeed, the corpus can be divided into three main periods. The first period contains texts in classical Arabic, the second period with texts in middle-age Arabic and the third period with texts in modern Arabic. Then, for each period, idf is calculated using the following formula [21]:

$$idf(w) = \log \frac{S}{WS}, \tag{1}$$

where w is a word that appeared in a specephic period, S is the total number of sentences in this period and WS is the number of sentences containing the word w. The similarity is subsequently measured by using a cosine distance metric.

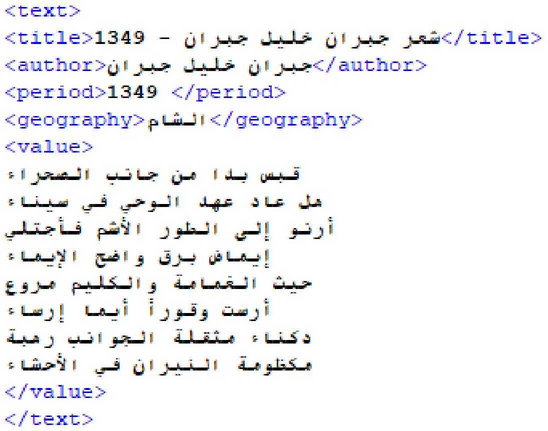
Word mover distance (WMD) method. WMD was introduced by [14]. WMD is a method that allows us to measure the distance between two documents (two sentences in our case). It takes into account the word's similarities in word embedding space. Indeed, we have used WMD to calculate the similarity between the context of use of the ambiguous word and its senses definition.

3.2 Methodology for Extraction of Historical Information for a Word

After the extraction of the meaning of the given word, we must determine the historical period when this word was first used. More precisely, our aim is to determine, for a given Arabic word, the date of its first appearance and the date of its sense's transformation. Then, we will store the history of Arabic words in an XML format [15].

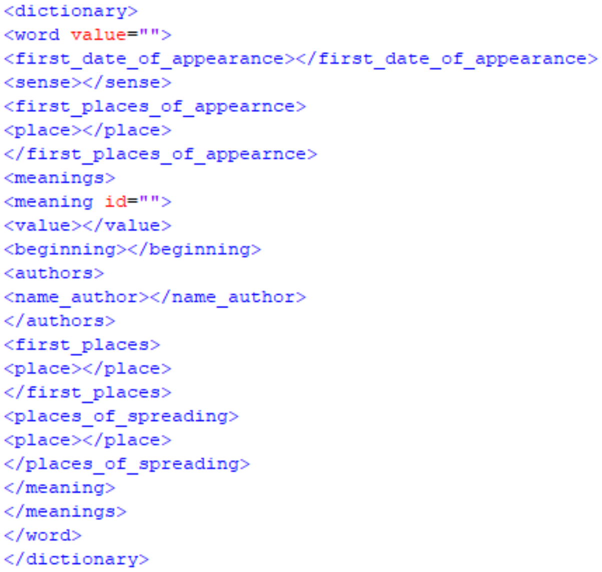
The first step is to represent the texts of the corpus in an XML format. The XML format allows us to capture the historical information of each document. In fact, the title of each document in the corpus is saved under the headings of: Author's name - Date of death. Therefore, the XML format of each document is automatically created by using the title of the document, the author and the period representing the date of death of the author. In fact, the date of death describes the historical period when that specific meaning was used. This could be explained with the fact that the date of the author's death gives a precise idea about the person's details.

However, the geography of appearance is extracted from Arabic Wikipedia. This step involves an XML description of all the text figures in the HADC corpus. The texts of the corpus are thus represented in the form of two files under different extension (TXT and XML). The TXT extension contains the value of the text and the XML extension contains the title, the author, the period and the geography of appearance of the text as well as its value (see figure 4).



```
<text>
<title>1349 - شعر جبران خليل جبران</title>
<author>جبران خليل جبران</author>
<period>1349 </period>
<geography>الشام</geography>
<value>
قبس بدا من جانب الصحراء
هل عاد عهد الوحي في سيناء
أرنو إلي الطور الأشم فأجتلي
إيماض برق واضح الإيماء
حيث الغمامة والكليم مروع
أرست وقوراً أيما إرساء
دكناء مثقلة الجوانب رهبة
مكظومة النيران في الأحشاء
</value>
</text>
```

Fig. 4. Structure of a XML file of a text extracted from the corpus.



```
<dictionary>
<word value="">
<first_date_of_appearance></first_date_of_appearance>
<sense></sense>
<first_places_of_appearnce>
<place></place>
</first_places_of_appearnce>
<meanings>
<meaning id="">
<value></value>
<beginning></beginning>
<authors>
<name_author></name_author>
</authors>
<first_places>
<place></place>
</first_places>
<places_of_spreading>
<place></place>
</places_of_spreading>
</meaning>
</meanings>
</word>
</dictionary>
```

Fig. 5. XML description of the dictionary of meaning.

The second step is to study the variation of the meanings of the word through time. Hence, the principle is as follows: for a given word, we first extract its meaning by applying our disambiguation method presented in section 3.1. Second, we stoke the historical information (period, place,user) related to the meaning of the word. This process will be repeated recursively for each document of the corpus containing the word being analyzed. Then, the meaning and the historical period in which this meaning was used is stoked in the XML format. It is worth pointing out that the XML model is automatically updated once the meaning of the word is found in an older document.

As for the word's first date of appearance, first place of appearance and origin, we will extract that automatically from the corpus. In fact, suppose that the texts of the corpus are historically classified from the oldest to the most recent ones, the first document containing the word in focus will reflect its first date of appearance, its origin and its user. We present an extract of the structure of our XML model in Fig. 5. It is necessary here to note that the texts of the corpus are lemmatized in order to identify the word's first occurrence. Finally, the XML model stores the following information:

– The oldest use of the word, together with its users, first date of appearance and first places of appearance,
– The meanings of the word historically classified according to its appearance in the corpus, specifying for each meaning the history of its oldest use, its oldest place of appearance and its users.

4 An Overview of the Developed System

In this section, we present the implementation details of our tool. The architecture of our system is shown in Fig 6.

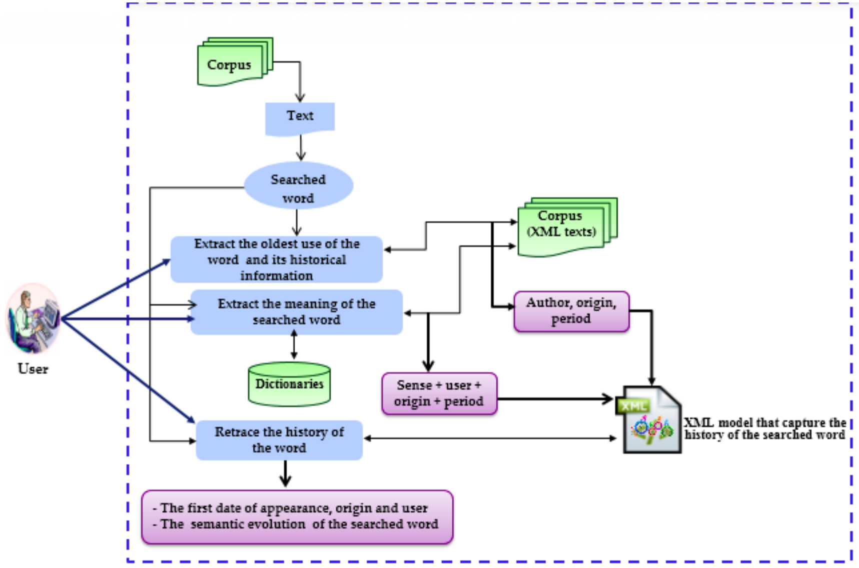

Corpus
Text
Searched word
Extract the oldest use of the word and its historical information
Corpus (XML texts)
Extract the meaning of the searched word
Author, origin, period
User
Dictionaries
Sense + user + origin + period
XML
XML model that capture the history of the searched word
Retrace the history of the word
- The first date of appearance, origin and user
- The semantic evolution of the searched word

Fig. 6. Architecture of the developed system.

To develop our tool, we used python programming language. We also used Gensim[6] toolkit to explore the pre-trained model, and PyQt4[7] toolkit to build our interface tool.

More precisely, the first version of our tool includes three main zones. The first one is called "The current text" in which the text will be uploaded to

[6] https://pypi.org/project/gensim/
[7] https://pypi.org/project/PyQt4/

display. The second one is called "Research operation" which represents the research possibilities (with criteria by entering a word to search, etc.). The last one contains three parts: "Historical information", "Research Sense", and "The Semantic evolution of a word".

Therefore, our tool allows users to execute the following tasks:

- Extract the first date of appearance of a given word, as well as its users and its first places of appearance. The historical information extraction interface is presented in Fig. 7.
- Extract the meaning of the given word according to the historical period in which it appeared. The word sense disambiguation interface is presented in Fig. 8.
- Generate the meanings of the word historically classified according to its appearance in the corpus, specifying for each meaning the history of its oldest use, its oldest place of appearance and its users.

5 Evaluation

To evaluate our work, we conduct a series of experiments regarding the ability of word embeddings to solve WSD problem.

- Our corpus of test consists of 183 texts that appear in different historical periods.
- These texts have been selected from the Historical Arab Corpus HADC and Open Source Arabic Corpora (OSAC) corpus [26].
- We have tested about 100 ambiguous words. For each ambiguous word, we have used AntConc[8] to extract its contexts of use from the test corpus.
- As we have previously mentioned, our test corpus contains documents that appear in different periods from Classical Arabic to Modern Arabic. Then, we have randomly extracted, for each period, 100 contexts of use for each ambiguous word.
- We have used 150-dimensional Skip gram word embeddings.
- Moreover, to extract the different meanings of the ambiguous word taking into account the historical period in which the word appeared in the document, we have used four Arabic dictionaries that describe the different historical periods of the Arabic language.
  - Tahdhib Allougha Dictionary[9], for Old Arabic by Abou Mansour Azhari,
  - Tej Alarous Dictionary[10] by Murtadha Zbidi, For Intermediate Arabic Dictionaries,

[8] `http://www.laurenceanthony.net/software/antconc/`

[9] AlAzhari, Abu Mansour, Refining the Language. Dar Alamaarif, Cairo, 1976.

[10] Zabidi, Sayed Mortadha, Tej Alarous, Kuwait Government Press and the National Council for Culture and Arts, Kuwait from 1965 to 2002.

- Alwassit dictionary[11] and dictionary of contemporary Arabic language[12] for modern Arabic.

Table 2. The average precision obtained with stop words removal from the corpus when training words vector.

Pre-processing step	Precision		
	Old Arabic	Classical Arabic	MSA
With stop words removal	42,5%	43,9%	49%
Without stop words removal	48.54%	47.48%	59.42%

Table 3. Results of disambiguation words according to its appearance in the document.

Method	Precision		
	Old Arabic	Classical Arabic	MSA
No weighting method	48.54%	47.48%	59.42%
MDA distance method	46,94%	44,10%	50,26%
IDF weighting method	48,89%	48,57%	63,44%

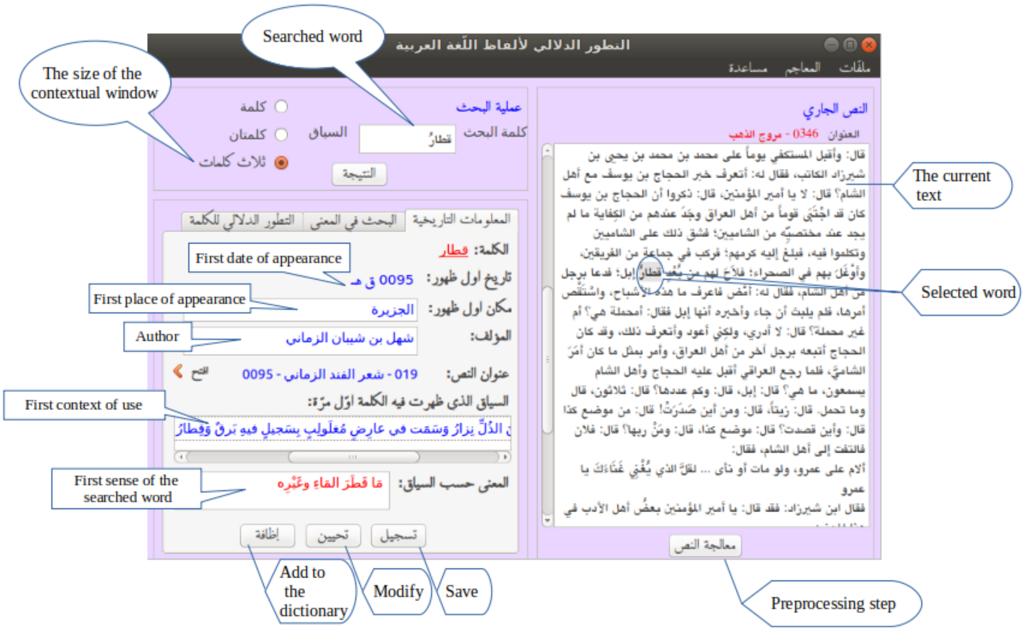
We have semi-automatically developed a structured electronic dictionary with an XML format containing the glosses of 100 ambiguous words in the Old Arabic. Similarly, we have developed a dictionary that contains the glosses of 100 ambiguous words extracted from Tej Alarouss. Thus, the last two dictionaries Tahdhib Alougha and Tej Alarous are manually structured because they have complex structures, which varies from one entry to another and are characterized by a quasi-absence of marker. For words in Modern Standard Arabic, the two dictionaries Alwaseet and Almouasera are used. Indeed, we have an HTML version of these two dictionaries.

These two dictionaries are distinguished by a set of markers facilitating the transformation of their raw content to a structured version in XML. Then we have automatically transformed them into a structured electronic XML format.

The first part of this evaluation is to test the impact of removing stop words from the corpus when training word vectors. Results are shown in table 2. For the word sense disambiguation task, we have noticed that without removal stop words from the corpus, our trained model exhibits stronger performances compared to the model obtained with trained corpus without stop words.

[11] The 5th Edition of the Alwasseet Dictionary published by the Arabic Language Complex, in Cairo in 2011.

[12] Mokhtar, Omar Ahmed, Modern Arabic Language, The Universe of Books, Cairo, 2008.



عملية البحث

كلمة البحث: قطارُ

السياق: ○ كلمة ○ كلمتان ● ثلاث كلمات

الكلمة: قطار

تاريخ اول ظهور: 0095 ق هـ

مكان اول ظهور: الجزيرة

المؤلف: شهل بن شيبان الزماني

عنوان النص: 019 - شعر الفند الزماني - 0095

السياق الذي ظهرت فيه الكلمة اوّل مرّة:

ن الذُلِّ نِزارٌ وسَمت في عارِضٍ مُغلَولِبٍ بِسَجيلٍ فيهِ بَرقٌ وَقِطارُ

المعنى حسب السياق: مَا قطَرَ الماءِ وغَيْرِه

النص الجاري

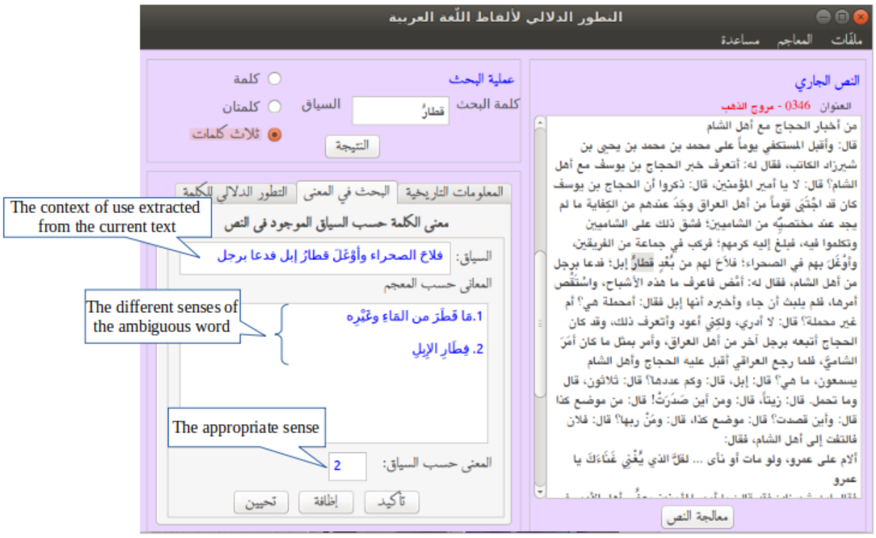
العنوان 0346 - مروج الذهب

قال: وأقبل المستكفي يوماً على محمد بن محمد بن يحيى بن شيرزاد الكاتب، فقال له: أتعرف خبر الحجاج بن يوسف مع أهل الشام؟ قال: لا يا أمير المؤمنين، قال: ذكروا أن الحجاج بن يوسف كان قد اجْتَبَى قوماً من أهل العراق وجَدَ عندهم من الكِفاية ما لم يجد عند مختصيِّه من الشاميين؛ فشق ذلك على الشاميين وتكلموا فيه، فبلغ إليه كرمهم؛ فركب في جماعة من الفريقين، وأوْغَلَ بهم في الصحراء؛ فلاَحَ لهم من بُعْدٍ قطارٌ إبل؛ فدعا برجل من أهل الشام، فقال له: أمْض فاعرف ما هذه الأشباح، واسْتَقْص أمرها، فلم يلبث أن جاء وأخبره أنها إبل فقال: أمحملة هي؟ أم غير محملة؟ قال: لا أدري، ولكِنِي أعود وأتعرف ذلك، وقد كان الحجاج أتبعه برجل آخر من أهل العراق، وأمر بمثل ما كان أمَرَ الشاميَّ، فلما رجع العراقي أقبل عليه الحجاج وأهل الشام يسمعون، ما هي؟ قال: إبل، قال: وكم عددها؟ قال: ثلاثون، قال وما تحمل. قال: زيتاً، قال: ومن أين صَدَرَتْ! قال: من موضع كذا قال: وأين قصدت؟ قال: موضع كذا، قال: ومَنْ ربها؟ قال: فلان فالتفت إلى أهل الشام، فقال:

ألام على عمرو، ولو مات أو نأى ... لقَلَّ الذي يُغْنِي غَنَاءَكَ يا عمرو

فقال ابن شيرزاد: فقد قال: يا أمير المؤمنين بعضُ أهل الأدب في

Fig. 7. Historical information extraction interface.


The context of use extracted from the current text
The different senses of the ambiguous word
The appropriate sense
التطور الدلالي لألفاظ اللّغة العربية
ملفّات
المعاجم
مساعدة
التتيجة
المعلومات التاريخية
البحث في المعنى
التطور الدلالي للكلمة
تأكيد
إظافة
تحيين
معالجة النص

عملية البحث

كلمة البحث: قطارُ

السياق: ○ كلمة ○ كلمتان ● ثلاث كلمات

معنى الكلمة حسب السياق الموجود فى النص

السياق: فلاح الصحراء وأوْغَلَ قطارٌ إبل فدعا برجل

المعانى حسب المعجم

1.مَا قطَرَ من الماءِ وغَيْرِه

2. قِطارِ الإبِلِ

المعنى حسب السياق: 2

النص الجاري

العنوان 0346 - مروج الذهب

من أخبار الحجاج مع أهل الشام

قال: وأقبل المستكفي يوماً على محمد بن محمد بن يحيى بن شيرزاد الكاتب، فقال له: أتعرف خبر الحجاج بن يوسف مع أهل الشام؟ قال: لا يا أمير المؤمنين، قال: ذكروا أن الحجاج بن يوسف كان قد اجْتَبَى قوماً من أهل العراق وجَدَ عندهم من الكِفاية ما لم يجد عند مختصيِّه من الشاميين؛ فشق ذلك على الشاميين وتكلموا فيه، فبلغ إليه كرمهم؛ فركب في جماعة من الفريقين، وأوْغَلَ بهم في الصحراء؛ فلاَحَ لهم من بُعْدٍ قطارٌ إبل؛ فدعا برجل من أهل الشام، فقال له: أمْض فاعرف ما هذه الأشباح، واسْتَقْص أمرها، فلم يلبث أن جاء وأخبره أنها إبل فقال: أمحملة هي؟ أم غير محملة؟ قال: لا أدري، ولكِنِي أعود وأتعرف ذلك، وقد كان الحجاج أتبعه برجل آخر من أهل العراق، وأمر بمثل ما كان أمَرَ الشاميَّ، فلما رجع العراقي أقبل عليه الحجاج وأهل الشام يسمعون، ما هي؟ قال: إبل، قال: وكم عددها؟ قال: ثلاثون، قال وما تحمل. قال: زيتاً، قال: ومن أين صَدَرَتْ! قال: من موضع كذا قال: وأين قصدت؟ قال: موضع كذا، قال: ومَنْ ربها؟ قال: فلان فالتفت إلى أهل الشام، فقال:

ألام على عمرو، ولو مات أو نأى ... لقَلَّ الذي يُغْنِي غَنَاءَكَ يا عمرو

Fig. 8. Word sense disambiguation interface.

Accordingly, we considered the word embeddings obtained without removing stop words from the corpus when trained words vectors. The second part is to evaluate the capacity of word embeddings to represent the sentence containing the ambiguous word and its sense definitions. Specifically, we consider three method: no weighting method, IDF weighting method and WMD distance method.

As illustrated in table 3, IDF weighting method achieves better results on WSD tasks. Unexpectedly, the cosine distance between average word vectors (No weighting method) is more powerful than WMD metric as it captures the meaning similarities between the context of use of the ambiguous word and its sense definitions.

6 Conclusion

In this paper, we presented our tool that aims to help linguists to create a historical dictionary of Arabic. The implemented method consists of two steps: in the first one we extract the meaning of an ambiguous Arabic word according to the historical period in which it appeared together with its first date of appearance, its users and its first place of appearance. Second, we generate the different meanings of a given word historically classified according to its appearance in the corpus. 100 ambiguous words have been chosen for the test. Each context of use has been extracted according to to three specific periods. Experiments have shown an accuracy of 48,89% for the Old Arabic, 48,57% for the Classical Arabic and 63,44% for the Modern Standard Arabic.

References

1. Abdel-Aziz, M.H.: A Historical Dictionary for Arabic Language: Documents and examples (2008)
2. Al-Said, A.B.: A Corpus-based Historical Arabic Dictionary: Linguistic and Computational processing. Ph.D. thesis, Cairo University (2011)
3. Al-Said, A.B.: The historical arabic dictionary resources. Journal Of the Arab languages (2015)
4. Al-Said, A.B., Medea-García, L.: The historical arabic dictionary corpus and its suitability for a grammaticalization approach. In: 5^{th} international conference in linguistics (2014)
5. Alian, M., Awajan, A., Al-Kouz, A.: Arabic word sense disambiguation using wikipedia. International Journal of Computing and Information Sciences (2016)
6. Alian, M., Awajan, A., Al-Kouz, A.: Arabic word sense disambiguation - survey. In: International Conference on New Trends in Computing Sciences (2017)
7. Alkhatlan, A., Kalita, J., Alhaddad, A.: Word sense disambiguation for arabic exploiting arabic wordnet and word embedding and word embedding. In: The 4th International Conference On Arabic Compitational Linguistics (ACLing) (2018)
8. Bouhriz, N., Benabbou, F., Lahmar, E.H.B.: Word sense disambiguation approach for arabic text. International Journal of Advanced Computer Science and Applications (2016)
9. Dahou, A., Xiong, S., Zhou, J., Haddoud, M.H., Duan, P.: Word embeddings and convolutional neural network for arabicsentiment classification. In: Proceedings of COLING (2016)

10. Frej, J., Chevallet, J.P., Schwab, D.: Enhancing translation language models with word embedding for information retrieval. The Computing Research Repository (CoRR) (2018)
11. Harris, Z.S.: Distributional structure. Word (1954)
12. Khalfallah, F., Aloulou, C., Belguith, L.H.: Had, a platform to create a historical dictionary. In: AICCSA (2016)
13. Khalfallah, F., Aloulou, C., Belguith, L.H.: A platform based anlp tools for the construction of an arabic historical dictionary. In: • (ed.) NLDB (2016)
14. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of the 32^{nd} International Conference on Machine Learning (2015)
15. Laatar, R., Aloulou, C., Hadrich-Belguith, L.: An xml model for an arabic historical dictionary. In: LPKM (2018)
16. Masmoudi, A., Bougares, F., Ellouze, M., Estève, Y., Belgui, L.: Automatic speech recognition system for tunisian dialect. Lang Resources and Evaluation (2017)
17. Menai, M.E.B.: Word sense disambiguation using evolutionary algorithms – application to arabic language. Computers in Human Behavior (2014)
18. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations (ICLR) (2013)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.A.: Distributed representations of words and phrases and their compositionality. In: The 26th International Conference on Neural Information Processing Systems (2013)
20. Nagoudi, E.M.B., Schwab, D.: Semantic similarity of arabic sentences with word embeddings. In: WANLP-EACL (2017)
21. Nagoudi, E.M.B., Schwab, D.: Semantic similarity of arabic sentences with word embeddings. In: WANLP@EACL (2017)
22. Navigli, R.: Word sense disambiguation : A survey. ACM Computing Surveys (2009)
23. Oele, D., van Noord, G.: Distributional lesk: Effective knowledge-based word sense disambiguation. In: International Conference on Computational Semantics (2017)
24. Pal, A.R., Saha, D.: Word sense disambiguation: A survey. International Journal of Control Theory and Computer Modeling (IJCTCM) (2015)
25. Pelevina, M., Arefiev, N., Biemann, C., Panchenko, A.: Making sense of word embeddings. In: Proceedings of the 1st Workshop on Representation Learning for NLP (2016)
26. Saad, M., Ashour, W.: Osac: Open source arabic corpora. In: International Conference on Electrical and Computer Systems (2010)
27. Soliman, A.B., Eissa, K., El-Beltagy, S.R.: Aravec: A set of arabic word embedding models for use in arabic nlp. In: 3rd International Conference on Arabic Computational Linguistics, ACLing (2017)
28. Ustalov, D., Teslenko, D., Panchenko, A., Chernoskutov, M., Biemann, C., Ponzetto, S.P.: An unsupervised word sense disambiguation system for under-resourced languages. In: In Proceedings of the 11^th Conference on Language Resources and Evaluation (2018)
29. Zahran, M.A., Magooda, A., Mahgoub, A.Y., Raafat, H.: Word representations in vector space and their applications for arabic. In: CICLing (2015)
30. Zouaghi, A., Merhbene, L., Zrigui, M.: Combination of information retrieval methods with lesk algorithm for arabic word sense disambiguation. Artificial Intelligence Review (2012)

31. Zouaghi, A., Merhbene, L., Zrigui, M.: A hybrid approach for arabic word sense disambiguation. International Journal of Computer Processing Of Languages (2012)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx