ISSN 2395-8618

Editorial

THIS ISSUE of the Polytechnic Open Library International Bulletin of Information Technology and Science (POLIBITS) includes ten papers by authors from eight different countries: Cuba, Ecuador, Mexico, Nepal, Peru, Switzerland, Ukraine, and USA. The majority of the papers included in this issue are devoted to the general topic of emerging challenges and trends in business intelligence, including such specific topics as software development, pattern recognition, natural language processing, forecasting, Internet of things, time series analysis, as well as optimization and multi-objective optimization.

A. Castro-Hernández et al. from **USA** in their paper "Collaboration and Content-Based Measures to Predict Task Cohesion in Global Software Development Teams" explore the use of collaboration and content-based measures to examine task cohesion within global software development teams, which a key component of team performance. They use a machine-learning classifier to derive content measures by categorizing the teams' message interactions as social, planning, or work. They show that content-based measures are more effective in predicting individual-level cohesion, and collaboration-based metrics are more effective at the group level.

Bijaya Kumar Hatuwal et al. from **Nepal** in their paper "Plant Leaf Disease Recognition Using Random Forest, KNN, SVM and CNN" classify and predict crop diseases by plant images with a number of machine learning models. For SVM, KNN and Random Forest algorithms, they extracted image features such as contrast, correlation, entropy, inverse difference moments using Haralick texture features algorithm. In contrast, CNN took directly the images as input. Of the classification algorithms they tested, CNN produced the best accuracy by a wide margin.

Olga Kolesnikova from **Mexico** in her paper "Complex System View on Natural Language" studies natural language as a complex adaptive system. She shows that the complex system view on natural language is a powerful tool not only for incorporating linguistic knowledge, but also for studying many open issues in phonetics, grammar, lexicon, language origin and evolution, and first language acquisition and development by simulating the structure and functioning of language using computational models.

Dmitriy A. Klyushin from **Ukraine** in his paper "Non-Parametric k-Sample Tests for Comparing Forecasting Models" provides an overview of the non-parametric tests used in business analysis for pairwise and group comparisons and describes a new, highly reliable, sensitive, and specific, non-parametric statistical test. His test is based on assessing the deviation of the observed relative frequency of an event from its a priori known probability. The prior probability is given by Hill's assumption, and the confidence intervals for the binomial success rate in the Bernoulli scheme are used to estimate its difference from the observed relative frequency. The paper presents the results of computer modeling and comparison of the proposed test with the alternative Kruskal-Wallis test and the Friedman test on artificial and real examples.

Rodolfo A. Pazos et al. from **Mexico** in their paper "Processing Natural Language Queries via a Natural Language Interface to Databases with Design Anomalies" describe an improvement to the processing performed by a domain-independent interface to treat databases with design anomalies and for the interface to be able to process correctly queries involving such anomalies. The problem is novel: existing literature on Natural language interfaces to databases (NLIDBs) has not even mentioned this problem, and much less addressed it. The importance of this problem, and of its solution, stems from the fact that existing NLIDBs would not work correctly for these databases with design anomalies, since these NLIDBs were designed under the assumption that they would be used with databases without anomalies.

Hermann M. Klusmann and **Renzo M. Carnero** from **Peru** and **USA** in their paper "Multicriteria Analysis for IoT Selection in a Telemetry System" propose a multi-criteria optimization study aimed to find the most suitable Internet of Things (IoT) technology for a telemetry network of water meters in a city. Basing on the types of IoT available in the given area, they balance accordingly to technical, social, and economic criteria and obtain the most appropriate IoT technology for the case study.

G. Rivera et al. from **Mexico** in their paper "Forecasting the Demand of Parts in an Assembly Plant Warehouse using Time-Series Models" describe the implementation of time-series models to forecast the demand for parts that could improve the relocation process. For this purpose, they implement different Holt-Winters Seasonal and SARIMA models. For the implementation of the SARIMA models, the follow the Box-Jenkins methodology. They use AIC and BIC metrics to identify the best Holt-Winters Seasonal model and the best SARIMA model. To check that the model is fit to the data, they test their models on the residual series. They use RMSE and MAPE metrics to evaluate the performance of Holt-Winters Seasonal and SARIMA models. They show that the SARIMA model outperforms the Holt-Winters Seasonal model.

ISSN 2395-8618

Eduardo Sánchez-Ansola et al. from **Cuba** in their paper "School Bus Routing Problem with Fuzzy Walking Distance" introduce a fuzzy model for the School Bus Routing Problem, particularly with the maximum student walking distance as a fuzzy element. This fuzzy version of the School Bus Routing Problem allows obtaining a set of solutions with different trade-offs between cost and relaxation of the original conditions. The authors analyze the results obtained in 31 instances by using the parametric approach taking into account three characteristics of the problem: number of bus stops, number of students, and walking distance. They show that the fuzzy version of the problem that they introduce is useful for decision-makers, since it provides relaxed alternative solutions with significant cost savings.

Lorena R. Rosas-Solórzano et al. from **Mexico** in their paper "Optimization of Many Objectives with Intervals Applying the MOEA/D Algorithm" propose I-MOEA/D, a new multi-criteria optimization method based on a MOEA/D approach, to deal with decision maker's uncertainty in costs and benefits of portfolios' projects. The novel features of their method include (a) handling large numbers of objectives; (b) a method to generate the initial population; and (c) handling the uncertainty of resources, costs, and benefits through intervals. The authors show competitiveness of their I-MOEA/D approach, which in their experiments with two to fifteen objectives improved the quality of solution of the state-of-the-art I-NSGA-II method in most instances.

Jhonny Pincay et al. from **Switzerland** and **Ecuador** in their work-in-progress paper "Towards a Computational Intelligence Framework to Smartify the Last-Mile Delivery" propose a framework for the improvement of the first-try delivery by studying traffic on the streets and past delivery success as a way of approximating customers' presence at home. In contrast to existing solutions, they work only with data that do not compromise the customers' privacy and get insights about traffic features in cities without the need of deploying expensive equipment to obtain data. Their main goal is to provide a route plan to the delivery team and route planners, which allows finishing the distribution of the parcels in the minimum amount of time, while being able to deliver effectively the highest amount of parcels. This will be translated into lower resource consumption and increased customer satisfaction. Their work follows the principles of design science for information systems.

This issue of the journal will be useful to researchers, students, and practitioners working in the corresponding areas, as well as to public in general interested in advances in computer science, artificial intelligence, and computer engineering.

Laura Cruz-Reyes,
Tecnológico Nacional de México /
Instituto Tecnológico de Cd. Madero (TecNM / ITCM),

Felix Castro-Espinoza,
Universidad Autónoma del Estado de Hidalgo (UAEH),

Gilberto Rivera,
Universidad Autónoma de Ciudad Juárez (UACJ),

Guest Editors

ISSN 2395-8618

Collaboration and Content-Based Measures to Predict Task Cohesion in Global Software Development Teams

A. Castro-Hernández, V. Pérez-Rosas, and K. Swigger

***Abstract*—Task cohesion is a key component of team performance. This paper explored the use of collaboration and content-based measures to examine task cohesion within global software development teams. The study aimed to predict the perception of task cohesion among teams involving students from two different countries. The study applied collaboration from previous work and also proposed new metrics such as Reply Similarity and Reply Rate. In addition, a machine learning classifier is used to derive content measures by categorizing teams' message interactions as social, planning, or work. Correlation analyses are conducted to examine whether collaboration and metrics are predictive of task cohesion. The analyses are conducted at the individual and group levels and used the culture factor as a control variable since cohesion has been found previously affected by location. The research findings suggest that content-based measures were more effective in predicting individual-level cohesion while collaboration-based metrics were more effective at the group-level.**

***Index Terms*—Component, formatting, style, styling, insert.**

1. Introduction

WORKING with global teams presents important challenges associated with managing time, distance, and communication technologies. Most of the information is, in many cases, stored in databases. However, for a user to obtain information from a database (DB), he must have knowledge of a query language for databases (such as SQL). [1].

Particularly, in educational settings, global teams provide students with many valuable experiences but also pose several challenges such as having to deal with people from diverse cultures, coping with different perceptions of time and relationships, and finding effective communication tools that allow distributed groups to work together.

Needless to say, any one of these challenges can have a significant effect on team performance. Among several team processes affected by computer-mediated communication (such as cohesiveness, status, and authority relations), cohesion has remained a critical issue for all types of work teams. In global teams, positive cohesion levels have been directly linked to group performance [1]. However, when compared with co-located teams, virtual teams tend to be less cohesive [2]. Thus, making it important to develop effective methods to measure cohesion that allow opportune team interventions.

While researchers have proposed several ways to measure task cohesion levels within groups, very few have been tested on or designed for virtual learning teams.

In this paper, we examine several existing cohesion metrics that characterize different degrees of similarity among group members (e.g., word category usage, reply behavior, etc.). We also proposed new measures that capture interaction aspects such as word and reply rates. We explored the question of whether quantitative group measures are better at predicting cohesion than those that are associated with the individual. Thus, our main objective was to determine whether similarity measures are better at predicting cohesion levels than quantity-based measures and whether individual measures are more related to cohesion perception than group measures.

2. Related Work on Cohesion Measures

Cohesion is usually defined as "a dynamic process which is reflected in the tendency for a group to stick together and remain united in the pursuit of its goals and objectives" [3]. It has been studied at both the individual and group levels [4] and has been linked to group performance [5]. Group cohesiveness in any type of team seems to increase over time, particularly when there is a leader in the group [6]. Other elements that appear to affect a group's cohesiveness include team size, degree of democratic behavior within a group, participation, and satisfaction [7].

The Group Environment Questionnaire (GEQ) [8] is a survey instrument commonly used to measure individual perceptions of group cohesion. GEQ consists of 18 items that measure group and individual factors, i.e., group integration and individual attraction to the group. These factors are further divided into tasks and social dimensions, which describe general motivation toward achieving group objectives and developing social relationships.

Manuscript received on October 21, 2020, accepted for publication on November 25, 2020, published on December 30, 2020.

A. Castro-Hernández is with the Department of Computer Science and Engineering, Wayne State University, Detroit, USA (e-mail: hg3246@wayne.edu).

V. Pérez-Rosas is with the Department of Computer Science and Engineering, University of Michigan, Ann Arbor, USA (e-mail: vrncapr@umich.edu).

K. Swigger is with the Department of Computer Science and Engineering, University of North Texas, Denton, USA (e-mail: kathy@cse.unt.edu).

ISSN 2395-8618

In addition to surveys, researchers have also developed more objective measures for characterizing a group's cohesiveness. These measures can be helpful in the analysis of team cohesion in remote work, which has particularly increased due to the impact of COVID-19 on our society [9]. For instance, [10] described group cohesion in terms of a density measure expressed as a ratio of the number of connections among members over all possible links. However, this approach did not take into account the weight or intensity of those connections. Thus, [11] calculated group cohesion (density) by looking at only the links that have weights higher or equal to a pre-defined number, which may vary according to different contexts.

Another type of cohesion measure, called Linguistic Style Matching (LSM), was developed by [12]. This particular measure is based on the similarity of the use of function words between two individuals. Once all paired similarities among group members are computed, the paired values in the group are then averaged, and this number becomes the group's cohesiveness score. Using this technique, the researchers found a correlation between LSM and a cohesion construct (obtained through a survey) and a limited relation between LSM and performance. This particular study tested the LSM measure using chat communications generated during a one-hour session from single-gender teams. However, researchers who applied LSM to the analysis of email messages among team members over an extended period of time were unable to duplicate the significant relationships between cohesion and performance [13]. Moreover, the use of function words in a non-native speaker group setting might also affect LSM's predictive capabilities.

In another study, [14] proposed a measure called *Individual cohesion* that is designed to predict group cohesion. The *Individual cohesion* measure is calculated by summing messages between each pair of individuals on a team and then averaging those counts. Members' individual cohesion scores are then correlated with task cohesion. Although the study found a significant relationship between *Individual cohesion* and a group's cohesion level, it was noted that this similarity approach might be affected by individuals who perform poorly but have similar interaction scores. Thus, the authors suggested that a measure based on communication intensity might be a better predictor of cohesion in a virtual setting because of the low interaction rates often found among group members in this type of setting.

The above research served as an important tool for defining the major factors that were deemed important for this study. These factors include both similarity and quantitative measures. The general question asked was which of these factors tends to be a better predictor of cohesion levels among global software learners. The measures mentioned above represent some of the factors that were included in the experiments for this study.

TABLE I
ACTIVITIES PER PROJECT AND INSTITUTION.

	US institution	MX institution
Project 1	Museum website	Database functionality
Project 2	Museum website redesign	Database functionality
Project 3	Learning website	Learning website functionality

3. DATASET

Our data is drawn from a global software development study involving students from three higher education institutions located in the US and Mexico (MX). Data collection occurred from 2014 to 2015 and included software development projects from three different undergraduate-level courses. Participants consisted of 116 males and 62 females who were 22-years old on average.

All participant's communications happened in English and used an online collaborative tool called Redmine [15]. Teams were formed by pairing students from the American institution and either of the Mexican institutions. Participants were randomly assigned to their respective groups. The teams worked on three different projects involving either the design of a museum or a learning website as well as implementing their functionality. Activities were distributed as shown in Table III.

We created a project management web application in Redmine that allowed us to record communications occurring among the team's members. These include participation in chats, forums, and wikis, as well as file sharing. We also recorded the date and time in which each activity occurred as well as the author of each online activity.

A. Collaboration Measures

We explored two main team factors in relation to cohesion: communication similarity and communication processes. We obtained assessments of *Task Cohesion* using individual questionnaires completed by most of the team members. The questions in the survey are derived from the multidimensional cohesion model developed by [8] and also from work by [16].

B. Communication similarity metrics

We derive communication similarity metrics from the graph representation of communication replies among team members. Figure 1 shows a sample graph derived from the interaction shown in Table II. In this figure, the conversation is shown as a directed graph, where vertices represent the different team members, and edges represent the number of replies received by participants. For example, u_2 replied to u_1's initial message, while u_1 replied to u_2's message about studying engineering. The reply counts are extracted from the forum and chat exchanges and reflect when a message is delivered to a specific participant (i.e., the last participant in the communication).

ISSN 2395-8618

TABLE II
EXAMPLE OF COMMUNICATION AMONG THREE TEAM MEMBERS.

User	Message
$u1$	Hello. My name is Bryan and I am a student at UNT...
$u2$	Hi, I am Carlos.
$u2$	I live in Panama.
$u2$	I study Computer Engineering.
$u1$	Nice two meet you.
$u1$	I just read the project description and it seems that...
$u3$	Hello I live in Texas, and I study Computer Science.

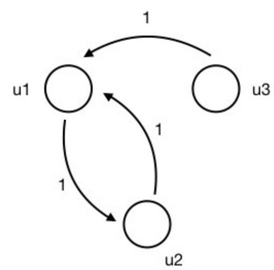

Fig. 1. Reply graph of conversation in Table II

We generate reply graphs for each team and project and then derive three similarity metrics as described below. Density is a group measure that describes the degree of interconnectedness among group members, with high density indicating a high degree of interconnectedness and low density indicating a low degree of interconnectedness [10]. The original density metric is calculated at the group level; however, we redefine it at the individual level. Where the *individual density* of a team participant (D_i) is the count of people with whom the participant interacted with (*actual_links$_i$*) over the number of people in the rest of the team (*possible links*).

$$D_i = \frac{actual\_links_i}{possible\_links} \quad (1)$$

Reply similarity We also measure the similarity on reply behaviors at the individual and group levels. These measures are derived from the cohesion metric proposed by [17] and are based on the idea that people view their group's cohesiveness as being a combination of their own participation as well as from others. We thus calculate the reply rate at the team level (*reply similarity*, where $reply_{ij}$ is the number of replies from person i to person j, and $reply_{ji}$ is the number of replies from person j to person i.

$$replysimilarity_i = 1 - \frac{|reply_{ij} - reply_{ji}|}{reply_{ij} + reply_{ji}} \quad (2)$$

The reply similarity metric produces a score between 0 and 1, indicating the degree of reply similarity among team members. The *individual reply similarity* indicates whether participation in group discussions increased the individual's perception of group cohesion and was simply calculated as the total number of replies that an individual sent.

Linguistic Style Matching (LSM) is a metric used to determine whether individuals match their language use. Its use as a cohesion measure was first proposed by [12]. We calculated LSM using the formula shown in Equation 3, where x is a function word from the LIWC lexicon (e.g., auxiliary verbs, articles, common adverbs, personal pronouns) and x_i denotes their use frequency by the i team member (or j). Using this formula, we determined the similar usage of function words between member i and each of the other members in the team.

$$LSMx_{ij} = 1 - \frac{|x_i - x_j|}{x_i + x_j} \quad (3)$$

Information Exchange We derived two metrics for information exchange among team members: similarity and rate. The first (*Information Exchange Similarity*) is calculated by simply counting the number of words typed by each participant, with the assumption that these words were being transmitted to every other team member in the group (i.e., each word in each message is perceived as some type of participatory exchange). This calculation was computed as follows:

$$informationexchange_i = 1 - \frac{|wc_i - wc_j|}{wc_i + wc_j} \quad (4)$$

where wc_i and wc_j are the numbers of words used by user i and j, respectively.

The second metric, called Information Exchange Rate, is calculated as the total word count by a team participant.

C. Content Measures

Previous research showed that words associated with word categories such as contribution, seeking input, reflection, social, and planning are highly related to cohesion perception [18]. For example, the use of words related to social content seems to influence the level of trust among team members as well as create a more pleasant environment within a team. Although these word categories may not have a direct effect on *Task Cohesion*, it is natural to assume that they may have some indirect effect on *Task Cohesion*; for example, the absence of social behavior may result in a decrease in *Task Cohesion* among group members.

We thus derived a set of content measures that capture the use of these categories.

Social: number of messages using language related to social interactions.

Planning number of messages containing language related to organizing the project development, i.e., verbs, nouns, or dates. Work messages containing information about the project organization and management. Borrowing from the original research [19], the work category included words related to *Contribution* and *Seeking input*.

4. RESULTS

A total of 5,583 messages were transmitted during the three projects. A total of 167, out of a possible 180, *Task Cohesion*

ISSN 2395-8618

surveys were collected. Since we had 23 missing surveys, we kept only messages by students who had completed the questionnaire; thus, our dataset included a total of 5446 messages.

A. The Culture Effect

A previous study on *Task cohesion* found that location (or country of birth) affects *Task cohesion* perception [14]; and since our experiments included participants studying in different locations (i.e., US or Mexico) but born in different countries, we anticipated that culture might have an effect on individuals' perceived cohesion.

TABLE III
TASK COHESION VALUES BY CULTURE. *$p < 0.05$.

Country	n	mean	India	US
India	55	8.01		
US	20	6.21	1.807*	
Mexico	78	6.48	1.5310*	0.2746

TABLE IV
PARTIAL CORRELATIONS OF SIMILARITY MEASURES AND DENSITY TO TASK COHESION, CONTROLLED BY CULTURE. †$p < 0.10$., *$p < 0.05$.

Measure	Task cohesion
Individual Density	0.037
Reply similarity	0.060
Linguistic Style Matching	0.114†

Our data contained survey responses from people who were born in eleven different countries. However, most of the surveys were completed from students born in India, the US, and Mexico (i.e., $n > 20$), while the rest of the countries were represented by only a few surveys (i.e., $n < 4$). As a result, we reduced our dataset even further and used data from only students born in any of the three main countries represented in our study; thus, ending up with a final count of 4,849 messages sent by 153 participants. We then conducted a preliminary analysis to evaluate whether *Task Cohesion* assessments differed among respondents. We compared the Task cohesion mean values between countries and found that students from India tended to have higher *Task Cohesion* perceptions than either US or Mexican students (see Table III). Thus, we decided to use the *Culture* factor as a control variable in our analyses.

B. Similarity Metrics in Task Cohesion

In order to determine the relationship between our similarity measures and *Task Cohesion*, we computed a partial correlation between each measure (i.e., *Individual Density*, *Reply Similarity*, *Linguistic Style Matching*, and *Task Cohesion*), controlling for the *Culture* factor. The results of these correlations are reported in Table IV.

Results suggest that the *Density* measure at the individual level is unrelated to *Task Cohesion* (r=0.037, p=0.325). Since weights assigned to different edges in the reply graphs tended to vary widely (ranging from 1 to 70), we suspect that the large variance among participants' exchanges may have affected the correlation between *Density* and *Task Cohesion*. Similarly, we found no correlation between *Reply Similarity* and *Task Cohesion* (r=0.060, p=0.230). The lack of a correlation between these two variables may be explained by the inactivity of one or two group members. For example, we noted that in cases where at least one of the group members was not participating in group discussions, the *Reply Similarity* score was low.

TABLE V
PARTIAL CORRELATIONS OF QUANTITY-BASED MEASURES TO TASK COHESION, CONTROLLED BY CULTURE. †$p < 0.10$., *$p < 0.05$.

Measure	Similarity	Rate
Reply	0.060	0.108†
Information exchange	0.152*	0.175*

On the other hand, *Linguistic Style Matching* has a positive, although low, significant correlation with *Task Cohesion* (r=0.114, p=0.081). An analysis of the use of function words by culture for the first week of each project shows students who participated in the Spring 2014 project had a significant difference in their use of personal pronouns (p=0.027) and quantitative words (p=0.054); also, students who participated in the Spring 2015 project show a significant difference in the use of impersonal pronouns (p=0.055); however, all 2014 participants showed no difference in function-word usage by country. It is important to note that the Fall 2014 project consisted largely of students who were born in either Mexico or the US, while the other two projects consisted mainly of students born in either Mexico or India. Thus, it appears that US and Mexican students have more similar linguistic patterns than Mexican and Indian students.

Finally, *Information Exchange Similarity* has a positive and significant correlation with *Task Cohesion* (r=0.152, p=0.031). Although *Information Exchange Similarity* and *Reply Similarity* are somewhat related, in terms of what they are measuring, the simple word-based metric of *Information Exchange Similarity* shows a better correlation with cohesion perception, possibly because words rather than replies (which tend to be sentences) produce more data. The amount of data that is used to analyze the relationship among variables may affect the degree of significance. Word counts may also be a better metric of different levels of interaction since individuals who tend to be more engaged in the project will probably communicate more, which in turn may affect the perception of the group's cohesiveness.

C. The Effect of Quantity-Based Measures in Task Cohesion

We next tried to determine if the intensity of the team's communications, as measured by *Reply Rate* and *Information Exchange Rate*, has an effect on Task cohesion. Again, we

ISSN 2395-8618

conducted a partial correlation analysis controlled by the *culture* factor between these two variables and Task Cohesion. Results are shown in Table V. The first column shows the correlation scores for both *Reply Similarity* and *Information Exchange Similarity*, while the second column shows the correlation scores for *Reply Rate* and *Information Exchange Rate*.

Our results indicate that *Reply Rate* has a positive, although low, significant correlation with *Task Cohesion* (r=0.108, p=0.093). *Information Exchange Rate* also has a significant positive correlation with *Task Cohesion* (r=0.175, p=0.016). In comparison with the regular similarity measures, the more quantitative-based measures have higher correlation values.

TABLE VI
PARTIAL CORRELATIONS OF GROUP MEASURES TO TASK COHESION, CONTROLLED BY CULTURE. †P < 0.10.

Measure	Task cohesion
Group reply rate	0.113†
Group information exchange rate	0.112†

This seems to suggest that quantity-based measures that capture amounts, and perhaps engagement, may be better predictors of group cohesion than measures that try to assess different similarity constructs in a team's exchanges.

D. Effect of Group Measures

We then took the two variables that appeared to be significantly related to task cohesion and calculated group-level scores for each of these two measures. Thus, we created a *Group Reply Rate* and *Group Information Exchange Rate* variable and examined the relationship between these two factors and *Task Cohesion*. Table VI shows the correlations for each of these two variables.

Group Reply Rate and *Group Information Exchange Rate* show a positive but low, significant correlation with *Task cohesion*, i.e., r=0.113, p=0.083 and r=0.112, p=0.085, respectively. These results suggest that an individuals' and a group's engagement in communication is a predictor of task cohesion perception.

E. Content-based Features for Task Cohesion

To enable these analyses, we devised a data-driven approach to extract content features from student's messages. We thus aimed to categorize participant messages into the social, planning, and work categories. Note that we used this approach as custom lexicons for these categories are not readily available. We thus developed an automated text analysis program that could classify students' messages into three different categories. To conduct our experiments, we used a dataset consisting of 1,866 messages that had been manually annotated in previous work [19], [20] with the planning, contributing, seeking input, reflections, monitoring, and social categories.

Our study only used four of these categories (i.e., social, planning and work) with the work category including the contribution and seeking input categories. Thus, messages that contained the *agreement* label were removed so that these types of short messages were not included in any of our counts for the proposed categories. Agreement messages are those that confirm or deny some previous message such as "ok," "sure," "good," etc. This resulted in a dataset consisting of 305 messages under the social category, as well as 166 and 1279 messages in the planning and working categories, respectively.

The features used to help seed the classification process were those found in research related to the LIWC software tool. Using this tool, we investigated 73 features. We also computed the unigrams of each message, obtaining a dictionary of more than 3000 entries. We tested the ability of both feature sets (i.e., Unigrams & LIWC) to predict the target label. Similarly, we tested the use of LIWC features only during the classification task.

We compared the performance of three classifiers: Support Vector Machines, Random Forest, and Naive Bayes. The performance metric used for comparison was the F1-score, which is the harmonic mean of precision and recall. Also, we conducted 10-fold cross-validation during these experiments.

TABLE VII
F-SCORE VALUES FOR CLASSIFICATION ALGORITHMS.

Feature set	Support Vector Machines	Random Forest	Naive Bayes
LIWC+Unigrams	0.821	0.777	0.522
LIWC	0.752	0.776	0.495

As seen in Table VII, all classifiers obtained better results by using both the LIWC and Unigrams features as compared to using only LIWC features. In addition, the best performances were obtained by the Support Vector Machines classifier (F1score=0.821). Hence, we used the Support Vector Machine classifier to label messages in our dataset. The result of applying the classifier to the messages in the experimental dataset can be found in Table VIII.

TABLE VIII
DISTRIBUTION OF MESSAGE CONTENT CATEGORIES IN THIS STUDY

Class	Instances
Social	936
Planning	4528
Work	392

Some of the features that provided more information for classifying each message into a category are shown in Table IX. The Work category was found to be related to unigrams that are sometimes associated with performing activities such as give, talk, data, instructor, right. However, the

ISSN 2395-8618

Work label was also found to be related to three LIWC categories: 1) function words *LIWC.FUNCTION* (e.g., prepositions, articles, and common adverbs), which are often related to the idea of formal thinking [21]; 2) work *LIWC.WORK* (e.g., accomplish, work, and success), which are often related to the idea of well-performing teams [14]; 3) Interrogatives *LIWC.INTERROG*, (e.g., how, when, what, which are often used to represent the students' seeking input process.

TABLE IX
RELEVANT WORDS BY CATEGORY, ACCORDING TO CLASSIFIER.

Work	Social	Planning
Right	LIWC.INFORMAL	meeting
LIWC.FUNCTION	fun	night
LIWC.WORK	Lets	tonight
LIWC.INTERROG	Thanks	Thursday
data	Oh	whenever
talk	LIWC.FOCUSPRESENT	must
instructor	later	yet
could	LIWC.SOCIAL	schema
modification	Nice	yours
give	hello	Monday

Similarly, words related to the Social category included words such as fun, lets, thanks, hello, later, nice - all of which seem to suggest characteristics related to social interaction. The Social category also included more concrete relations with specific LIWC categories: 1) informal *LIWC.INFORMAL*, a category that consists of words such as netspeak (lol, btw, thx), swear words (fuck, damn, shit), non-fluencies (err, mmm); 2) *Present tense* (*LIWC.FOCUSPRESENT*), a category that includes words such as today, is, now, which are words related to more personal information sharing; 3) social (*LIWC.SOCIAL*), a category that consists of words such as mate and they. Again, LIWC provided the research with relevant key categories related to specific conversation types.

On the other hand, the main word features that comprised the Planning label did not include any specific LIWC category. Instead, the classifier produced unigrams related to project management activities such as night, whenever, days of the week (Thursday, Friday, Monday). We also saw patterns that included words related to specific management activities such as meeting and schema (which was probably because of the type of projects that were assigned, e.g., database schema). The Planning label also included words related to future events, which are often used in planning tasks. Given this particular list of words, it should be possible, at some time in the future, to create a LIWC category that would automatically identify these types of communication.

These results show the importance of the LIWC tool and the use of unigrams to obtain the appropriate label for the message's exchanges by team members.

After placing the messages into their various content categories, we computed the correlations between *Social Similarity*, *Planning Similarity*, and *Work Similarity* and *Task Cohesion*. In addition, we computed correlations between *Social Rate*, *Planning Rate*, and *Work Rate* and the same target variable.

TABLE X
CORRELATIONS OF CONTENT-BASED VARIABLES WITH TASK COHESION. †P < 0.1, *P < 0.05

Content type	Similarity	Rate
Social	0.052	0.136*
Planning	0.001	0.060
Work	0.101†	0.063

Table X shows the correlations of these variables when controlled by team size and culture. The *Work Similarity* variable shows a nearly statistically significant correlation with *Task Cohesion* (r=0.101). However, neither *Social Similarity* or *Planning Similarity* are correlated with the cohesion construct.

On the other hand, there exists a statistically significant correlation (r=0.136) between the *Social Rate* variable and *Task Cohesion*. But, again, *Work Rate* and *Planning Rate* do not show a statistically significant correlation.

TABLE XI
AVERAGE AND SUM OF SIMILARITY AND RATE CONTENT MEASURES

Content type	Similarity (average)	Rate (sum)
Social	0.216	559
Planning	0.081	203
Work	0.408	2046

The statistically significant *Work Similarity* correlation can be explained by looking at previous results that show the perception of Group Cohesiveness is often affected by a group's perception that all members are doing their fair share of the work. So, if evidence shows that all members are participating in the communication, then an individual's perception of the group's cohesiveness should be higher. Moreover, the similarity between members' conversations about work (i.e., *Work Similarity*) seems to be more important than the rate at which these exchanges occur.

Interestingly, *Social Rate* shows a stronger correlation with *Task Cohesion* than *Work Rate*. This result may demonstrate the importance of having some social communications among group members within a virtual environment. Despite the possibility that participants' social interactions may sometimes affect the accurate assessment of *Task Cohesion*, as discussed in [7], we believe that this did not occur in the context studied in this research, as evidenced by a large number of work-related communications as compared to social messages that were transmitted among group members, as shown in Table XI.

ISSN 2395-8618

In contrast, there were no correlations between *Planning Similarity* and *Task Cohesion*, nor between *Planning Rate* and *Task Cohesion*. This lack of statistical significance between *Task Cohesion* and the planning construct may be due to the relatively small number of messages that were generated in this category.

We also evaluated the performance of these same metrics at the group-level, i.e., we calculated *Group Social Similarity*, *Group Work Similarity*, *Group Planning Similarity*, *Group Social Rate*, *Group Work Rate*, *Group Planning Rate*. We first removed teams that had one or more participants who were not identified with one of the major countries, i.e., the United States, India, and Mexico. Also, we computed the *Group Task cohesion* measure by averaging the *Task Cohesion* perception of the team members. When a survey was not completed by a student, *Group Task Cohesion* was estimated using a missing data technique based on systematic non-response [22]. A total of 18 teams (out of 35) were used in the group-level analysis.

TABLE XII
CORRELATIONS OF CONTENT-BASED VARIABLES AT THE GROUP-LEVEL WITH TASK COHESION.

Content type	Similarity	Rate
Social	–0.235	0.088
Planning	–0.124	0.192
Work	–0.062	–0.038

Results shown in Table XII indicate that none of the group variables had a statistically significant correlation with *Task Cohesion*. This result may have occurred because of the small sample size for the comparison. Thus, more data might be required to determine whether the individual-level metrics, which we found to be related to *Task Cohesion*, i.e., whether (*Work Similarity* and *Social Rate*), have similar effects when aggregated at the group level.

5 CONCLUSIONS

This research examined a number of different measures to determine whether certain types of factors were better predictors of *Task Cohesion* than others. Similarity measures, drawn from previous research, and new quantitative measures created for this study were examined. The measures were applied to data generated from three global software learning projects that took place between students located in Mexico and the US.

Individual similarity measures intended to predict *Task Cohesion* had mixed results. For example, *Individual Density* was not correlated with *Task Cohesion*. At the same time, *Reply Similarity* had no significant relation to cohesion. The lack of significant results for these two variables can be explained by looking at the number of exchanges between members in the co-located teams. Students seemed to have fewer exchanges between team members in their own country as opposed to team members in the remote country. Possibly, co-located team members may have had offline conversations that were not measured. It may also be the case that the measure itself needs to be redefined since scores for both *Density* and *Reply similarity* are highly affected by inactive members. On the other hand, *Linguistic Style Matching* had a small but significant relationship to group cohesiveness. We speculate that the significance was low because of the large difference between the use of function words by Indian students as compared to Mexican students. *Information Exchange Similarity*, which is a word-based calculation, also achieved a significant correlation with *Task Cohesion*. This seems to suggest that a simple measure of the number of words exchanged provides a better representation of cohesion than either messages or replies, perhaps because such a measure generates more data for the analysis.

Since communication within a virtual learning team tends to vary, metrics based on interaction intensity were also proposed, i.e., *Reply Rate* and *Information Exchange Rate*.

Both seemed to predict *Task Cohesion* much better than the similarity version of these two variables. It should be noted that the word-based factor of *Information Exchange Rate* was a better predictor of *Task Cohesion* than *Reply Rate*. Although we computed group versions of *Reply Rate* and *Information Exchange Rate*, we found only low significant relationships between either one of the variables and *Task Cohesion*.

These results seem to indicate that previously cited similarity measures used to predict task cohesion may be limited to analyzing groups that are highly interactive and that work on short-term tasks. Data from this study suggests that similarity measures may be affected by less-active members and the presence of cross-cultural teams, both of which can impact message length and word usage. Since global software teams have both of these characteristics, similarity measures may have limited value in computing cohesiveness for distributed learning projects. However, this study also found that several quantitative measures that captured both reply and word rates were useful predictors of a group's cohesiveness within a global software development learning team.

Furthermore, measures based on the content within the communications were also developed. First, we generated a message classifier that performed accurately. Using output from the classifier, we were able to analyze their relationship with *Task Cohesion*. These measures showed mixed results. Only *Work Similarity* and *Social Rate* were found correlated to *Task Cohesion*. The statistically significant relation between *Work Similarity* and *Task Cohesion* can be explained by looking at previous results that show that the perception of cohesiveness is often affected by a group's perception that all members are doing their fair share of the work. So, if evidence shows that all members are participating in the communication, then an individual's perception of the group's cohesiveness should also be higher. On the other hand, the strong

ISSN 2395-8618

correlation between *Social Rate* and *Task Cohesion* shows that it is important to have at least some social communications during a software development project since these types of interactions can lead to an increase in trust within the group, which may result in an increase in group cohesiveness [13].

Further research is needed to determine whether a temporal approach (which may help to identify when good-performance teams start working) can produce better predictions of the perceived cohesiveness within a distributed team. Until that research is completed, we believe that similarity and, to a larger extent, content measures can be used to predict group cohesiveness within a global software student project.

REFERENCES

[1] P. Goodman, E. Ravlin, and M. Schminke, "Understanding Groups in Organizations," *Tepper School of Business*, Jan. 1987.

[2] K. Furumo and J. Pearson, "An Empirical Investigation of How Trust, Cohesion, and Performance Vary in Virtual and Face-to-Face Teams," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006. HICSS '06*, vol. 1, Jan. 2006, pp. 26c–26c.

[3] A. V. Carron, "Cohesiveness in sport groups: Interpretations and considerations." *Journal of Sport psychology*, 1982.

[4] S. M. Gully, D. J. Devine, and D. J. Whitney, "A meta-analysis of cohesion and performance effects of level of analysis and task interdependence," *Small Group Research*, vol. 26, no. 4, pp. 497–520, 1995.

[5] S. S. Webber and L. M. Donahue, "Impact of highly and less job-related diversity on work group cohesion and performance: A meta-analysis," *Journal of management*, vol. 27, no. 2, pp. 141–162, 2001.

[6] V. L. Schwanda, K. Barron, J. Lien, G. Schroeder, A. Vernon, and J. T. Hancock, "Temporal Patterns of Cohesiveness in Virtual Groups," in *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, ser. CSCW '11. New York, NY, USA: ACM, 2011, pp. 709–712.

[7] E. Rovio, J. Eskola, S. A. Kozub, J. L. Duda, and T. Lintunen, "Can High Group Cohesion Be Harmful? A Case Study of a Junior Ice-Hockey Team," *Small Group Research*, vol. 40, no. 4, pp. 421–435, Aug. 2009.

[8] A. V. Carron, W. Widmeyer, and L. R. Brawley, "The development of an instrument to assess cohesion in sport teams: The Group Environment Questionnaire." *Journal of sport psychology*, 1985.

[9] A. Ozimek, "The future of remote work," *Available at SSRN 3638597*, 2020.

[10] S. Wasserman, *Social network analysis: Methods and applications*. Cambridge University Press, 1994, vol. 8.

[11] C. Reffay and T. Chanier, "How Social Network Analysis can help to Measure Cohesion in Collaborative Distance-Learning," in *Designing for Change in Networked Learning Environments*, ser. Computer-Supported Collaborative Learning, B. Wasson, S. Ludvigsen, and U. Hoppe, Eds. Springer Netherlands, Jan. 2003, no. 2, pp. 343–352.

[12] A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker, "Language Style Matching as a Predictor of Social Dynamics in Small Groups," *Communication Research*, vol. 37, no. 1, pp. 3–19, Feb. 2010.

[13] S. A. Munson, K. Kervin, and L. P. Robert Jr, "Monitoring Email to Indicate Project Team Performance and Mutual Attraction," in *17th ACM conference on Computer supported cooperative work & social computing*. New York, NY, USA: ACM, 2014, pp. 542–549.

[14] A. Castro-Hernandez, K. Swigger, and M. P. Ponce-Flores, "Effects of´ cohesion-based feedback on the collaborations in global software development teams," in *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2014 International Conference on*. IEEE, 2014, pp. 74–83.

[15] Redmine. (2021) Redmine overview. [Online]. Available: www.redmine.org

[16] S. A. Carless and C. D. Paola, "The Measurement of Cohesion in Work Teams," *Small Group Research*, vol. 31, no. 1, pp. 71–88, Feb. 2000.

[17] A. Castro-Hernandez, K. Swigger, M. P. Ponce-Flores, and J. D.´ Tern-Villanueva, "Measures for Predicting Task Cohesion in a Global Collaborative Learning Environment," in *Global Software Engineering Workshops (ICGSEW), 2016 IEEE 11th International Conference on*. IEEE, 2016, pp. 31–36.

[18] K. Swigger, M. Hoyt, F. C. Sere, V. Lopez, and F. N. Alpaslan, "The temporal communication behaviors of global software development student teams," *Computers in Human Behavior*, vol. 28, no. 2, pp. 384–392, Mar. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0747563211002160

[19] D. D. Curtis and M. J. Lawson, "Exploring collaborative online learning," *Journal of Asynchronous learning networks*, vol. 5, no. 1, pp. 21–34, 2001.

[20] M. R. Hoyt, "Automatic Tagging of Communication Data," Ph.D. dissertation, University of North Texas, Denton Texas, Aug. 2012. [Online]. Available: http://digital.library.unt.edu/ark:/67531/metadc149611/

[21] J. Pennebaker, *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury USA, 2013.

[22] D. A. Newman and H.-P. Sin, "How Do Missing Data Bias Estimates of Within-Group Agreement? Sensitivity of SD WG, CVWG, rWG(J), rWG(J) * , and ICC to Systematic Nonresponse," *Organizational Research Methods*, vol. 12, no. 1, pp. 113–147, Jan. 2009.

ISSN 2395-8618

Plant Leaf Disease Recognition Using Random Forest, KNN, SVM and CNN

Bijaya Kumar Hatuwal, Aman Shakya, and Basanta Joshi

***Abstract*—Huge loss in crop production occurs every year due to late identification of pant diseases in developing countries like Nepal. Timely and correct identification of such diseases with less dependency in related field expert can be more effective solution to the problem. Plants suffer from various diseases and correctly identifying them by observing the leaves is major challenge especially if they have similar texture. Consideration of plant leaf color and various texture features is extremely important to correctly predict the defect in plant. The aim of this work is to classify and predict given disease for plant images using different machine learning models like Support Vector Machine(SVM), k-Nearest Neighbors (KNN), Random forest Classifier (RFC), Convolutional Neural Network and compare the results. Image features like contrast, correlation, entropy, inverse difference moments are extracted using Haralick texture features algorithm which are fed to SVM, KNN and Random Forest Algorithms whereas CNN directly feeds upon images as input. Among the used models CNN produced highest level of accuracy of 97.89% and RFC, SVM and KNN had accuracy of 87.43%, 78.61% and 76.96% respectively for sixteen different image categories used.**

***Index Terms*—Plant disease, Haralick texture, support vector machine, k-nearest neighbor, random forest classifier, convolutional neural network.**

I. Introduction

FOR BETTER agricultural productivity health of plant is a primary concern. Diseases hinder the normal state of a plant and potentially modify or interrupt its vital mechanisms such as fertilization, photosynthesis, transpiration, germination etc. Time and again plants get various diseases depending upon the factors like environment, season, soil, bacteria and others. The conventional process of plants disease detection with bare eyes observation method is tiresome and is non-effective. So identifying the plant disease for farmers requires help of related filed specialist most of the time. Hiring the related filed experts may cost farmers heavily and use of pesticides without knowledge will degrade the quality land and harm the living organisms. So utilizing available technology to identify the plant diseases may be a viable solution.

Plant disease recognition by visual method is cumbersome task, less accurate and can be applied only in limited areas. Some general diseases in plants are early and late scorch, yellow and brown spots and some are bacterial and virus diseases. Some research works are done identifying plant disease but they have not covered the broader categories of plant diseases and image features for training purpose to obtain much accurate results for large set of images. Depending upon the texture (shape, size, roughness, intensity, etc.) plant diseases can be predicted in some cases and in others it may require further test. This research is focused on classifying and recognizing images based on the plant leaf textures such as contrast, correlation, entropy, inverse difference moments and leaf colors (red, green and blue). There are sixteen different plant leaf categories considered as healthy or other predefined disease for this work. The models for CNN, KNN, RFC and KNN will be developed and the results will be compared.

This papaer is organized with introduction to emphasize the need of this research work to be done which is followed by reivew section to provide the overview of the works done previously, their limitaions and the further enhancements that will be incorporated by this research work in the plant leaf diesease recgonition. The working meachinism, variuos algorithms used and mathematical represenations used in the work is explanined in the methodology section. The result and discussion section describes the findings of the research work. The overall findings and the possible future enhacements is summarized in the conclusion and future works section.

II. Literature review

Various researchers have used machine learning and image processing techniques for identifying the diseases on different types of plants. The Authors of paper [1] investigated using k-means clustering method for Brinjal leaves with image processing techniques to identify plant leaf disease. The authors performed histogram equalization to increase image quality prior clustering process. Color Co-occurrence Method (CCM method) was used to extract the color and texture features. The features were trained using k-means clustering algorithm with three clusters as infected object, infected leaf and the black background of leaf. However, the features are not sufficient to classify much larger classes of images and clusters with subtle change in colors.

The authors of the paper [2] proposed converting RGB image into HSV and perform color based subtraction of unwanted background by retaining pixel having G value more than R and B values for plant leaf disease classification. The

Manuscript received on June 20, 2020, accepted for publication on July 31, 2020, published on December 30, 2020.

Bijaya Kumar Hatuwal is with the Department of Computer and Electronics Engineering, Himalaya College of Engineering (of Tribhuvan University), Chyasal, Lalitpur, Nepal (e-mail: bjkat28@gmail.com).

Aman Shakya and Basanta Joshi are with the Department of Computer and Electronics Engineering, Institute of Engineering Pulchowk (of Tribhuvan University), Lalitpur, Nepal (e-mail: aman.shakya@ioe.edu.np, basanta@ioe.edu.np).

ISSN 2395-8618

connected elements in the image are discovered out from the cluster based background subtraction and the immense part of the image is kept and other part is removed. They used SVM which created the hyper planes in high dimensional space for categorizing the data points into different classes.

The authors of the paper [3] proposed KNN as an effective method in identifying leaf diseases for agronomical crop images. They used luminance and linear characteristics image to detect skeleton of leaves to determine whether the leaf is of grape or not. Then, GLCM (Gray-Level Co-Occurrence Matrix) features are extracted and diseases are classified by using the obtained grape leaf images. However the detection and recognition was only for grape specific and could not perform well for other species of plant.

The authors of the paper [4] performed Convolutional Neural Network operation for plant disease detection using python API They resized image to 96x96 resolution for image processing. Data augmentation technique was used to rotate, flip, shit images horizontally and vertically. Adam optimizer was incorporated using categorical cross-entropy. They trained the image with 75 epochs using 32 batch sizes for 35000 images. Similarly, the authors of paper [5], proposed framework ResNet50, ResNet101, DenseNet161, and DenseNet169 as their Deep Neural Network (DNN) framework to detect disease in rice plant. Images were resized as 224×224 pixels, the batch size was set to 64 , epoch to 15 and the learning rate was set a constantly of 0.0001. The DenseNet161 produced the best results with an accuracy of 95.74%.

The authors of the paper [6] investigated on using k-means clustering for the image segmentation of grape leaf disease. Shape, color and texture were extracted as main features. Linear Support Vector Machine (LSVM) was used for classification purpose. The images were classified into two classes Downy and Powderly using the extracted nine texture features and nine color features for all three segmented parts of single leaf image.

The authors of the paper [7] proposed using three feature descriptors Hu moments, Haralick Texture and color histogram for plant disease classification using various machine learning algorithms Logistic regression, Support vector machine, k-nearest neighbor, CART, Random Forests and Naive Bayes. The accuracy of machine learning models Support vector machine 40.33%, k-nearest neighbor 66.76% and random forest 70.14 was quite low.

Most of the papers focused on same species of plant for the disease prediction which has almost similar texture. Also comparison among the various algorithms to obtain the better results is done by few authors where the number of image feature considered was fewer and the overall accuracy was low. This study will consider multiple features like contrast, correlation, inverse difference moments, entropy and red, green and blue colors for training and prediction purpose of plant leaf images to get higher accuracy. The sixteen different plant image classification categories will be considered. The machine learning models like SVM, KNN, CNN and RF will be used for training and testing purpose and the accuracy results among the models will be compared.

III. Methodology

There are four major stages involved in the proposed approach as image acquisitions and preprocessing, image features extraction, model training and testing and given input image prediction. Fig. 1 and Fig. 2 represents the detailed flow diagram of the methodology.

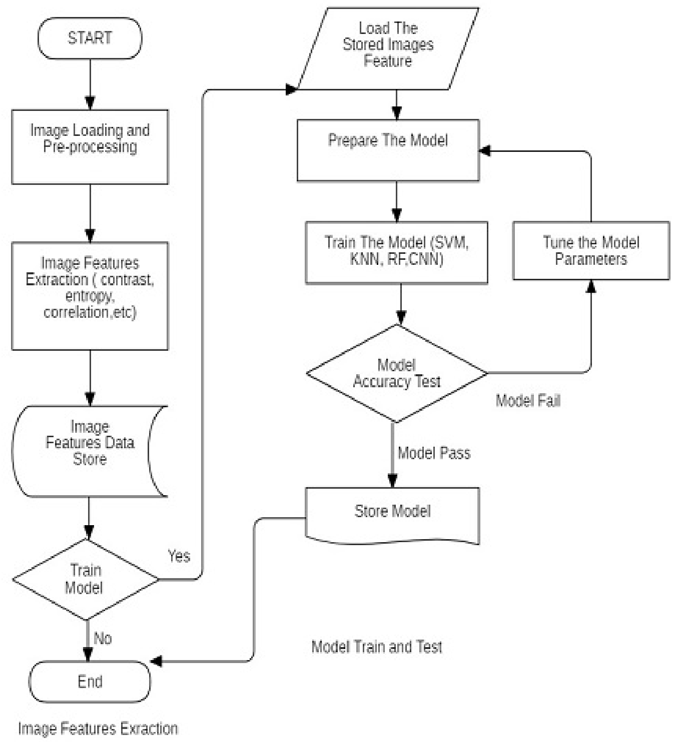

Fig. 1. Image feature extraction and model preparation

A. Image Acquisitions

The images were collected for various plants species and diseases. The images were placed in jpg format. The source images were taken from the Kaggle plant village dataset. The train (folder containing images for training purpose of the models) and valid (folder containing images for validation purpose of the model) folder consists of images in ratio of 80 to 20 for training and testing purpose respectively for sixteen different categories which is shown in Fig. 3' .

B. Image Features Extraction

In total ten properties from color and textures are generated as the features from the images. The mean and standard deviation of each color channel red (R), green (G) and blue (B) are calculated. Then blurring is done after converting

ISSN 2395-8618

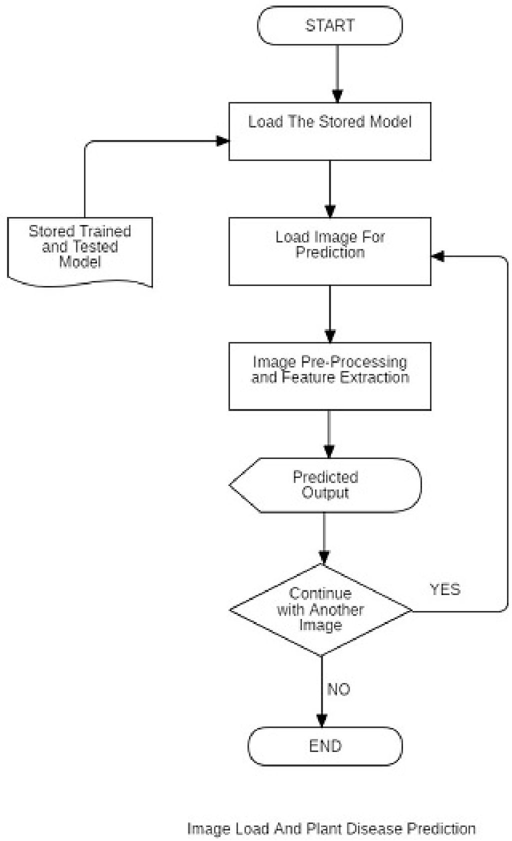

Fig. 2. Plant disease prediction with given image

S.N	Plant Categories as sub folders in root folders Train and Valid	Images count in respective sub folders of root folder (Train)	Images count in respective sub folders of root folder (Valid)
1	Apple___Apple_scab	2016	504
2	Apple___Black_rot	1987	497
3	Apple___Cedar_apple_rust	1760	440
4	Apple___healthy	2008	502
5	Cherry_(including_sour)___healthy	1826	456
6	Cherry_(including_sour)___Powdery_mildew	1683	421
7	Grape___Black_rot	1888	472
8	Grape___Esca_(Black_Measles)	1920	480
9	Grape___healthy	1692	423
10	Grape___Leaf_blight_(Isariopsis_Leaf_Spot)	1722	430
11	Peach___Bacterial_spot	1838	459
12	Peach___healthy	1728	432
13	Pepper,_bell___Bacterial_spot	1913	478
14	Pepper,_bell___healthy	1988	497
15	Strawberry___healthy	1824	456
16	Strawberry___Leaf_scorch	1774	444

Fig. 3. Plant disease images counts in different categories for training and testing

the image into gray scale to reduce the noise level in the image. Gaussian noise is very common kind of noise that is likely to arise for case of any image due to poor illumination or high temperature or transmission [8]. The texture based feature extraction is performed using Haralick texture features algorithm which extracts contrast, correlation, inverse difference moments, and entropy from the images converted as grayscale. The Haralick texture algorithm uses gray-level co-occurrence matrix (GLCM) to calculate the features. GLCM is a matrix that represents the relative frequencies of a pair of grey levels present at certain distance d apart and at a particular angle θ [9]. Extraction of textural information from images containing highly directional characteristics is majorly dependent on selection of correct angle θ [9]. The mathematical equatioin of the features used are represented by Equations (2)–(5).

$$G = \begin{bmatrix} P(1,1), & P(1,2), & \cdots & ,P(1,D_g) \\ P(2,1), & P(2,2), & \cdots & ,P(2,D_g) \\ \vdots & \vdots & \ddots & \vdots \\ P(D_g,1) & P(D_g,2) & \cdots & P(D_g,D_g) \end{bmatrix} \quad (1)$$

$$Contrast = \sum_{n=0}^{Dg-1} x^2 \left\{ \sum_{i=1}^{Dg} \sum_{j=1}^{Dg} P(i,j) \right\}, |i-j| = n \quad (2)$$

where Dg is numbers of gray levels that can be represented by a matrix G having dimension Dg as shown in Equation (1) with any pixel point (i,j) and P(i,j) represents the probability of presence of pixel pairs at certain distance d at angle θ in GLCM image.

$$Correlation = \frac{\sum_{i=1}^{Dg} \sum_{j=1}^{Dg} (i,j) P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (3)$$

where $\mu_x \mu_y$ are means and $\sigma_x \sigma_y$ are standard deviations of Px and Py the partial derivative function.

$$Inverse\ Difference\ Moments = \sum_{i=1}^{Dg} \sum_{j=1}^{Dg} \frac{1}{1+(1-j)^2} P(i,j) \quad (4)$$

$$Entropy = -\sum_{i=1}^{Dg} \sum_{j=1}^{Dg} P(i,j) \log [P(i,j)] \quad (5)$$

C. Model Training and Testing

The image features extracted were split in the ratio eighty to twenty for the training and testing purpose for the SVM, KNN, K-means clustering and random forest.

There are many ways to solve Multi-class classification problems for SVM such as Directed Acyclic Graph (DAG), Binary Tree (BT), One Against-One (OAO) and One-Against-All (OAA) classifiers [10]. Constructing an optimal hyper plane regarded as the decision surface using the input samples to make the two sides' margin largest is the main mission of support vector machine [11]. To

ISSN 2395-8618

perform the multi class classification using SVM we have used one-versus-rest method and Gaussian radial basis function. Radial basis function (RBF) kernel (is a positive parameter for controlling the radius) [10] which is given by Equation (6): As SVM cannot perform multiclass classification at once, one-versus-rest method is used by SVM which does binary operation with each dataset to finally make multiclass classification.

$$K(xi, xj) = exp\left(\frac{-\|xi - xj\|^2}{2\sigma^2}\right) \tag{6}$$

where k is the kernel function , $x_i = (x_{i1}, x_{i2}, \ldots, x_{iN})$ corresponds to the attribute set for the ith sample in each sample tuple represented by (x_i, x_j) in N training data of a binary classification.

Random Forest can be used for classification and regression. The proposed methodology uses classification using Random forest. It is an ensemble method as uses the average or voting from multiple decision trees to reach the decision. The problem of over fitting is also reduced drastically by reducing variance in this algorithm.

K Nearest Neighbor (KNN) can be used for both classification and regression operation. In pattern acknowledgement, the (KNN) k nearest neighbors algorithm is a non-parametric method used for classification and regression [12]. KNN performs classification based on the majority voting on similarity to K nearest number of neighbors calculated using distance functions. Some most common distance functions are Euclidean, Manhattan, Minkowski and Hamming distance. For categorical variables Hamming distance is used and for continuous variables Euclidean, Manhattan and Minkowski distance calculation are used.

$$Euclidean\ Distance = \left(\sum_{i=1}^{k}(x_i - y_i)^2\right)^{1/2} \tag{7}$$

$$Hamming\ Distance\ (H_D) = \sum_{i=1}^{k}\|x_i - y_i\| \tag{8}$$

where $x = y \Rightarrow D = 0, x \neq y \Rightarrow D = 1$, x =($x_1$,$x_2$,$x_3$,. . .,$x_k$) and y=($y_1$,$y_2$,$y_3$,...,$y_k$) are the points in the space and D is the distance between the two points x and y.

Convolutional Neural Network (CNN) is preferred as a deep learning method in this study. CNN, which can easily identify and classify objects with minimal pre-processing, is successful in analyzing visual images and can easily separate the required features with its multi-layered structure [13]. The major layers in CNN consist of convolutional layer, pooling layer, activation function layer and fully connected layer. Python library scikit-learn was used with initial image width and height equal to 180px. Multiclass classification with sixteen number of classes was done. The number of epoch used was calculated as: (Total number of images) mod (Batch size). Convolutional Kernel of size (3, 3), max pooling matrix of size (2, 2) and ReLu was used as an activation function in each convolutional layer. The convolutional layer output was flattened with 0.5 dropout value. Sigmoid activation function was used in fully connected layer with dense value as 16 was used for the experiment.

D. Given Input Image Prediction

The input image absolute file path is provided as an input. For the given image, feature extraction is performed. The extracted featured is used for the prediction using the previously saved machine learning models. The predicted plant category (healthy or disease) name as shown in Fig. 5 is displayed on the integrated development environment console or in browser.

IV. Results And Discussion

For support vector machine we achieved the accuracy of 78.61% where regularization parameter C is set to 100, gamma to 0.0001 and tolerance in optimization to 0.001. The given hyper parameters values were changed manually to get the best possible accuracy. Fig. 4 and Fig. 5 show the graphical user interface for image feature extraction and plant disease prediction. Table. I. Shows the precision, recall, f1-score and support of testing images for SVM based classification

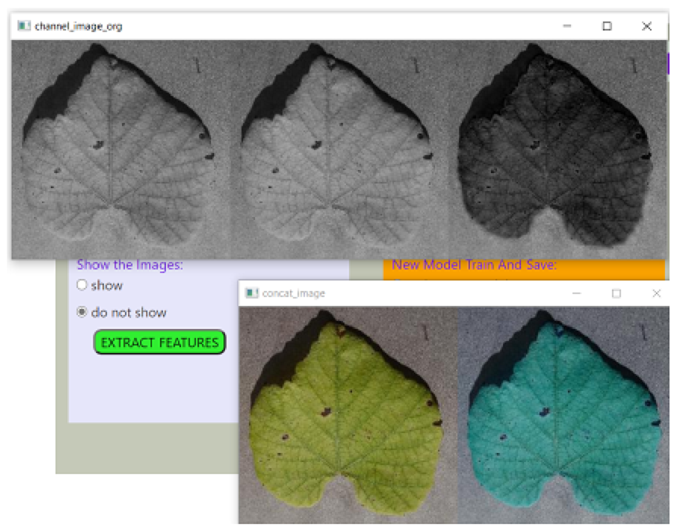

Fig. 4. Image feature extraction

Feature extraction process of image with different color channel (red, green and blue) and original image is shown in Fig. 4.

In KNN the nearest number of neighbors k value is used as 5 and an accuracy of 76.969% is obtained. Though best accuracy can be obtained at k=1 we used k=5 to prevent use of single value voting for prediction. The plot for KNN is show in Fig. 6 and Fig. 7.

The plot Fig. 6 shows the multiclass classification of the images of sixteen different categories in testing. It shows the number of times the image of given category is categorized in which category moving along the truth and predicted axes as

ISSN 2395-8618

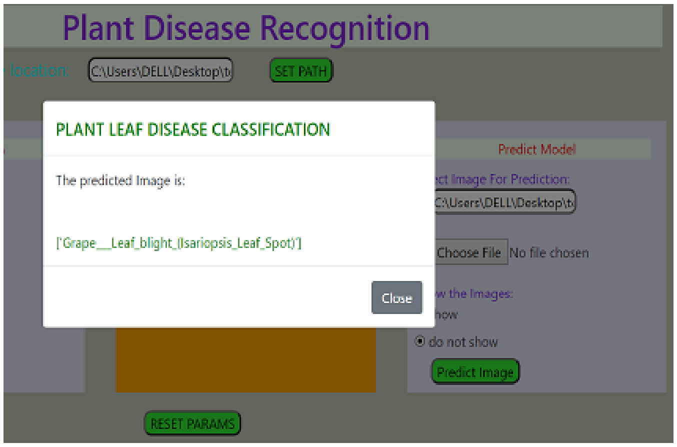

Fig. 5. Plant disease prediction using SVM

TABLE I
SVM CLASSIFICATION REPORT FOR TESTING DATA

Labels	***Precision***	***Recall***	***F1-Score***	***Support***
Apple___Apple_scab	0.75	0.68	0.71	418
Apple___Black_rot	0.62	0.73	0.67	380
Apple___Cedar_apple_rust	0.79	0.78	0.79	326
Apple___healthy	0.88	0.85	0.87	411
Chery__Powdery	0.82	0.88	0.85	346
Cherry__healthy	0.91	0.96	0.93	356
Grape___Black_rot	0.63	0.75	0.69	356
Grape__(Black_Measles)	0.75	0.71	0.73	365
Grape___blight_(Isariopsis)	0.81	0.83	0.82	327
Grape___healthy	0.83	0.87	0.85	340
Peach___Bacterial_spot	0.81	0.75	0.78	384
Peach___healthy	0.91	0.92	0.91	341
Pepper_bell_Bacterial_spot	0.68	0.59	0.63	388
Pepper_bell_healthy	0.76	0.68	0.72	410
Strawberry___Leaf_scorch	0.89	0.85	0.87	371
Strawberry___healthy	0.84	0.88	0.86	395

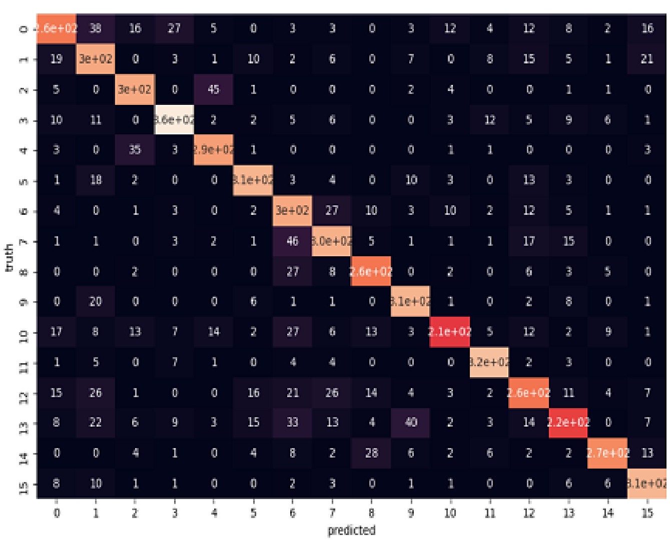

Fig. 6. Heatmap plot for KNN

shown in the figure. The weighted average value for precision, recall, f1-score and support are 0.78, 0.77, 0.77 and 5914 respectively for testing images for KNN.

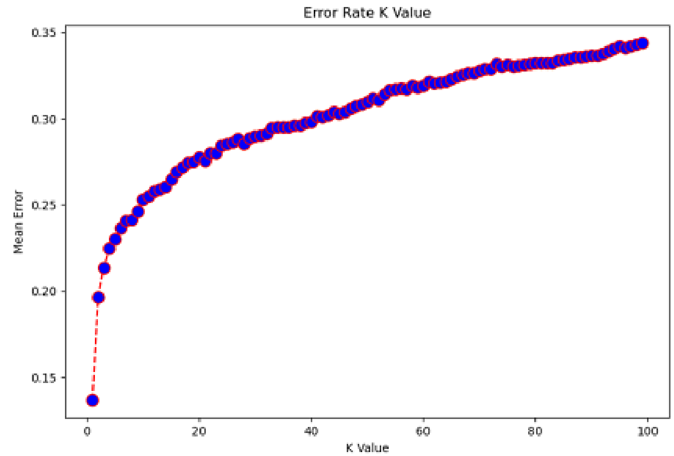

Fig. 7. Elbow criterion plot for KNN

The Elbow criterion plot in Fig. 7 shows the mean error for the given K value in the iteration. The least mean error value as shown is for k 1 but we took K value as 5 to prevent single value voting which might consider the outlier data as true positive value.

Random Forest model with accuracy 87.436% is created with 250 numbers of estimators and the heatmap plot as shown in Fig. 8. The weighted average value for precision, recall and, f1-score is 0.88 and support value is 5914 for testing images for KNN.

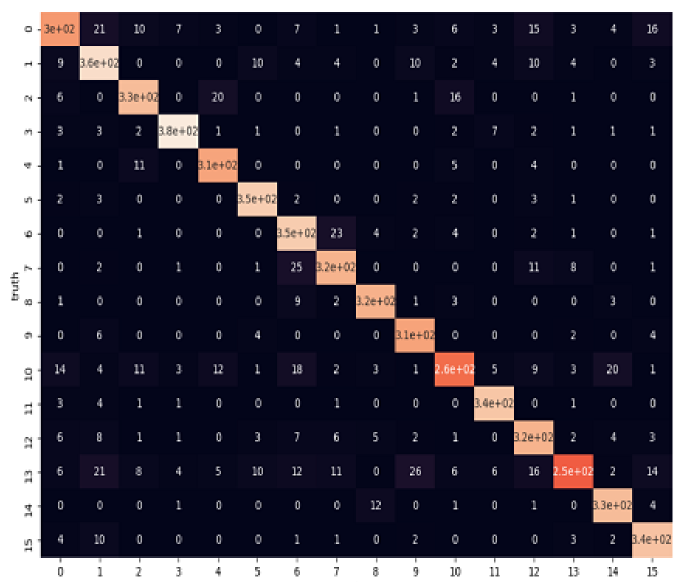

Fig. 8. Heatmap plot for random forest

Convolutional Neural Network model has training accuracy of 97.89% and validation accuracy of 99.01% which is trained

ISSN 2395-8618

for 147 epochs with 29567 and 7391 images for training and validation respectively. The CNN model was trained using the Google Colaboratory GPU named as device:GPU:0. The model accuracy and loss plot for training and testing is shown in Fig. 9.

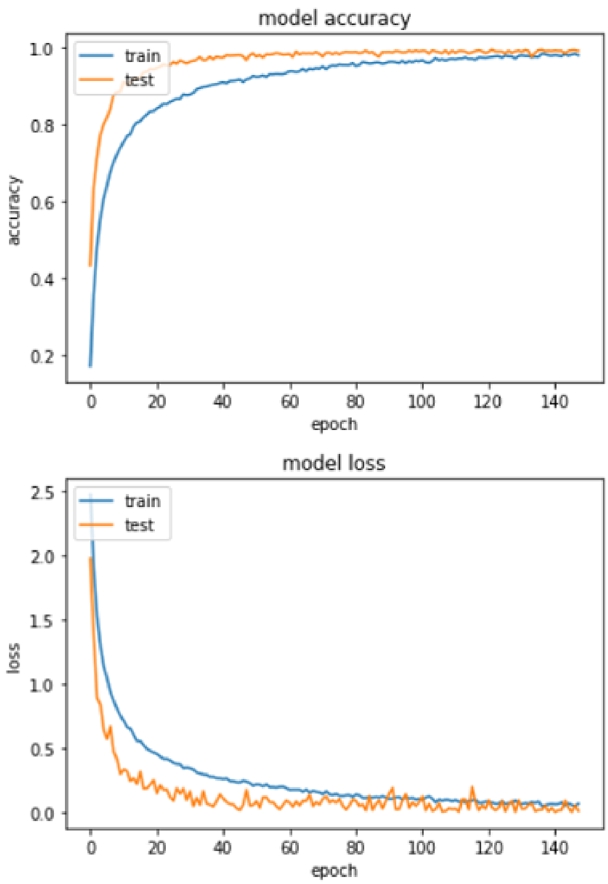

Fig. 9. CNN accuray and losss plot for train and test

TABLE II
TESTING DATA OUTPUT COMPARISON OF DIFFERENT MODELS

Model Name	***Accuracy in %***
Convolutional Neural Network (CNN)	97.89%
Random forest (RF)	87.43%
Support Vector Machine (SVM)	78.61%
K Nearest Neighbor (KNN)	76.96%

The CNN produced the result with much accuracy than other machine learning model but the time and the physical resource like RAM and CPU (Google Colaboratory GPU named as device:GPU:0) used by CNN is high compared to other models. The obtained accuracy of work for CNN is 2.52% higher than the accuracy obtained by author of paper [4] which was 94.74%. The accuracy of SVM, KNN and RF obtained in our work is higher than the authors of paper [7] where SVM, KNN and RF produced the highest accuracy of 40.33%, 66.76% and 70.14% respectively.

V. CONCLUSION AND FUTURE WORKS

This work presents the various plant leaf diseases recognition using Haralick feature extraction technique and machine learning models like SVM, KNN, Random forest and CNN. Contrast, correlation, inverse difference moments, entropy and images RGB color standard deviation are extracted image features for this work. Among the given models CNN produced the highest level of accuracy 97.89% followed by Random forest 87.436%, SVM 78.61% and KNN 76.969%.

Future works can be done to include more plant species with different diseases and texture characteristics. Further improvements in the given models prediction can be done by extracting much distinct features from the plants leaves. Grid search or other algorithms can be used to find the best optimal value of the hyper parameters used in different models used in efficient way.

REFERENCES

[1] R. Anand, S. Veni, and J. Aravinth, "An application of image processing techniques for detection of diseases on brinjal leaves using k-means clustering method," in *2016 International Conference on Recent Trends in Information Technology (ICRTIT)*, Chennai, 2016, pp. 1–6, doi: 10.1109/ICRTIT.2016.7569531.

[2] C. G. Dhaware and K. H. Wanjale, "A modern approach for plant leaf disease classification which depends on leaf image processing," in *2017 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, 2017, pp. 1–4, doi: 10.1109/ICCCI.2017.8117733.

[3] N. Krithika and A. G. Selvarani, "An individual grape leaf disease identification using leaf skeletons and KNN classification," in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, 2017, pp. 1–5, doi: 10.1109/ICIIECS.2017.8275951.

[4] S. V. Militante, B. D. Gerardo, and N. V. Dionisio, "Plant leaf detection and disease recognition using deep learning," in *2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, Yunlin, Taiwan, 2019, pp. 579–582, doi: 10.1109/ECICE47484.2019.8942686.

[5] S. Mathulaprangsan, K. Lanthong, D. Jetpipattanapong, S. Sateanpattanakul, and S. Patarapuwadol, "Rice diseases recognition using effective deep learning models," in *2020 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, Pattaya, Thailand, 2020, pp. 386–389, doi: 10.1109/ECTIDAMTNCON48261.2020.9090709.

[6] P. B. Padol and A. A. Yadav, "SVM classifier based grape leaf disease detection," in *2016 Conference on Advances in Signal Processing (CASP)*, Pune, 2016, pp. 175–179, doi: 10.1109/CASP.2016.7746160.

[7] S. Ramesh, R. Hebbar, M. Niveditha, R. Pooja, B. N. Prasad, N. Shashank, and P. Vinod, "Plant disease detection using machine learning," in *2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C)*, Bangalore, 2018, pp. 41–45, doi: 10.1109/ICDI3C.2018.00017.

[8] G. Mukherjee, A. Chatterjee, and B. Tudu, "Study on the potential of combined GLCM features towards medicinal plant classification," in *2016 2nd International Conference on Control, Instrumentation, Energy & Communication (CIEC)*, Kolkata, 2016, pp. 98–102, doi: 10.1109/CIEC.2016.7513746.

[9] S. Singh, D. Srivastava, and S. Agarwal, "GLCM and its application in pattern recognition," in *2017 5th International Symposium on Computational and Business Intelligence (ISCBI)*, Dubai, 2017, pp. 20–25, doi: 10.1109/ISCBI.2017.8053537.

[10] F. F. Chamasemani and Y. P. Singh, "Multi-class support vector machine (SVM) classifiers—an application in hypothyroid detection and classification," in *2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications*, Penang, 2011, pp. 351–356, doi: 10.1109/BIC-TA.2011.51.

ISSN 2395-8618

[11] S. Xiaowu, L. Lizhen, W. Hanshi, S. Wei, and L. Jingli, “Image classification via support vector machine,” in *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, Harbin, 2015, pp. 485–489, doi: 10.1109/ICCSNT.2015.7490795.

[12] M. P. Vaishnnave, K. S. Devi, P. Srinivasan, and G. A. P. Jothi, “Detection and classification of groundnut leaf diseases using KNN classifier,” in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India, 2019, pp. 1–5, doi: 10.1109/ICSCAN.2019.8878733.

[13] M. Sardogan, A. Tuncer, and Y. Ozen, “Plant leaf disease detection and classification based on CNN with LVQ algorithm,” in *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, Sarajevo, 2018, pp. 382–385, doi: 10.1109/UBMK.2018.8566635.

ISSN 2395-8618

Complex System View on Natural Language

Olga Kolesnikova

***Abstract*—The study of complex systems in many sciences such as physics, chemistry, biology, engineering, economics, psychology, among others, has demonstrated itself as a powerful approach to resolve many hard issues and to contribute to a fuller and more realistic description of various phenomena. As different to other types of systems, complex systems are characterized by such properties as self-organization, emergence, openness, dynamic nature, chaoticity, fractality, catastrophism, nonlinearity, and fuzziness. It turns out that natural language as a system possess many indispensable properties of complex systems, so it can be viewed and studied as a complex adaptive system. We show that the complex system view on natural language is powerful not only to incorporate knowledge of language accumulated by traditional linguistics, but further make significant discoveries on many open issues in phonetics, grammar, lexicon, language origin and evolution, first language acquisition and development, simulating language functionalities by computational models.**

***Index Terms*—Adaptive dynamic complex systems, agent-based model, cognitive model, emergence, usage-based grammar.**

I. Introduction

IT has been a long and well-grounded scientific approach to study a human, or natural, language as a system comprised of several layers: phonetic, phonological, morphological, syntactic, semantic, and pragmatic. In its turn, each layer is also a system of its proper interrelated elements, so the layers are subsystems within the system. Traditionally, these subsystems are objects of research in the respective fundamental linguistic subdisciplines: phonetics, phonology, morphology, syntax, semantics, and pragmatics. There are many other branches in linguistics focusing on specific language elements or aspects, giving a closer look at selected finer language details; examples of such branches are articulatory and acoustic phonetics, lexicology, conversation analysis, text linguistics, stylistics, among others.

There are also other dimensions of our interest in language: we desire to explore its diversity in language typology, its variation and change over time in historical linguistics and evolutionary linguistics, its usage in different locations in dialectology and by different social groups in sociolinguistics and ethnolinguistics. Language is analyzed with its relation to culture in linguistic anthropology, to brain/mind activities in neurolinguistics and psycholinguistics, to human cognition in cognitive linguistics, its acquisition as first or second language is another area of linguistics.

Manuscript received on August 17, 2020, accepted for publication on September 23, 2020, published on December 30, 2020.

The author is with the Escuela Superior de Cómputo, Instituto Politécnico Nacional, Mexico (e-mail: kolesolga@gmail.com).

Many interesting discoveries have been made in all these disciplines, many language facets and properties have been described and analyzed, indeed, an enormous amount of knowledge have been gained and applied in miscellaneous tasks. However, the knowledge we have is segmental and uncoordinated, each linguistic branch seems to be separate and isolated to some degree from the other branches, and still little is understood how all the branches interconnect and interact to explain and represent language in its totality and wholeness. The accumulated evidence and facts still remain in detached “storehouses” whose walls prevent them all from composing the complete picture.

Nowadays, a promising approach which can lead to a solution of the above issue is adopting the view on language as a complex system. Complex system framework has been proven successful in many other sciences: physics, chemistry, biology, engineering, computer science, economics, psychology, health studies, education. In this article we explain the concept of a complex system, its properties, and review recent works on modeling a natural language as a complex system mentioning some applications this model might have.

A. Systems and their Types

The study of complex systems is an area of systems theory where a system is defined as a construct with “the following characteristics: it consists of a set of objects, a set of relations between the objects, and a structure of layers, and it interacts with its environment” [1, p. 25].

There are many types of systems distinguished depending on their properties. Systems can be open or closed: an open system interacts with its environment, it has input/output and reacts to events occurring in its surroundings, and the opposite is a closed or isolated system (see http://pespmc1.vub.ac.be). According to their time-related behavior, systems can be static or dynamic: a static system is not time-dependent, dynamic systems evolve over time exhibiting continuous, discrete, or hybrid change; such systems can be stable or unstable, and an extreme case of an unstable system is a chaotic system [2]. Depending on the characteristics of output with respect to input, dynamic systems can be linear or non-linear, in a non-linear system output is not directly proportional to input [3]. With respect to system’s reaction to environmental stimuli, it can be adaptive, customizing its structure and comportment to its changing context, otherwise it is non-adaptive [4].

B. Complex Systems

In what way does a complex system differ from any other kind of system? While there has not yet been developed a single answer to this question accepted by the whole research

ISSN 2395-8618

community, specialists in physics, biology and social sciences—three research fields from which this new science of complex systems has emerged [5]—come up with their own vision on what makes a system complex.

For instance, the physicists Amaral and Ottino [6] distinguish among simple, complicated, and complex systems in the following way. A system is simple if it contains a small number of components which act according to well-understood laws, giving a prototypic example of a pendulum. Then, a system with a large number of components which have well-defined roles and are governed by well-understood rules is defined as complicated, for example, a Boeing 747-400. For a complex system, the authors give an example of a migrating geese flock, which at first glance looks much simpler than the Boeing, however, it is different from the Boeing in the sense that it is an adaptive system, i.e., its behavior is emergent, it changes depending on the environment, and the flock is also self-organizing, i.e., the roles of the components (geese) may also change.

To summarize the considerations in [6], a complex system, as different from the other types of systems, is characterized by self-organization and emergence: it is a set of components (agents) whose roles may be fluid and the relations between the components may be plastic, it means that the rules that govern the behavior and connectivity of the components as well as their roles may change over time. Besides, complex systems also possess other properties which may be not unique to them but can be observed in other types of systems: openness, dynamic nature, chaoticity, fractality, catastrophism, nonlinearity, and fuzziness [7].

II. Complex System View vs. Traditional View on Natural Language

Knowing that a complex system is a structure with the properties mentioned in Section 1.2, and that in a complex system the components and rules of their interaction and behavior may change, can we observe such features in natural language? The researchers who call themselves the "Five Graces Group" [8] affirm that natural language has characteristics allowing to consider it as a complex adaptive system (CAS):

"Language as a CAS involves the following key features: The system consists of multiple agents (the speakers in the speech community) interacting with one another. The system is adaptive; that is, speakers' behavior is based on their past interactions, and current and past interactions together feed forward into future behavior. A speaker's behavior is the consequence of competing factors ranging from perceptual constraints to social motivations. The structures of language emerge from interrelated patterns of experience, social interaction, and cognitive mechanisms. The CAS approach reveals commonalities in many areas of language research, including first and second language acquisition, historical linguistics, psycholinguistics, language evolution, and computational modeling." [8, p. 1–2]

The authors go on to say that the processes involved in language acquisition, language use and change, although studied separately in different areas of linguistics and viewed as different systems, in fact are elements of only one complex system. This system is comprised of interacting speakers (system's agents) whose language behavior is adaptive as it depends on their experience of interaction among themselves and with the real world. The complex system of language also exhibits such characteristics as distributed control and collective emergence, intrinsic diversity, perpetual dynamics, adaptation through amplification and competition of factors, nonlinearity and phase transitions, sensitivity to and dependence on network structure, and local change [8].

Discourse
Pragmatics
Semantics
Syntax
Morphology
Phonology
Phonetics

Fig. 1. Linguistic subdisciplines respective to the levels of language they study Magnetization as a function of applied field.

Different from the traditional view on language as a system of layers abstracted from individual language users (Fig. 1), the complex system approach sees language as adaptive and self-organizing communication behavior of agents (speakers), emerging from interaction among them as well as with environment and developing according to cultural selection and structural coupling [9]. Compare Fig.1 with Fig. 2, where basic components, processes, and relations in a complex system are presented [10][1]. Although this diagram is general and represent a "skeleton" of any CAS, not developed specifically for language, it conveys fundamental principles of complexity and adaptivity found in any domain of reality. The vision of CAS expressed by Massip-Bonet [2] supplements the diagram in Fig. 2:

"Complex adaptive systems are systems that learn or evolve in the same manner as living beings. They seek patterns. They interact with the environment, they "learn" from experience and, as a result, they adapt. They contain information on the environment; in a special sense, they "know" it. The common trait of complex adaptive systems is that they all process information in some way" [2, p. 40].

[1] https://www.innovationlabs.com/summit/discovery1/

ISSN 2395-8618

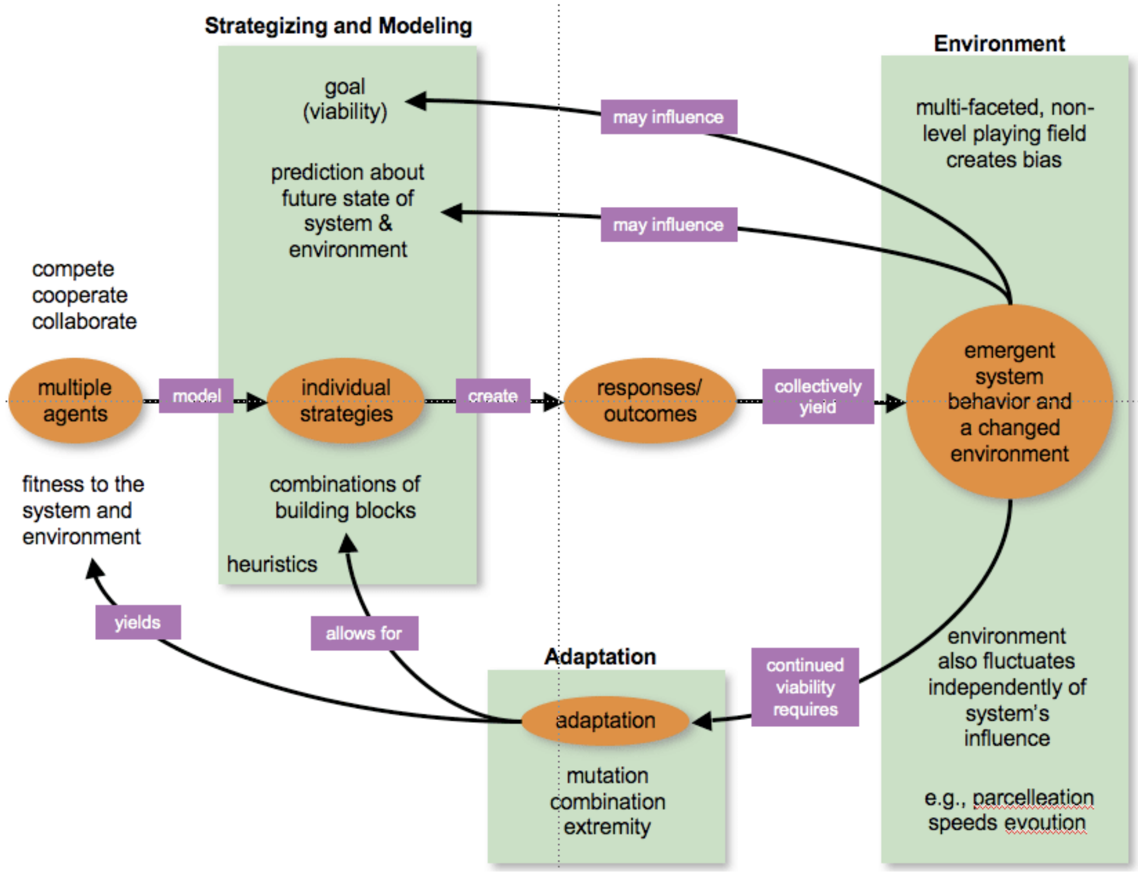

Fig. 2. Complex adaptive systems.

The traditional structural language modeling follows the strategy "divide and conquer" by segmenting its object of study into smaller thus more manageable parts, further analyzing their elements and relations. It seeks generalizations and comes up with well-defined concepts and categories [11]. Categories can be defined in terms of characteristic features or resemblance to a prototype. However, not all linguistic phenomena fit perfectly into elaborated categories, there are cases difficult to resolve and assign to a specific category.

As an example, let us consider the concept of collocation usually defined as a combination of words whose meaning cannot be derived from the meaning of its constituent words. Therefore, the distinctive feature of collocations is their semantic non-compositionality. Here are some examples of collocations with their meaning in parenthesis: *have breakfast* (consume a morning meal), *break a habit* (stop doing something that you do regularly, especially something that you should not do), *dead-end job* (a job in which there is no chance of progressing to a better, more important job), *call it a day* (bring something to an end), *find one's feet* (become more comfortable in whatever one is doing), *under the weather* (sick), *hit the sack* (go to sleep)[2].

Observing these examples, it can be noticed that the meaning of some collocations is closer to the composition of the meaning of their constituents (*have breakfast*, *break* a *habit*), while the meaning of others cannot be analyzed compositionally (*under the weather*, *hit the sack*). The remaining collocations (*dead-end job*, *call it a day*, *find one's feet*) can be situated within the range given by these two extremes: compositional semantics and completely fused semantics. It looks like non-compositionality is graded, it varies within an interval starting from zero non-compositionality in free word combinations with fully compositional semantics (*write a letter*, *buy a dress*) to maximum non-compositionality in idioms: *pull someone's leg* (joke with someone), *beat around the bush* (avoid saying what you mean because it is uncomfortable)[3]. Reddy *et al*. [12] studied compound nouns like *climate change*, *crash course*, *spelling bee*, *cash cow* and proved that their compositionality varied over a continuum of values, see Fig. 3 [12].

It means that linguistic categories are fuzzy rather than crisp sets, Fig. 4 [7]. To study fuzzy sets, we need tools not available within the traditional linguistic paradigm analysis but well suited for complex system approach [13].

[2] Examples from Merriam-Webster Dictionary https://www.merriam-webster.com and Cambridge Dictionary https://dictionary.cambridge.org.

[3] Examples from https://www.ef.com/wwen/english-resources/english-idioms.

ISSN 2395-8618

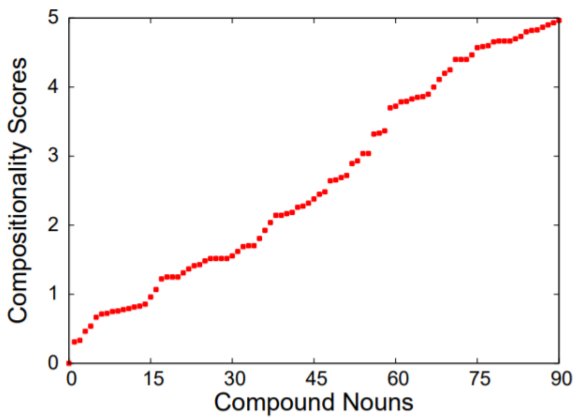

Fig. 3. Mean values of compositionality scores.

Fig. 4. Fuzzy vs. crisp set.

In theory, taking into account the above said, it can be expected that the complex system approach would allow for a more realistic description of natural language. We hope that it is able to assist in discovering previously unknown regularities and patterns useful in various tasks where spoken or written texts have to be interpreted manually or automatically. Besides, the complex system view can open a new perspective on observations and knowledge accumulated within the traditional linguistics, incorporating them into a wider linguistic map. This is how Cameron and Larsen-Freeman [14] put it: "Analysis or investigation of discourse from a complex systems perspective does not require us to throw away other approaches and their techniques. Indeed, multiple types of analysis are needed to work with information from systems at different scales, and new ways of blending methods are needed to explore simultaneous activity on several scales" [14, p. 236].

In the next section, we discuss some results achieved in fields of language study, where complex system strategies have begun to unfold their potential in helping researchers to acquire more insight and understanding.

III. Complex Models of Language

To model language as a complex adaptive system, we are to study not only its discrete constituents but also their interactions as they produce self-organized behaviors. Such interactions are difficult to describe using the traditional conceptual apparatus, so computational simulation and robotic applications come to rescue in this case [15].

Machine learning algorithms assist in finding and evaluating inter-dependency relations among system's constituents, artificial neural networks are used to model the brain activities of the system's agents, their interactions and collective performance. Cellular automata simulate emergence and evolvement of language components, robotic systems allow us to experiment with language phenomena which involve embodiment: perception and action of speakers. In the following sections we will consider complex models developed in scientific endeavors aimed to advance in resolving some interesting issues and challenges posed in language studies.

A. Language Origin and Evolution

Since long ago, researchers have been interested in how natural language emerged as a means of communication among human agents, why and how it undergoes changes and evolves over time. Recently, swarm robotic complex systems have been used to investigate the emergence of language intercourse. The objective of such systems is to model collective self-organizing communication through interactions of multiple simple robots. A tool proposed by Cambier *et al*. [16] to simulate the evolvement of language communication among robots is language games. The authors describe several language games such as imitation games, guessing games, and category games, further emphasizing that a best suited game for robotic swarms to develop communication activities is the minimal naming game explored earlier by other researches in different environments [17–20].

The minimal naming game is played by two or more robots; initiating the game, each robot is provided with a set of artificially generated words and then, in the course of the game, it takes the role of a transmitter/speaker or a recipient/hearer in turns. In each step of the game, the goal for the robots is to reach an agreement on the choice of a word from their sets of words, which is to be associated with a given object; selecting words in this way, the robots create a vocabulary of objects' names. Fig. 5 is an illustration of the game in its simple version of two participating robots [16].

First, the robot-transmitter selects a random word from its vocabulary and messages it to the robot-recipient. Then, if the recipient also has this word in its vocabulary, the agreement on the object's name is reached and the game is success; otherwise the game is failure, in such case the recipient has to insert the word in its vocabulary. Cambier *et al*. [16] argue that the minimal naming game possesses a big latent capability to generate an emergent language in systems of robotic swarms aiming at collectively fulfilling a specific task.

The language game approach has been shown productive in another work on language evolution, the research done by Vera *et al*. [21]. The authors focused on the emergence of the fundamental property of any natural language, namely, Zipfian distribution of words in vocabulary. They modeled human communicative decentralized interactions with a bipartite graph where words were mapped to their meaning. Phase transitions in numerical simulations executed on the model

ISSN 2395-8618

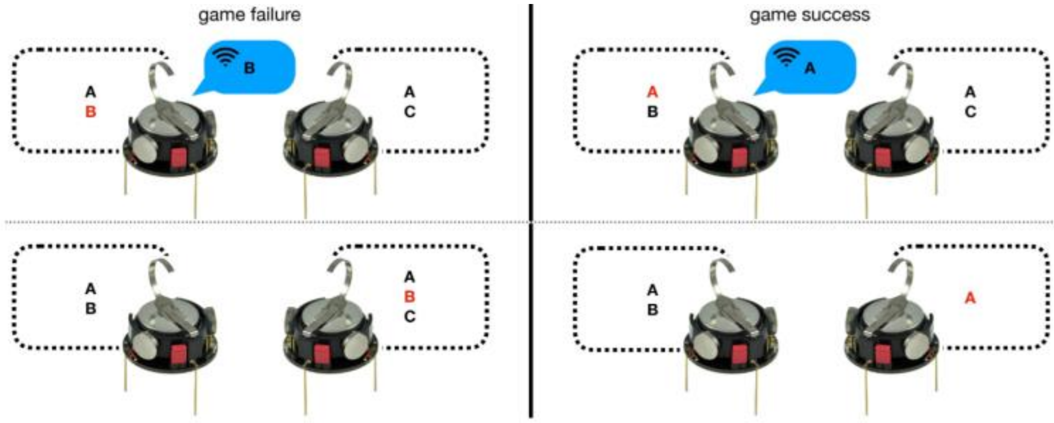

Fig. 5. The minimal naming game played by two robots.

converged to the state where the vocabulary revealed a word distribution in accordance with the Zipf's law, thus demonstrating how such distinctive lexicon property might have originated.

Another important question we can pose considering language origin and evolution is related to speech sounds. How did speech sounds appear, and simultaneously, how did they acquire the ability to form meaningful structures, i.e., words? The first attempt most cited in the literature to answer this tough question within the complex systems framework was the experiments in [22]. Using robotic simulation, de Boer modeled the origins of vowels by means of imitation games. Each robotic agent was equipped with a sound feature analyzer and a sound synthesizer, imitating human acoustic and articulatory apparatus. The agents in turns produced and perceived random sounds. If an agent who received the sound produced by another agent was able to imitate it, the outcome in the game was count successful, if it could not reproduce it, it was a failure. After 4,000 games, the sounds evolved in the system showed features very close to those of human language vowels. This result is a strong evidence for evolution of language as a self-organizing system, opposite to other theories on this topic, see a substantial review of these in [23].

Another model of phonological system evolution is presented in [24], showing how local interactions among speakers can cause the emergence of a global system. The authors efficiently modeled basic motor, auditory, and sensory-motor factors in play in the process of language formation. However, there are other factors—cognitive, environmental, social, cultural—present in the context of language development. Due to this multi-causal scenario, it is rather complex to incorporate all factors into a single model; to our knowledge, such model has not yet been designed. Besides, some researchers assume that there exist still other driving forces influencing language evolution, and among them, curiosity as an intrinsic human capability.

Speaking of curiosity, we need to note here that there have already been made several contributions to the effort of integrating this human faculty into complex evolutionary models. An interesting suggestion is [25]. These researchers conducted an experiment with a system of robotic agents using a strategy of curiosity-driven learning, interpreting curiosity as "an epistemic motivational mechanism that pushes an organism to explore activities for the primary sake of *gaining information* (as opposed to searching for information in service of achieving an external goal like finding food or shelter)" [25, p. 493]. The authors implemented curiosity as a mechanism using which the robots, in an active learning environment, were acquiring skills allowing them to decrease uncertainty. In the course of the experiment, the emergence of more complex behaviors adapting to the constraints of the environment was observed.

In this section we reviewed some of the works on the issue of language origin and evolution in the light of the complex system view. These have been initial attempts to model and analyze sophisticated processes involved in language development, trying to account for factors commonly neglected in traditional linguistic studies for the sake of simplicity. The proposed models still need further improvement, and we hope that in future, more elaborated complex models will be proposed.

B. *Phonology*

In this section we continue the discussion of sound system emergence and evolvement started briefly in Section 3.1, where we focused more on language in its totality. Here we will speak specifically of the phonological system as viewed and studied within the complex system framework.

de Boyer's model of vowel system emergence [22], introduced shortly in Section 3.1, inspired other scientists to apply the complex system approach to studying phonology. In fact, de Boyer himself extended his robotic vowel model to the sound system as a whole [26], since the former was found to explain the emergence of language sounds in a very realistic way. Now we will give more details on de Boyer's simulation.

In de Boyer's model, the evolvement of sounds was simulated by means of an imitation game with the basic rules

ISSN 2395-8618

shown in Fig. 6 [26], reproduced in [27, p. 141] through self-organization of robotic agents' sound production and perception activities. The simulation included two mathematical models. The first model was responsible for generating sounds whose format frequency values were computed based on a number of equations designed for this purpose. This model imitated the human articulatory mechanism. The second model was developed for imitating the human auditory system. Using the second model, the robots perceived sounds.

Initiator (teacher)	**Imitator (learner)**
If $V = \emptyset$ then Add random vowel to V	
Pick random v from V, Increment count of uses of v Produce signal A_1 from v	

	Receive A_1
	If $V = \emptyset$ then
	    v_{new} = Find phoneme(A_1)
	    $V = V \cup v_{new}$
	Calculate $v_{received}$
	Produce signal A_2 from $v_{received}$

Receive A_2
Calculate $v_{received}$
If $v = v_{received}$ then
    Send non-verbal feedback: success
    Increment count of successful use of v
Else
    Send non-verbal feedback: failure	

Do other updates of V (see below)	Receive non-verbal feedback Update V according to feedback signal Do other updates of V (see below)

Fig. 6. Basic rules of imitation game.

The game was played as follows. At the beginning, each robot was equipped with a set of vowels. Then, in the course of interactions, robots, in turns, were emitting sounds and perceiving them, as well as reproducing them as an evidence of perception. After perceiving and reproducing a sound, a robot compared it with the sounds in its set. If it was similar to a sound in the set, it was merged with this sound, otherwise it was added as a new sound to the set. Thus, while the game proceeded, the sound systems of robots went through a continuous process of developing and updating. Fig. 7 [26], reproduced in [27, p. 141] presents a part of the emergent sound system obtained by aggregating the sound systems of all robots into a single structure. Comparing the emerged system with the human sound system, the former was found to be very similar to the latter.

The work of de Boyer [22], [26] was dedicated to the issue of the sound system origin as such system emerged resulting from self-organizing activities of speakers-hearers. Another interesting issue in phonology is sound change over time. The questions we may ask here are why sounds change, what the causes of sound change are, in what way sounds change, and how this change is preserved and becomes fixed thus turning into a conventional pattern accepted by all language speakers, i.e., how a changed sound becomes a phoneme.

One of the works aiming at answering these tough questions is [28]. The authors explained the sound change with a model of human speech processing, studying this phenomenon in the context of interaction between the speaker and hearer, analyzing the processes of speech production, transmission, and perception. The authors reviewed a number of cognitive-computational models of sound change and proposed two contributions which could improve the models and shed more light on how phonetic variation leads to phonologization, that is, when a diverge from a given phoneme results in fixing the new phonetic changes in a phonemic contrast. The first contribution was emphasizing the necessity to add the aspect of hyper- and hypoarticulation in relation to pragmatic meaning, and the second contribution was the proposal that a sound change model could take advantage of associations between perception and production if they are introduced mathematically, dynamically mapping articulatory patterns to acoustic aspects.

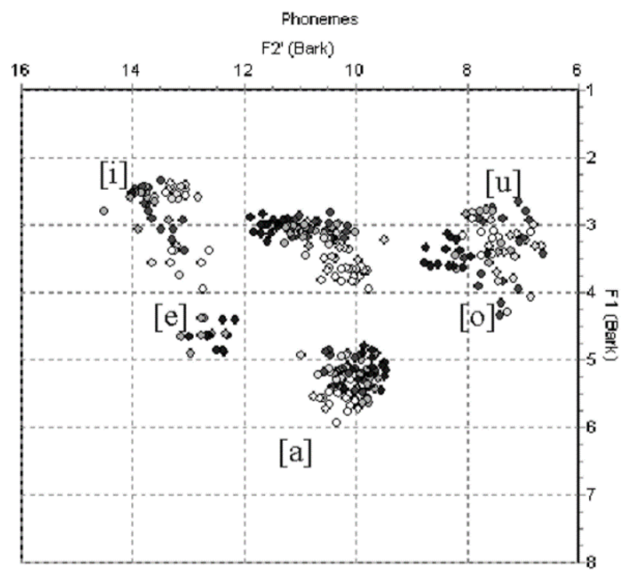

Fig. 7. Emergent vowel system: vowel phonemes are shown as clusters of similar vowels learnt by multiple robotic agents.

Latest achievements in attending to the phonological issues touched in this section as well as many others have resulted from neural networks. An interesting work where the neural network strategy was adopted to examine sound learning is [29]. The author implemented a Generative Adversarial Network (GAN) to learn aspirated and non-aspirated allophones of three English voiceless stop phonemes in an unsupervised fashion. The task was far from trivial as the network was trained on a corpus of continuous speech, i.e., the data was time-series. The basic principle of GAN is twofold: on the one hand, to learn phonetic features in the discriminator part of the network, and on the other hand, using such features, generalize various allophonic pronunciations to a single speech segment which corresponds to a phoneme, producing it in the generator part of the network. The author's experiments demonstrated a high potential of the network to learn allophonic distribution of phonemes. Alongside with other

ISSN 2395-8618

interesting contributions in [29], it is important to note that this is the first work where such distribution is uncovered in an unsupervised manner.

There have been other works done in the field of phonology, still there are many open issues left for future. However, as computer technologies are rapidly developing nowadays, we can expect a substantial progress in modeling the phonological layer of natural language as a complex system, which may lead to a richer and fuller representation and understanding of this phenomenon.

C. *Grammar*

The traditional approach to grammar consists in describing and categorizing language structures as well as formulating rules which generalize to as a broad range of language phenomena as possible; each rule is usually accompanied by a list of exceptions. Complexity theory does not neglect the knowledge accumulated by conventional grammars, but places it in the context of real-life experiences and interactions between humans as language users.

This usage-based approach looks at grammar as a dynamic adaptive network where grammatical constructs emerge through communicative experience of interrelating human agents. Thus, grammar is claimed to be “a cognitive organization of one’s experience with language” [30, p. 711], where the frequency of word usage and word associations plays a decisive role in forming grammar constructions. It turns out that frequencies and therefore probabilities of occurrence and co-occurrence of language constructions promote highly used forms turning them into conventionalized structures [8].

Going further in this direction of thought, one may ask a question of how and in what particular ways frequent usage gives shape to grammatical constructions, what mechanisms operate in the process of conventional pattern emergence. In other words, “what is the best model of constraint generalization”, as Dunn [31, p. 1] puts it, claiming that grammatical constructions are sequences of words that comply with some set of lexical, syntactic, and semantic constraints. So how these constraints emerge, that is, why some particular words and not others are combined to form structures accepted by a language community as grammatical or correct? With an intent to find an answer to this question, Dunn built two grammar generating models: one based on frequency and the other based on associations. Both models were applied to a corpus in order to encode it in a grammar, thus obtaining two grammars. The generalization capacity of the grammars was defined as the number of bits used to encode the corpus, and the quality of generalization was evaluated using the Minimum Description Length principle [32]. The association-based model was the winner, since the grammar it produced used a smaller number of bits.

Dunn’s work studied grammar as a whole; there have also been done studies of aspects or parts of grammar, focusing on their emergence as subsystems of the complex system of grammar. One of such works is [33] where the author studies the emergence of morphosyntactic patterns in complex words based on usage, proposing a theory called Construction Morphology, explaining that its basic idea “is that word formation patterns can be seen as abstract constructional schemas that generalize over sets of existing complex words with a systematic correlation between form and meaning” [33, p. 198]. An example of a constructional schema is derivation of nouns from verbs: *eat* – *eater*, *sing* – *singer*, *walk* – *walker*. All these deverbal nouns are formed according to the common pattern or constructional schema, which can be formalized in a simplified manner as verb + *er* = the verb’s agent. Analyzing many other examples as well as grounding on psycho- and neuro-linguistic evidence, Booij argues that schemas explain the relation between form and meaning of existing complex words and phrases on the one hand and on the other hand serve as a blueprint for creating novel words.

There are a lot of other research questions related to grammar as a complex system, to various facets of grammar and their functions, to acquiring grammar, and Ibbotson *et al*. [34] are interested in finding an answer to one of such questions: how does the complex system of syntax emerge through interactions and dependencies between many simpler units such as words, when used in utterances by small children feedbacked by their caregivers? The authors aim at evaluating how distributional properties of words make their impact on emerging grammatical categories and their relations. Their model is a dynamic, evolving over time network, where weights are assigned to links and which grows incrementally in complexity, thus simulating cognitive processes in human brain. To develop and study the model, the authors took advantage of analytic tools, proposed for doing research on complex adaptive systems in general, especially of those tools developed for community detection in networks. A network community is defined as a set of network units or agents whose connections among themselves are stronger than their connections with the other units in the system. Ibbotson *et al*. [34] detected network communities by link density measures and further studied with respect to their grammatical structure. In the experiments, a corpus of child directed speech was used. When the caregivers pronounced a word for the first time, it was added as a node to the network, when two words were used together as a bigram in an utterance, they were linked by an edge. If such words were pronounced together more than one time, the weight on their link grew in proportion of usage frequency. Fig. 8 [34] shows how the network grows using as input the utterances *John liked Mary*, *John liked Bob*.

After inputting the corpus, Ibbotson *et al*. [34] identified network communities and discovered that they corresponded to some specific grammatical patterns, thus explaining how grammatical constructions can be learnt as self-organizing structures evolving through usage by language speakers. This discovery showed how emergence of grammatical structures is tightly related to real-world contexts, such fact contrasts with the conventional rigid system of highly abstracted rules.

ISSN 2395-8618

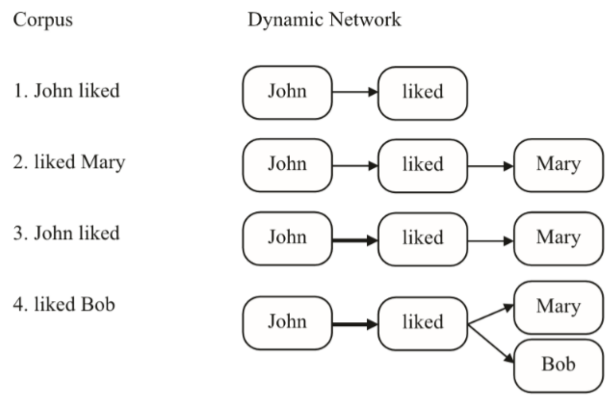

Fig. 8. The dynamic construction of network from speech '*John liked Mary*. *John liked Bob*.'

Another promising research line leading to a more realistic account of grammar than the conventional view is the theory of interactional linguistics. Its detailed review can be found in [35], here we will comment on it briefly. Interactional linguistics suggests to analyze grammar within a conversational setting, monitoring how grammatical constructions are used by speakers during their intercourse. The researchers, who adopted this approach, look for the ways grammar is shaped by human conversation and vice versa: how conversation is shaped by grammar. Therefore, grammar is examined within its real-world context; researchers observe the emergence and self-organization of linguistic structures and conversational traits when participants aim at understanding messages of each other. Moreover, during spoken discourse, language users may employ non-linguistic communication mechanisms such as sighs, sniffs, clicks, whistles, and the like, termed as liminal signs in [36]. In this research, the author stresses that the "language sciences need approaches that can deal just as well with the fluid, hybrid, and liminal aspects of language in interaction as they can deal with its better studied systematic, structural, and compositional aspects" [36, p. 194].

One more point of view on how grammar emerges is formulated by cognitive linguistics. This approach gives attention to a particular element of the context where language develops and operates, namely, general cognitive human abilities such as vision, attention, thinking, and reasoning, including categorization, analogy and entrenchment [37]. Cognitive linguistics is a challenge to generative (rule-based) grammarians who claim that a person is born with specific linguistic capacities, stipulations, and types, and their functioning results in grammatical competence. According to the cognitive approach, such competence emerges and develops through speakers' language experience by exercising general cognitive abilities mentioned above. During communicative practice of speakers, grammatical constructions come up through generalization of phrases that share structure and semantics. An example of such generalization in usage is given in [37] alongside with many other examples: the interrogative frame What AUX NP V? can be abstracted from the concrete questions used in child speech *What's Mommy doing? What's donkey doing? What's Mommy making?*

As we attempted to show in this section, there have been many fruitful trends in studying grammar as a complex dynamic system, functioning and developing within a wide context. In comparison with the traditional view on grammar, usage-based grammar and a complex view on grammar have been quite recent areas of research, where we can wait for many more works to appear.

D. *Lexicon*

Now speaking about grammar in Section 3.3. and lexicon in the present section may produce an impression that these aspects of language are separate but related strata, lexicon items being fillers in grammatical construction slots. In fact, much work has been done in this fashion. However, in real-world language interactions, grammar and lexicon function together in a tight coordination, so a better description of lexicon can be achieved when it is studied together with grammar. Indeed, the complex view on language suggests that all language elements as subsystems can be understood better within the whole complex system of language, which also include speakers as agents and all modalities of constantly changing contexts and environment. It means that the complex system approach can not only describe the existent lexicon in a more realistic way, but also account for its development and emergence of new words. Therefore, although this section is dedicated to lexicon, it will be considered in its relation with grammar and its usage by speakers.

One of many questions arising from such usage-based view on lexicon is how speakers choose words in each specific communicational context, or why a particular word is preferred to other words with similar meaning under certain circumstances.

Answering the above question, applied as a case study to four common Spanish verbs with the semantics of 'becoming', Bybee and Eddington [38] analyzed their 423 usage instances in spoken and written corpora representative enough of speakers' language experience. The verbs were *ponerse* 'to put (reflexive)', *volverse* 'to turn (reflexive)', *quedarse* 'to remain (reflexive)', and *hacerse* 'to make (reflexive)'. The typical frame in which these verbs were found is <an animate subject + verb + adjective>, for example, *El vecino se puso nervioso* ('The neighbor became nervous'). So, the question concerning the speaker's lexical choice would be the following: how does the speaker choose a particular verb out of these four verbs given their synonymy? How does she make the decision to use, for example, *hacerse* in some context, but not any of the other three verbs? Searching for a solution to this issue, Bybee and Eddington [38] fulfilled a comparative analysis based on verb frequencies and semantic similarity, discovering a number of interesting selection patterns, which can be consulted in their article as the space limits do not permit us to contemplate them here.

ISSN 2395-8618

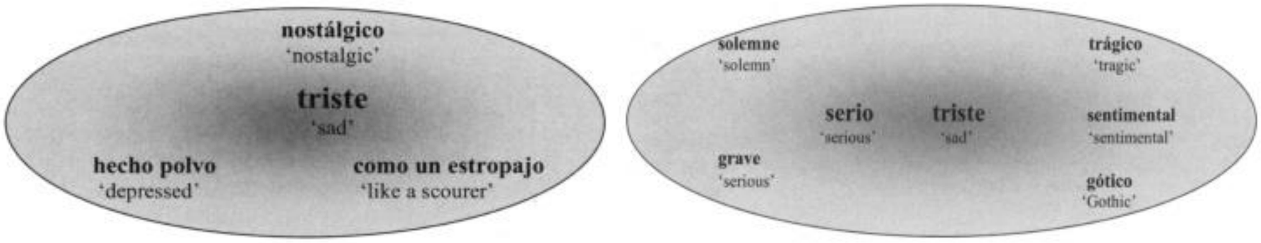

Fig. 9. Clusters centered on *quedarse triste* 'become sad' (left) and on *ponerse serio*/*triste* 'become serious/sad' (right).

Bybee and Eddington were not the first scientists to ask questions with regard to selectional decisions, there had been other research on this topic, mainly based on finding boundaries for each verb usage depending on semantic features. However, such boundaries resulted to be fuzzy, there is overlap among verb usage cases, moreover, it is hard to define semantic features formally and precisely.

A contribution of Bybee and Eddington [38] is their finding that verb choice does not depend on semantic features, but rather on the analogy to speaker's previous linguistic experience with verbal constructions. The authors formalized such choice by constructing a usage-based exemplar model. An exemplar is defined as a mental representation of an actual word or phrase usage; it differs from a prototype, which is a set of semantic features common in all realizations of the word or phrase. In [38], exemplars were retrieved by grouping semantically similar adjectives in all patterns <an animate subject + verb + adjective> found in corpora. In this manner, adjectives and, respectfully, verbs were categorized as clusters with adjectives of highest frequency as central members. As an example, Fig. 9 depicts clusters centered on *quedarse triste* 'become sad' and *ponerse serio*/*triste* 'become serious/sad'. These expressions are synonymous, so it is interesting to observe similarities and differences in adjectives used with the two verbs. All clusters found in the experiments formed a complex network representation of the section of Spanish lexicon that includes the four verbs mentioned previously.

Another important aspect of a complex system is its context. Jones *et al*. [39] studied the impact of diverse and distinctive contexts on mental lexicon organization posing the question: how do speakers arrive at similar organization of their mental lexicon which allows them to communicate and understand each other while having different individual linguistic experience? To answer this question, the authors used data on word recognition, semantic and episodic memory as general cognitive ability, and information retrieval for their analysis. It was found that not only word frequency is important in forming mental lexicon, but also the number of diverse contexts the word is encountered in. Words are learnt more rapidly and better if they are experienced by a learner as a constituent of other words or phrases, in a wide spectrum of combinations with other lexical items and in texts belonging to different domains. Such findings are important not only for enhancing language learning, development, and education, but also for developing more robust and precise mathematical and computational models and tools for more accurate text understanding and generation.

Speaking of lexicon, our language competence includes not only knowledge of individual words but also multiword expressions, various word combinations, and utterances. With such respect, Lieven *et al*. [40] studied how two-year-old children learn new multiword utterances relating them to known utterances. The research technique employed was the traceback method, in which a corpus of recorded utterances produced by four two-year-old children, two female and two male, was used and observations were made on the degree of relation between an utterance and previously produced utterances, which included it as a whole or as some segment of it. The analysis of observations revealed that what children said at the moment was closely related with the preceding speech. This discovery is in a complete accord with other research within usage-based approach; it was also confirmed that what children learn are multiword combinations, not just single words. Concerning grammatical schemas, the authors found that the slot for referents was most easily acquired by children when they compared it to other slots such as location, process, attribute, utterance or direction. This relation between speech and previously produced utterances showed that through such relations children develop schemas and their slots, starting with referring expressions and then, with further language experience, they develop more abstract slots in conventionalized constructions.

Usage-based approach to lexicon, which is the focus of this section, has been confirmed by many researchers, among them are Bruns *et al.* [41], who applied usage-based approach to familiar collocations (*I don't know*, *it's alright*, etc.) in non-fluent aphasia. The authors studied the frequency of producing such collocations by speakers with Broca's aphasia and their conversation partners in both dyadic and non-dyadic speech. Also, as a case study, the authors examined how one particular collocation *I don't know* was used in conversations. Their study showed that speakers with aphasia used much fewer word combinations then their conversational partners. However, the words within collocations were more strongly associated in the aphasic speech. Such stronger association confirm the usage-based view on language processing.

ISSN 2395-8618

Finally, the usage-based approach has been shown productive in studying artificial languages, for instance, sign language. Lepic [42] investigated how general-domain cognitive processes of chunking, entrenchment, and routinization were involved in lexicalization in American sign language. Particularly, the author studied multiword expressions, fingerspelled words, and morphologically complex signs. Analyzing sign language speech according to the usage-based position, the author demonstrated that conventionalized, or lexicalized, fixed multiword expressions, e.g. *interpreter bring-on* and *take-to hospital*, come from the ability to recognize and generalize relations between form and meaning while speakers exercise such ability in their individual experiences with language. Looking at fingerspelled words, e.g., *power* and *pen*, the author observed gradual and continuous processes of entrenchment and reduction, as these words were repeated in the same communication context. Concerning morphologically complex signs, e.g. *rent* and *hash-things-out*, Lepic found that they obtain their typical meaning by means of their contexts of use. In such contexts, complex signs gradually become related to additional senses by means of metonymy, diverging significantly from the original meaning of complex signs components.

The complex approach to lexicon, which claims that lexical constructions emerge through linguistic experience of speakers, has been successful in many other scientific works. Gradually, this approach has been forming a solid framework for analyzing lexicon, proposing effective formal language models, explaining development and function of lexical items. Such models further improve language analysis and processing, aiding in many aspects of human activities.

E. First Language Acquisition and Development

Another interesting but hard question to answer is how language is acquired, how it is transmitted from adult to infant, and how a child internalizes language as a global system, making its own individually distinctive version of this system.

In the course of time, there have been made several attempts to create theories explaining how a child develops her unique fashion of language. One of such theories is nativism, suggesting that the potential to acquire language is innate in the human nature and the ability to develop diverse language structures is genetically predetermined. The most prominent nativism advocate is Noam Chomsky [43]. In his opinion, specific neurophysiological attributes of human brain could be the only rationale accounting for the child amazing capacity of mastering such a sophisticated aggregate as language.

As opposed to nativism, behaviorism argued that linguistic forms like other stimuli are learnt through practice and repetition. This view was proposed by Skinner [44] and strongly criticized by Chomsky who claimed that the behaviorist approach cannot explain how the child develops adequate usage of irregular wordforms. The matter is that the process of acquiring such forms usage is not easily modeled with repetitive behavioral patterns: the child first utters these words correctly, however, in a while she starts to makes errors often giving no attention to adult corrections, eventually, this period terminates and the child attains to the proper usage.

Similar to behaviorism, empiricism (see, for instance, the classical work in [45]) criticizes nativism claiming that to acquire language, the child does not use genetically conditioned language-specific potential permitting to develop complex language forms, but instead, biologically inherited general cognitive abilities as in assimilating any kind of knowledge during the child's early lifetime. Researchers within this trend, ground their arguments in latest finding of neuroscience concerning the way children process information in different periods of their life [46].

Different from the previously mentioned theories, the emergentist approach to language acquisition [47] suggests that all factors—genetic, behavioral, and environmental—interplay in the learning process, leading to the continuous emergence and improvement of language proficiency in the child. Emergentism, allying with the complexity perspective, argues that language mastery results from local interactions between many factors and aspects embedded in the child-caregiver-real world complex system.

As a complex adaptive and dynamic system, language involves form, user, and usage [8]. Their relationships are difficult to study: it is practically impossible to make long-term observations of a child and her environment to investigate how she grows in language competence. Besides, learning takes its particular course in every child, so we have to observe continuously a representative sample of the children population which is also impracticable. These restrictions make computational models and simulation a valuable research tool on this complicated issue.

McCauley and Christiansen [48] suggested that language learning in fact is language use; they developed a computational model based on this assumption in a cross-linguistic ambient. Other underlying principles of the model design were incremental acquisition, memory constraints, and the child's ability to comprehend and produce language via discerning multiword units (chunks) rather than individual words, thus acquiring patterns of language structure. Incremental acquisition was modeled as a probabilistic word-by-word analysis in the phase of comprehension of adult speech directed to the child, and as chunk-by-chunk generation in the phase of language production based on immediate input rather on stored knowledge, thus modeling memory constraints, i.e., the word probabilities were computed over adjacent chunks. Multiword unit discovery was simulated by means of part-of-speech tagging and shallow parsing. Simulating child language acquisition, the model was learning language structure and producing utterances which were then compared to real children's real utterances in the corpus of English speech from CHILDES database [49] reaching the mean F-score of 74.5, thus successfully demonstrating the child's early linguistic comportment.

ISSN 2395-8618

Another side of language acquisition complexity was exposed by Ororbia *et al*. [50] who argued that early language learning takes place in multi-modal environment. The argument was underpinned by two experiments: first, learning was modeled by a neural network and, second, by the same network in conjunction with visual context. The model in the second experiment yielded better results than that in the first one, demonstrating that augmenting the predictive neural language model with images illustrating the sentences being learnt improved the model's predictive capacity. This result is supportive of the situated cognition theory [51], [52] whose claim is that language development goes in hand with learner's interaction with physical reality. Also, the results in [50] emphasized the significance of non-symbolic semantic representation combined with linguistic structures for language acquisition on the one hand, and on the other hand, served as an evidence of one of the principles of complex systems: the environment or context is to be taken into account as a part of the language system and not as its surroundings only, therefore, language cannot be abstracted from context without any loss of information.

In research community, there has been interest not only in modeling acquisition of the complex system of language as a whole, so to say, from the bird's eye view, but also in modeling the acquisition of language segments and subsystems. If we "take a loop" and give a closer look at language, we would observe further striking complexity in any language element we approach. One of such elements of the language development model is acquisition of verb inflections to express person, number, or gender. The computational model of verbal conjugation acquisition developed by Engelmann *et al*. [53] and applied to Finnish and Polish was built to test if rote storage and phonological analogy, two principles grounded in children's real-life linguistic activity, were sufficient for achieving verbal inflection mastery in contrast to traditional grammar rules. This assumption was verified in the experiments where the system reached almost adult-level expertise of six person/number forms in the verbal paradigm, chosen for the tests. The authors concluded that rote storage and phonological analogy were responsible not only for successful verbal form acquisition, but for attaining the knowledge of inflectional morphology in its totality and even for learning the whole body of language.

There are other works on complex modeling of morphology acquisition, different from [53]. However, the space for this article does not allow us to speak of them in detail, we will only make a brief mention of [54], where the researchers suggested that an end-to-end machine learning model based on a semantic vector space, linear transformation, and proportional analogy is potent enough for efficient lexicon management without the need of morphological analysis of words into phonemes and morphemes. In their experiments, the authors found that incremental implicit learning formalized by the proposed model, built as a two-layer linear network, explains quite well the fundamental aspects of understanding and producing words.

With the above survey of some recent papers on first language acquisition within the complex system modeling approach, we conclude our consideration of language as a complex system. We hope that all research works we reviewed in this article expose a high potential that the complex view on language, so different from traditional linguistic descriptions, has in exploring this multi-faceted and multi-dimensional phenomenon which is natural language.

IV. Conclusion

Human language possesses many features in common with complex systems: self-organization, emergence, dynamic and adaptive nature, fuzziness. This fact opens a new horizon in studying language since it can be viewed and investigated as a complex adaptive system. Traditional linguistics—also including methods in computational linguistics and natural language processing which develop models along the lines of the traditional approach—tries to avoid complexity by segmenting language into various layers (phonetics, phonology, lexicon, morphology, syntax, etc.) and studying or modeling them separately. In contrast, the complex view on language suggests to study language as a system of multiple agents-speakers involved in language interactions and relationships, where language structures emerge and develop. Research papers reviewed in this article present many significant discoveries in language studies made by using computational models and simulations within the complex system framework. We are convinced that in future we will witness many more contributions produced by the complex system approach to language.

Acknowledgments

The work was done under partial support of Mexican Government: SNI, COFAA-IPN, BEIFI-IPN, and SIP-IPN grant 20201948.

References

[1] J. Y. L. Forrest, J. Nicholls, K. Schimmel, and S. Liu, "Basics of systems science," in *Managerial Decision Making*. Springer, Cham, 2020, pp. 25–40.

[2] À. Massip-Bonet, "Language as a complex adaptive system: Towards an integrative linguistics," in *Complexity Perspectives on Language, Communication and Society*. Berlin: Springer, 2013, pp. 35–60.

[3] M. S. M. Saat, S. K. Nguang, and A. Nasiri, *Analysis and Synthesis of Polynomial Discrete-Time Systems: An SOS Approach*. Butterworth-Heinemann, 2017.

[4] L. Sabatucci, V. Seidita, and M. Cossentino, "The four types of self-adaptive systems: A metamodel," in *Int. Conf. Intelligent Interactive Multimedia Systems and Services*. Springer, Cham, May 2018, pp. 440–450.

[5] S. Thurner, R. Hanel, and P. Klimek, *Introduction to the Theory of Complex Systems*. Oxford University Press, 2018.

[6] L. A. Amaral, and J. M. Ottino, "Complex networks," *The European Physical J. B*, vol. 38, no. 2, pp. 147–162, 2004.

ISSN 2395-8618

[7] F. Munné, “The fuzzy complexity of language,” in *Complexity perspectives on language, communication and society*., Berlin: Springer, 2013, pp. 175–196.

[8] “Five Graces Group”, C. Beckner, R. Blythe, J. Bybee, M. H. Christiansen, W. Croft, N. C. Ellis, J. Holland, J. Ke, D. Larsen-Freeman, and T. Schoenemann, “Language is a complex adaptive system: Position paper,” *Language Learning*, vol. 59, pp. 1–26, 2009.

[9] L. Steels, “Language as a complex adaptive system,” in *Int. Conf. Parallel Problem Solving from Nature*. Berlin: Springer, 2000, pp. 17–26.

[10] B. Coffman, J. Smethurst, M. Kaufman, and L. Morris, “A brief introduction to complex adaptive systems,” *Discovery Workshop by InnovationLabs*, Chicago, IL, August 27–29, 2007.

[11] I. Ibarretxe-Antuñano, and J. Valenzuela, “Language as a complex system: A view from cognitive linguistics,” in *Language as a Complex System: Interdisciplinary Challenges*, G. Bel-Enguix, and M. Dolores Jiménez López, Eds. 2010, pp. 3–38.

[12] S. Reddy, D. McCarthy, and S. Manandhar, “An empirical study on compositionality in compound nouns,” in *Proc. 5th Int. Joint Conf. Natural Language Processing*, 2011, pp. 210–218.

[13] A. Gegov, *Fuzzy Networks for Complex Systems*. Berlin: Springer, 2010.

[14] L. Cameron, and D. Larsen-Freeman, “Complex systems and applied linguistics,” *Int. J. of Applied Linguistics*, vol. 17, no. 2, pp. 226–239, 2007.

[15] P. Y. Oudeyer, G. Kachergis, and W. Schueller, “Computational and robotic models of early language development: A review, ” *arXiv preprint arXiv:1903.10246*, 2019.

[16] N. Cambier, R. Miletitch, V. Frémont, M. Dorigo, E. Ferrante, and V. Trianni, “Language evolution in swarm robotics: A perspective,” *Frontiers in Robotics and AI*, vol. 7, no. 12, 2020.

[17] R. R. Liu, W. X. Wang, Y. C. Lai, G. Chen, and B. H. Wang, “Optimal convergence in naming game with geography-based negotiation on small-world networks,” *Physics Letters A*, vol. 375, no. 3, pp. 363–367, 2011.

[18] A. Baronchelli, “A gentle introduction to the minimal Naming Game,” *Belgian J. of Linguistics*, vol. 30, no. 1, pp. 171–192, 2016.

[19] N. Cambier, V. Frémont, V. Trianni, and E. Ferrante, “Embodied evolution of self-organised aggregation by cultural propagation,” in *Proc. Int. Conf. Swarm Intelligence*. Springer, Cham, October 2018.

[20] R. Miletitch, A. Reina, M. Dorigo, and V. Trianni, “Emergent naming of resources in a foraging robot swarm,” *arXiv preprint arXiv:1910.02274*, 2019.

[21] J. Vera, F. Urbina, and W. Palma, “Phase transitions in a decentralized graph-based approach to human language,”*arXiv preprint arXiv:2003.02639*, 2020.

[22] B. de Boer, “Emergence of vowel systems through self-organisation,” *AI Communications*, vol. 13, no. 1, pp. 27–39, 2000.

[23] B. Vaux, and B. Samuels, “Explaining vowel systems: dispersion theory vs natural selection,” *The Linguistic Review*, vol. 32, no. 3, pp. 573–599, 2015.

[24] C. Moulin-Frier, J. Diard, J. L. Schwartz, and P. Bessière, “COSMO (“Communicating about Objects using Sensory–Motor Operations”): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems,” *J. of Phonetics*, vol. 53, pp. 5–41, 2015.

[25] P. Y. Oudeyer, and L. B. Smith, “How evolution may work through curiosity-driven developmental process,” *Topics in Cognitive Science*, vol. 8, no. 2, pp. 492–502, 2016.

[26] B. de Boer, “Emergence of sound systems through self-organisation,” in *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pp. 177–198, 2000.

[27] D. J. *Livingstone*, *Computer Models of the Evolution of Language and Languages*. Doctoral dissertation, University of Paisley, 2003.

[28] J. Harrington, F. Kleber, U. Reubold, F. Schiel, M. Stevens, and P. Assmann, “The phonetic basis of the origin and spread of sound change,” in *The Routledge Handbook of Phonetics*. 2019.

[29] G. Beguš, “Modeling unsupervised phonetic and phonological learning in Generative Adversarial Phonology,” in *Proc. Society for Computation in Linguistics*, vol. 3, no. 1, pp. 138–148, 2020.

[30] J. Bybee, “From usage to grammar: The mind's response to repetition,” *Language*, pp. 711–733, 2006.

[31] J. Dunn, “Frequency vs. association for constraint selection in usage-based construction grammar,” *arXiv preprint arXiv:1904.05529*, 2019.

[32] P. D. Grünwald, and A. Grunwald, *The Minimum Description Length Principle*. MIT press, 2007.

[33] G. Booij, “Language use and the architecture of grammar: A Construction Morphology perspective,” *Suvremena lingvistika*, vol. 40, no. 78, pp. 193–212, 2014.

[34] P. Ibbotson, V. Salnikov, and R. Walker, “A dynamic network analysis of emergent grammar,” *First Language*, vol. 39, no. 6, pp. 652–680, 2019.

[35] E. Couper-Kuhlen, and M. Selting, *Interactional Linguistics: Studying Language in Social Interaction*. Cambridge University Press, 2017.

[36] M. Dingemanse, “Between sound and speech: Liminal signs in interaction,” *Research on Language and Social Interaction*, vol. 53, no. 1, pp. 188–196, 2020.

[37] H. Diessel, “Usage-based construction grammar”, in *Cognitive Linguistics-A Survey of Linguistic Subfields*, E. Dąbrowska and D. Divjak, Eds. Berlin: Mouton de Gruyter, 2019, pp. 50–80.

[38] J. Bybee, and D. Eddington, “A usage-based approach to Spanish verbs of ‘becoming’,” *Language*, pp. 323–355, 2006

[39] M. N. Jones, M. Dye, and B. T. Johns, “Context as an organizing principle of the lexicon,” in *Psychology of Learning and Motivation*, vol. 67. Academic Press, 2017.

[40] E. Lieven, D. Salomo, and M. Tomasello, “Two-year-old children's production of multiword utterances: A usage-based analysis,” *Cognitive Linguistics*, vol. 20, no. 3, pp. 481–507, 2009.

[41] C. Bruns, R. Varley, V. C. Zimmerer, M. Carragher, G. Brekelmans, and S. Beeke, “I don’t know”: a usage-based approach to familiar collocations in non-fluent aphasia,” *Aphasiology*, vol. 33, no. 2, pp. 140–162, 2019.

[42] R. Lepic, “A usage-based alternative to “lexicalization” in sign language linguistics,” *Glossa: A Journal of General Linguistics*, vol. 4, no. 1, 2019.

[43] N. Chomsky, *Knowledge of Language: Its Nature, Origin, and Use*. Greenwood Publishing Group, 1986.

[44] B. F. Skinner, *Verbal behavior*. New York: Appleton-Century-Crofts, 1957).

[45] M. Ramscar, and N. Gitcho, “Developmental change and the nature of learning in childhood,” *Trends in Cognitive Sciences*, vol. 11, no. 7, pp. 274–279, 2007.

[46] J. L. Mueller, A. D. Friederici, and C. Männel, “Developmental changes in automatic rule-learning mechanisms across early childhood,” *Developmental Science*, vol. 22, no. 1, 2019.

[47] B. MacWhinney, “Emergentism,” in *Cognitive Linguistics-Key Topics*, E. Dabrowska, and D. Divjak, Eds. Berlin: Mouton-DeGruyter, 2019, pp. 275–294.

[48] S. M. McCauley and M. H. Christiansen, “Language learning as language use: A cross-linguistic model of child language development,” *Psychological Review*, vol. 126, no. 1, 2019.

[49] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk*, vol. I: *Transcription Format and Programs*, vol. II: *The Database*. 2000.

[50] A. Ororbia, A. Mali, M. Kelly, and D. Reitter, “Like a baby: Visually situated neural language acquisition.” *Proc. 57th Annu. Meeting of the Association for Computational Linguistics*, July 2019, pp. 5127–5136.

[51] J. G. Greeno, and J. L. Moore, “Situativity and symbols: Response to Vera and Simon,” *Cognitive Science*, vol. 17, no. 1, pp. 49–59, 1993.

[52] K. A. Overmann, and L. Malafouris, “Situated cognition,” in *The International Encyclopedia of Anthropology*, 2018, pp. 1–8.

[53] F. Engelmann, S. Granlund, J. Kolak, M. Szreder, B. Ambridge, J. Pine, A. Theakston, and E. Lieven, “How the input shapes the acquisition of verb morphology: Elicited production and computational modelling in two highly inflected languages,” *Cognitive Psychology*, vol. 110, pp. 30–69, 2019.

[54] R. H. Baayen, Y. Y. Chuang, E. Shafaei-Bajestan, and J. P. Blevins, “The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning,” *Complexity*, 2019.

ISSN 2395-8618

Non-Parametric *k*-Sample Tests for Comparing Forecasting Models

Dmitriy A. Klyushin

***Abstract*—Business intelligence is impossible without practical tools for assessing the quality of forecasts and comparing forecast models. A naive approach to comparing models by comparing predicted values with observable ones ignores the probabilistic nature of errors. Many models with varying degrees of accuracy are statistically equivalent. Hence, before ranking the models for accuracy, it is necessary to test the statistical hypothesis about the homogeneity of the distributions of their errors. In the presence of several models, the problems of their pairwise and group comparison arise. This chapter provides an overview of the non-parametric tests used in business analysis for pairwise and group comparisons and describes a new non-parametric statistical test that is highly reliable, sensitive, and specific. This test is based on assessing the deviation of the observed relative frequency of an event from its a priori known probability. The prior probability is given by Hill's assumption, and the confidence intervals for the binomial success rate in the Bernoulli scheme are used to estimate its difference from the observed relative frequency. The paper presents the results of computer modeling and comparison of the proposed test with the alternative Kruskal-Wallis test and the Friedman test on artificial and real examples.**

***Index Terms*—Business process modeling, error analysis, modeling and prediction, non-parametric statistics.**

1. Introduction

BUSINESS intelligence combines technologies and methods for the collection, analysis, and prediction of business information. It may be represented as a hierarchical scheme consisting of four layers.

At the foundation layer, business data are collected and analyzed by descriptive data analytics methods. Here we answer the question: "What happened?" The examples of these data are the signals from the sensors or exchange rate. Analysis of these data allows detecting the symptoms of failure of devices and identifying change-point of time series describing exchange rate.

At the second level, we analyze diagnostic information to answer the question: "Why did it happen?" This problem is harder than the first-level problem because it requests to detect and recognize patterns. Using statistical methods, we establish intrinsic relations between data and discover reasons for events. For example, we may find the correlation between events and exchange rate changes.

At the third level, we try to answer the question: "What will happen?" Using data obtained at the first two levels and methods of predictive analytics, we forecast events that may occur in the future to determine what consequences can follow. The fundamental complexity of forecasting requires the use of very complex tools of mathematical statistics, machine learning, data mining, and simulation. For example, predictive analytics can be used for forecasting the future stock price in the stock market or determine the optimal time for repairs to prevent breakdowns of process equipment.

At the fourth level using methods of prescriptive analytics, we answer the question: "What to do?" for making the optimal decision. This level demands very complex mathematical methods of optimization. For example, we can find an optimal moment for buying or selling stocks or the beginning of a repair.

The subject of this chapter is the non-parametric approach to the estimation of predictive models. Predictive analytics opens up broad perspectives in commercial, financial, and industrial applications. For example, it allows you to optimize recommendation systems in online stores, as well as segment your customer base to improve the effectiveness of direct marketing and targeted advertising. In banking and insurance, predictive analytics has become a necessary tool in assessing an applicant's creditworthiness and detecting fraud. In the financial sector, predictive analytics can improve investment performance and optimize risk management.

In industrial applications, predictive analytics allows solving the problems of forecasting product quality, optimization of repair schedules, recognition of abnormal symptoms, and many other tasks. Sophisticated high-tech enterprises use automatic control of operational processes. To optimize this control, it is necessary to organize an automated collection and analysis of indicators in order to predict resource consumption and product output. By continuously analyzing large volumes of production data, you can prevent line failures and shutdowns, minimize costs and maximize product quality.

The variety of numerical data found in all these calculations does not allow making educated assumptions about the form of their distribution; therefore, the most suitable methods for analysis in these areas are non-parametric methods, which do not imply a certain form of data distribution. These methods include non-parametric hypothesis testing methods.

Manuscript received on October 17, 2020, accepted for publication on November 24, 2020, published on December 30, 2020.

The author is with the Taras Shevchenko National University of Kyiv, 03680, Kyiv, prospect akademika Glushkova 4D, Ukraine, (e-mail: dokmed5@gmail.com).

ISSN 2395-8618

The purpose of this chapter is to survey non-parametric methods for two- and *k*-sample used in predictive analytics and propose a new approach to test homogeneity of two- and *k*-samples for estimation forecasting model effectiveness.

The paper is organized in the following way. Section 1 describes the purpose of the chapter. Section 2 describes widely used non-parametric tests for homogeneity. Subsection 2.1 contains a short survey of two-sample tests, as subsection 2.2 is devoted to *k*-sample tests. Section 3 describes the application of non-parametric tests in business intelligence. Section 4 describes the Klyushin–Petunin non-parametric test for two-sample homogeneity without ties. The two-sample case without ties is considered in Subsection 4.1, and the version with ties is described in Subsection 4.2. The extension of the above-mentioned tests on the *k*-sample case is considered in Subsection 4.3. Section 5 describes the results of numerical experiments. Section 6 contains conclusions and directions for future work.

2. Non-Parametric Homogeneity Tests

Non-parametric homogeneity tests for two and *k*-samples were studied in many scientific papers.

2.1 Two-Sample Tests

Let samples $x=(x_1,x_2,...,x_n)$ and $y=(y_1,y_2,...,y_m)$ be drawn from populations G_1 and G_2 which follow absolutely continuous distribution functions F_1 and F_2. The null hypothesis H_0 states that samples are homogeneous, i.e., follow the same distribution function, $F_1=F_2$, and the alternative hypothesis H_1 states the opposite ($F_1 \neq F_2$). The tests for homogeneity of two samples are subdivided into permutation tests, rank tests, randomization tests, and distance tests. Also, these tests form the group of universal tests that are valid against any pair of alternatives (e.g., the Kolmogorov–Smirnov test [1], the Kuiper test [2]), and tests that are valid against pairs of different alternatives of a particular class (Dickson [3], Wald and Wolfowitz [4], Mathisen [5], Wilcoxon [6] Mann–Whitney [6], Wilks [8], etc.). Also, the tests may be classified as pure non-parametric and conditionally non-parametric ones. The pure non-parametric tests do not depend on the assumptions of the distribution function (e.g., all the tests mentioned above). The conditionally nonparametric tests put some assumptions of distributions (Pitman [9], Lehmann [10], Rosenblatt [11], Dwass [12], Fisz [13], Barnard [14], Birnbaum [15], Jockel [16], Allen [17], Efron and Tibshirani [18], Dufour and Farhat [19]).

2.2 k-Sample Tests

Let samples $x^{(k)}=\left(x_1^{(k)},x_2^{(k)},...,x_n^{(k)}\right)\in G_k$ be drawn from populations $G_k, k=\overline{1,K}$ following distributions F_k. The null hypothesis H_0 states that samples are homogeneous, i.e. $F_1=F_2=...=F_K$, and the alternative hypothesis H_1 states the opposite, i.e., there are such i and j that $F_i \neq F_j$, $i \neq j$, $i, j = 1,\dots, K$.

k-sample tests based on the distance between empirical distributions functions include the Kolmogorov–Smirnov test [1], [2], the Cramer–Von-Mises tests [20], the Anderson–Darling test [21], etc. These tests use different distances between empirical functions, e.g. L_1 (the Manhattan distance), L_2 (the root-mean-square deviation), and L_∞ (the Chebyshev distance). The sensitivity of these tests depends on the sample size; the larger, the better. Obviously, this condition can rarely be satisfied in real applications.

k-sample tests based on the likelihood ratio are differentiated depending on a function used in the ratio: empirical distributions functions (Zhang test [22], dynamic slicing [23], energy distance [24], etc.), empirical characteristics functions (Sźekely and Rizzo [25], [26], Fernández et al. [27], Hušková and Meintanis [28], etc.), and kernel density estimations (Uña-Alvarez et al. [29], [30]).

Despite the strengths of these tests, their power and consistency depend on some parameters and assumptions that sometimes are hard to satisfy in practice. Also, to compute critical values of these tests, it is often necessary to make permutations that increase a computational burden [31].

Non-parametric rank-based tests are more powerful alternatives to the above-mentioned ones. They do not require an assumption on the distribution function (e.g., normality) to determine the p-value of the test, can be very powerful in the cases when other tests would fail and are robust to outliers [32]. That is why we shall use as benchmarks the Kruskal-Wallis test and The Friedman test [33–37] implemented in all popular statistical computing packages.

3. Non-parametric used in forecasting models estimation

Analyzing time series, the analyst has a wide range of predictive model options. The choice of the best model depends on many factors, primarily on its accuracy. Meanwhile, the accuracy of the model is a random variable that has a certain distribution. Two models having different accuracy can be compared with each other only by testing the statistical hypothesis that the error of one of them is stochastically less than the error of the other. If the distributions of errors are identical, then the models should be considered equivalent, despite the fact that they have different indicators of accuracy. Usually, the standard error (MSE), the mean absolute percentage error (MAPE) and the mean absolute deviation (MAD) are used for estimating the accuracy of forecasting models. Thus, analyzing the quality of forecasting models, we must test a hypothesis that samples of forecast errors are homogeneous as a rule model errors are supposed to be stationary and unbiased.

ISSN 2395-8618

The testing equality of forecasting model accuracy is clearly described in [38]. Let $M_j, j=1,...,m$ be forecasting models producing predictions $x_i^{(j)}$ of data sequence $x_i, i=1,...,n$ and $\varepsilon_i^{(j)}, i=1,...,n; j=1,...,m$ be errors of the model M_j following distribution F_i. Consider a loss function $g\left(\varepsilon_i^j\right)$ describing model accuracy, e.g., standard deviation. The hypothesis about similar accuracy of models M_k and M_l is equivalent to the hypothesis that the mathematical expectation of $d_i^{(k,l)} = g\left(\varepsilon_i^{(k)}\right) - g\left(\varepsilon_i^{(l)}\right)$ is zero. Thus, $E\left(g\left(\varepsilon_i^{(k)}\right)\right) = E\left(g\left(\varepsilon_i^{(l)}\right)\right)$. If a loss function is the standard deviation $\varepsilon_i^{(k)} - \varepsilon_i^{(l)}$, the problem is reduced to testing the hypothesis $E\left(\varepsilon_i^{(k)}\right) = E\left(\varepsilon_i^{(l)}\right)$. Usually, testing equality of forecasting models accuracy is reduced to the testing of this hypothesis [38]. However, the hypothesis that accuracy measures follow the same distribution is more general. Thus, to test homogeneity, it is necessary to test the null hypothesis not only in partial case $E\left(\varepsilon_i^{(k)}\right) = E\left(\varepsilon_i^{(l)}\right)$ but in the general case $F_k = F_l$.

4. Klyushin–Petunin Test for Homogeneity

The Klyushin–Petunin test has two versions: for samples without ties and for samples containing ties.

4.1 Two-sample case without ties

Let samples $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_m)$ be drawn from populations G_1 and G_2 which follows absolutely continuous distribution functions F_1 and F_2. According to Hill's assumption $A_{(n)}$ [39], if random values $x_1, x_2, ..., x_n$ are exchangeable and belong to absolutely continuous distribution, then

$$P\left(x_{n+1} \in \left(x_{(i)}, x_{(j)}\right)\right) = p_{ij} = \frac{j-i}{n+1}, \quad j > i, \tag{1}$$

where x_{n+1} is a random value following the same distribution as $x_1, x_2, ..., x_n$, and $x_{(i)}$ is the *i*-th value of the ordered sample. On the ground of this fact, non-parametric tests for homogeneity of samples without ties [40] and with ties [41] were developed.

Let $A_{ij}^{(k)}$ be an event when the elements of y are greater than $x_{(i)}$ and less than $x_{(j)}$, and h_{ij} is its relative frequency. Knowing the a priori probability (1) and the observed relative frequency h_{ij}, we can estimate how much h_{ij} deviates from p_{ij} using Wilson confidence intervals for binomial proportions:

$$\begin{aligned} p_{ij}^{(1)} &= \frac{h_{ij}m + z^2/2 - z\sqrt{h_{ij}(1-h_{ij})m + z^2/4}}{m+z^2}, \\ p_{ij}^{(2)} &= \frac{h_{ij}m + z^2/2 + z\sqrt{h_{ij}(1-h_{ij})m + z^2/4}}{m+z^2}. \end{aligned} \tag{2}$$

Then, we compute the lower and upper bounds of the confidence interval $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ with the parameter *z* depending on the desired significance level. If *z* is equal to 3, then the significance level of $I_{ij}^{(n,m)}$ is less than 0.05 [40]. In this case, (2) is a so-called *3s-rule interval*. This rule is based on the Petunin-Vysochanskii inequality [42, 43], stating that if *X* be a random variable with unimodal distribution, mean μ and finite, non-zero variance σ^2, than $P\left(|X-\mu| \geq \lambda\sigma\right) \leq \frac{4}{9\lambda^2}$ for any $\lambda > \sqrt{\frac{8}{3}}$. Replacing the mean μ and the variance σ by the sample mean $\bar{x}$ and the standard deviation *s* and setting λ equal to 3, we obtain the *3s-rule*: $P\left(|X-\bar{x}| \geq 3s\right) \leq \frac{4}{81} \approx 0.494 < 0.05$.

Denote $N = (n-1)n/2$ and $L = \#\left\{p_{ij} = \frac{j-i}{n+1} \in I_{ij}^{(n,m)}\right\}$. Then, $h = L/N$ is a homogeneity measure of samples *x* and *y*, which we shall call *p-statistics*, and a Wilson confidence interval (2) where we set *h* instead of h_{ij} and *N* instead of *m* is the confidence interval for the probability $p\left(\frac{j-i}{n+1} \in I_{ij}^{(n,m)}\right)$.

The scheme of events $A_{ij}^{(k)}$ when the null hypothesis is true is called a generalized Bernoulli scheme [44], [45]. If the null hypothesis is false, this scheme is called a modified Bernoulli scheme. In the general case, when the null hypothesis can be either true or false, this scheme is called Matveichuk–Petunin scheme [46]. Thus, the test for the null hypothesis $F_1 = F_2$ with a significance level, which is less than 0.05, maybe formulated in the following way: construct the Wilson confidence interval $I_n = (p_1, p_2)$ for p-statistics; if I_n contains 0.95, the null hypothesis is accepted, else the null hypothesis is rejected.

4.2 Two-Sample Case with Ties

In practice, due to imprecise measurement, samples often contain ties, i.e., repeated elements. A sample x containing absolutely precise elements we shall call hypothetical. The sample $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ containing approximations of the hypothetical elements of the hypothetical x we shall call empirical. Population $\tilde{G}$ we shall call an empirical population corresponding to the hypothetical population G. Let $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ and $\tilde{x}_{(1)} \leq \tilde{x}_{(2)} \leq ... \leq \tilde{x}_{(n)}$ be variational series constructed by hypothetical and empirical samples.

For a sample value x^* that is drawn from G independently from x the Hill assumption holds [39]:

$$p\left(x^* \in \left[x_{(k)}, x_{(k+1)}\right)\right) = \frac{1}{n+1}, \tag{3}$$

where $k = 0, 1, ..., n$, $x_{(0)} = -\infty$, and $x_{(n+1)} = \infty$. Hence,

ISSN 2395-8618

$$p_{ij} = p\left(A_{ij}\right) = p\left(\tilde{x}^* \in \left[\tilde{x}_{(i)}, \tilde{x}_{(j)}\right)\right) \approx \approx \gamma_i + \gamma_{i+1} + ... + \gamma_{j-1} + \frac{j-i}{n+1}, \quad (4)$$

where $\gamma_k = \frac{t\left(\tilde{x}_{(k)}\right)-1}{n+1}$, $t\left(\tilde{x}_{(l)}\right)$ is the number of repetitions of $\tilde{x}_{(l)}$, $A_{ij} = \left\{\tilde{x}^* \in \left[\tilde{x}_{(i)}, \tilde{x}_{(j)}\right]\right\}$. If $\tilde{x} = \left(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\right)$ does not contain ties, then (4) transforms to (1).

Let the null hypothesis states that hypothetical continuous distribution functions F_1 and F_2 of hypothetical populations G_1 and G_2 are identical. Consider empirical samples $\tilde{x} = \left(\tilde{x}_1,...,\tilde{x}_n\right) \in \tilde{G}_1$ and $\tilde{y} = \left(\tilde{y}_1,...,\tilde{y}_m\right) \in \tilde{G}_2$, where $\tilde{G}_1$ and $\tilde{G}_2$ are the empirical populations corresponding to hypothetical populations G_1 and G_2. Suppose that $F_1 = F_2$ and denote $A_{ij}^{(k)} = \left\{\tilde{x}_k \in \left(\tilde{x}_{(i)}, \tilde{x}_{(j)}\right)\right\}$, where $\tilde{x}_{(i)}$ is the *i*-th value of the ordered sample. If $F_1 = F_2$, then the probability of $A_{ij}^{(k)}$ is equal to (4). Construct the Wilson confidence interval (2) $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ for the unknown probability of $A_{ij}^{(k)}$ using its observed relative frequency. The number of all confidence intervals $I_{ij}^{(n,m)} = \left(p_{ij}^{(1)}, p_{ij}^{(2)}\right)$ is equal to $N = (n-1)n/2$. Put $h = \frac{1}{N}\#\left\{\gamma_i + \gamma_{i+1} + ... + \gamma_{j-1} + \frac{j-i}{n+1} \in I_{ij}^{(n)}\right\}$. Compute a confidence interval $I^{(n,m)} = \left(p^{(1)}, p^{(2)}\right)$ for the probability $p\left(\gamma_i + \gamma_{i+1} + ... + \gamma_{j-1} + \frac{j-i}{n+1} \in I_{ij}^{(n,m)}\right)$ using (2) as described in Section 4.2. The statistics h is called the *empirical p-statistics*. It estimates the homogeneity of empirical samples $\tilde{x}$ and $\tilde{y}$, Wilson confidence interval (2), where we set h instead of h_{ij} and N instead of m is the confidence interval for the probability $p\left(\gamma_i + \gamma_{i+1} + ... + \gamma_{j-1} + \frac{j-i}{n+1} \in I_{ij}^{(n,m)}\right)$.

4.3 k-sample case

The two-sample Klyushin–Petunin test may be expanded on the *k*-sample case using the scheme one-vs-rest. Suppose that all *k* samples follow the same distribution. Then, if we select a sample and join other samples into one sample, we shall reduce the *k*-sample case to the two-sample case. Joining the samples following the same distribution, we obtain a sample following this distribution. Thus, if the p-statistics between selected and joined samples is greater than 0.95, the null hypothesis about identical distributions is accepted; otherwise, the *k* samples are heterogeneous.

5. Numerical Experiments

The strengths of the proposed approach are justified by numerical experiments comparing the p-statistics with the Kruskal–Wallis test and the Friedman test. We test the location shift hypothesis (samples have different means and the same variances) and the scale shift hypothesis (sample have the same means and different variances) using samples, which do not contain ties and samples with a single tie. For experiments, we selected samples from the Gaussian distributions $N(\mu, \sigma)$ with different overlapping, where μ is the mean and σ is the standard deviation.

5.1 Location Shift Hypothesis without Ties

To test a location shift hypothesis we 100 times generated by C++ pseudo random number generator 5 samples containing 10 real numbers that follows the distributions N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), and N(0.4, 1), and computed p-statistics by the scheme one-vs.-others. If the p-statistics between these samples was greater than 0.95, we concluded that they are homogeneous, otherwise the *k* samples were considered as heterogeneous. Average p-statistics was equal to 0.756. The Wilson confidence interval for the p-statistics, constructed as indicated in Section 4.1, was (0.532, 0.894). Hereinafter, Var1–Var5 denotes N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), and N(0.4, 1) in Tables 1, 3, 5, 11, 13, 15, and N(0,1), N(0, 2), N(0, 3), N(0, 4), and N(0, 5) in Tables 6, 8, 10, 16, 18 and 20.

TABLE 1
UPPER BOUNDS OF THE CONFIDENCE INTERVALS OF THE PAIRWISE P-STATISTICS FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITHOUT TIES

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.767	0.767	0.676	0.767
Var2	–	1.000	0.907	0.801	0.907
Var3	–	–	1.000	0.864	0.921
Var4	–	–	–	1.000	0.695
Var5	–	–	–	–	1.000

TABLE 2
SUMMARY OF THE KRUSKAL–WALLIS TEST FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITHOUT TIES

Kruskal–Wallis statistics (Observed value)	5.781
Kruskal–Wallis statistics (Critical value)	9.488
Degree of freedom	4
p-value (one-tailed)	0.216
Significance level	0.05

TABLE 3
PAIRWISE P-VALUES OF THE KRUSKAL–WALLIS TEST FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITHOUT TIES

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.974	0.505	1.000	1.000
Var2	0.974	1.000	0.984	0.653	0.974
Var3	0.505	0.984	1.000	0.110	0.505
Var4	1.000	0.653	0.110	1.000	0.700
Var5	0.789	1.000	0.921	0.700	1.000

ISSN 2395-8618

TABLE 4
SUMMARY OF THE FRIEDMAN TEST FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITHOUT TIES

Friedman statistics (Observed value)	4.080
Friedman statistics (Critical value)	9.488
Number of degrees of freedom	4
p-value (one-tailed)	0.395
Significance level	0.05

TABLE 5
PAIRWISE P-VALUES OF THE FRIEDMAN TEST FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITHOUT TIES

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	1.000	0.708	0.955	1.000
Var2	1.000	1.000	0.790	0.915	1.000
Var3	0.708	0.790	1.000	0.279	0.708
Var4	0.955	0.915	0.279	1.000	0.955
Var5	1.000	1.000	0.708	0.955	1.000

TABLE 6
UPPER BOUNDS OF THE CONFIDENCE INTERVALS OF THE PAIRWISE P-STATISTICS FOR DISTRIBUTIONS N(0,1), N(0,2), N(0, 3), N(0, 4), AND N(0, 5) WITHOUT TIES

Upper bound	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.732	0.732	0.767	0.749
Var2	–	1.000	0.945	0.957	0.957
Var3	–	–	1.000	0.695	0.741
Var4	–	–	–	1.000	0.921
Var5	–	–	–	–	1.000

The results cited in Table 1 suggest that any sample was not recognized as homogeneous with others. The results cited in Table 2 and Table 3 indicate that all the samples were recognized as homogeneous with others. Thus, we may ascertain that the Kruskal–Wallis fails in this case. The results cited in Table 4 and Table 5 show that the Friedman test fails in this case.

5.2 Scale Shift Hypothesis without Ties

To test a scale hypothesis, we 100 times generated five samples containing ten real numbers that follow the distributions N(0,1), N(0, 2), N(0, 3), N(0, 4), and N(0, 5), and computed p-statistics by the scheme one-vs-others. If the p-statistics between these samples is greater than 0.95, we concluded that they are homogeneous; otherwise, the *k* samples were considered heterogeneous. The average p-statistics was equal to 0.822. The confidence interval was (0.603, 0.934). This interval does not contain 0.95; thus, the samples in whole may be considered heterogeneous. The corresponding results are provided in Tables 6–10. Note that the Klyushin–Petunin test recognized the samples as heterogeneous in 8 cases of 10. Meantime, the Kruskal–Wallis and the Friedman tests failed.

TABLE 7
SUMMARY OF THE KRUSKAL–WALLIS TEST FOR DISTRIBUTIONS N(0,1), N(0, 2), N(0, 3), N(0, 4), AND N(0, 5) WITHOUT TIES

Kruskal–Wallis statistics (Observed value)	5.512
Kruskal–Wallis statistics (Critical value)	9.488
Number of degrees of freedom	4
Kruskal–Wallis statistics (Observed value)	0.239
Kruskal–Wallis statistics (Critical value)	0.05

TABLE 8
PAIRWISE P-VALUES OF THE KRUSKAL–WALLIS TEST FOR DISTRIBUTIONS N(0,1), N(0, 2), N(0, 3), N(0, 4), AND N(0, 5) WITHOUT TIES

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.155	1.000	0.700	0.961
Var2	0.155	1.000	0.213	0.984	0.996
Var3	1.000	0.213	1.000	0.700	0.974
Var4	0.700	0.984	0.700	1.000	1.000
Var5	0.961	0.996	0.974	1.000	1.000

TABLE 9
SUMMARY OF THE FRIEDMAN TEST FOR DISTRIBUTIONS N(0,1), N(0, 2), N(0, 3), N(0, 4), AND N(0, 5) WITHOUT TIES

Friedman statistics (Observed value)	7.600
Friedman statistics (Critical value)	9.488
Number of degrees of freedom	4
p-value (one-tailed)	0.107
Significance level	0.05

TABLE 10
PAIRWISE P-VALUES OF THE FRIEDMAN TEST FOR DISTRIBUTIONS N(0,1), N(0, 2), N(0, 3), N(0, 4), AND N(0, 5) WITHOUT TIES

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.118	0.993	0.279	0.527
Var2	0.118	1.000	0.279	0.993	0.915
Var3	0.993	0.279	1.000	0.527	0.790
Var4	0.279	0.993	0.527	1.000	0.993
Var5	0.527	0.915	0.790	0.993	1.000

5.3 Location shift Hypothesis with Ties

To test a shift hypothesis for samples with ties, we 100 times generated five samples containing ten real numbers with a single tie that follow the distributions N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), and N(0.4, 1), were two sample values were the same, and compute p-statistics by the scheme one-vs-others. The average p-statistics was equal to 0.533. The Wilson confidence interval constructed using the 3s-rule was (0.324, 0.731). This interval does not contain 0.95; thus, the sample in whole may be considered heterogeneous.

The results cited in Table 11 suggest that only 3 cases were recognized as homogeneous with others. The results cited in Tables 12–15 indicate that all the samples were recognized as homogeneous. Thus, we may ascertain that the Kruskal–Wallis test and the Friedman test fail in this case.

ISSN 2395-8618

TABLE 11
UPPER BOUNDS OF THE CONFIDENCE INTERVALS OF THE PAIRWISE P-STATISTICS FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITH SINGLE TIE

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.988	0.986	0.968	0.992
Var2	–	1.000	0.998	0.921	0.997
Var3	–	–	1.000	0.849	0.992
Var4	–	–	–	1.000	0.934
Var5	–	–	–	–	1.000

TABLE 12
SUMMARY OF THE KRUSKAL–WALLIS TEST FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITH SINGLE TIE

Kruskal–Wallis statistics (Observed value)	5,808
Kruskal–Wallis statistics (Critical value)	9,488
Degree of freedom	4
p-value (one-tailed)	0,214
Significance level	0,05

TABLE 13
PAIRWISE P-VALUES OF THE KRUSKAL–WALLIS TEST FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITH SINGLE TIE

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.984	0.505	1.000	0.745
Var2	0.984	1.000	0.984	0.652	0.998
Var3	0.505	0.984	1.000	0.110	0.921
Var4	1.000	0.652	0.110	1.000	0.700
Var5	0.745	0.998	0.921	0.700	1.000

TABLE 14
SUMMARY OF THE FRIEDMAN TEST FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITH SINGLE TIE

Friedman statistics (Observed value)	4,080
Friedman statistics (Critical value)	9,488
Number of degrees of freedom	4
p-value (one-tailed)	0,395
Significance level	0,05

TABLE 15
PAIRWISE P-VALUES OF THE FRIEDMAN TEST FOR DISTRIBUTIONS N(0,1), N(0.1, 1), N(0.2, 1), N(0.3, 1), AND N(0.4, 1) WITH SINGLE TIE

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	1.000	0.708	0.955	1.000
Var2	1.000	1.000	0.708	0.955	1.000
Var3	0.708	0.708	1.000	0.279	0.790
Var4	0.955	0.955	0.279	1.000	0.915
Var5	1.000	1.000	0.790	0.915	1.000

TABLE 16
UPPER BOUNDS OF THE CONFIDENCE INTERVALS OF THE PAIRWISE P-STATICTICS FOR DISTRIBUTIONS N(0,1), N(0,2), N(0, 3), N(0, 4), AND N(0, 5) WITH SINGLE TIE

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.921	0.985	0.946	0.921
Var2	–	1.000	0.894	0.993	0.879
Var3	–	–	1.000	0.948	0.934
Var4	–	–	–	1.000	0.946
Var5	–	–	–	–	1.000

TABLE 17
SUMMARY OF THE KRUSKAL–WALLIS TEST FOR DISTRIBUTIONS N(0,1), N(0, 2), N(0, 3), N(0, 4), AND N(0, 5) WITH SINGLE TIE

Kruskal–Wallis statistics (Observed value)	6,145
Kruskal–Wallis statistics (Critical value)	9,488
Degree of freedom	4
p-value (one-tailed)	0,189
Significance level	0,05

TABLE 18
PAIRWISE P-VALUES OF THE KRUSKAL–WALLIS TEST FOR DISTRIBUTIONS N(0,1), N(0, 2), N(0, 3), N(0, 4), AND N(0, 5) WITH SINGLE TIE

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.155	0.999	0.700	0.921
Var2	0.155	1.000	0.182	0.974	0.999
Var3	0.999	0.182	1.000	0.652	0.863
Var4	0.700	0.974	0.652	1	1.000
Var5	0.921	0.999	0.863	1.000	1.000

TABLE 19
SUMMARY OF THE FRIEDMAN TEST FOR DISTRIBUTIONS N(0,1), N(0, 2), N(0, 3), N(0, 4), AND N(0, 5) WITH SINGLE TIE

Friedman statistics (Observed value)	8,400
Friedman statistics (Critical value)	9,488
Number of degrees of freedom	4
p-value (one-tailed)	0,078
Significance level	0,05

TABLE 20
PAIRWISE P-VALUES OF THE FRIEDMAN TEST FOR DISTRIBUTIONS N(0,1), N(0, 2), N(0, 3), N(0, 4), AND N(0, 5) WITH SINGLE TIE

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.118	0.993	0.279	0.527
Var2	0.118	1.000	0.279	0.993	0.915
Var3	0.993	0.279	1.000	0.527	0.790
Var4	0.279	0.993	0.527	1.000	0.993
Var5	0.527	0.915	0.790	0.993	1.000

ISSN 2395-8618

TABLE 21
RESULTS OF ANN AND ARIMA MODELS FOR DELL STOCK INDEX [47]

Forecast error			Forecast error		
Time	ARIMA	ANN	Time	ARIMA	ANN
01.03.2010	0.0302	0.0162	18.03.2010	0.0206	−0.0110
02.03.2010	0.0095	0.0161	19.03.2010	0.0090	−0.0278
03.03.2010	0.0007	0.0110	22.03.2010	−0.0034	−0.0198
04.03.2010	0.0088	0.0000	23.03.2010	0.0019	0.01380
05.03.2010	0.0252	0.0072	24.03.2010	0.0220	−0.0080
08.03.2010	0.0086	0.0093	25.03.2010	−0.006	−0.0229
09.03.2010	0.0183	0.0134	26.03.2010	0.0160	−0.02135
10.03.2010	0.0325	0.0154	29.03.2010	0.0047	−0.0294
11.03.2010	0.0021	0.0007	26.03.2010	0.0160	−0.02135
12.03.2010	−0.0035	−0.0028	29.03.2010	0.0047	−0.0294
15.03.2010	0.0077	−0.0098	30.03.2010	−0.003	−0.0354
16.03.2010	−0.0133	−0.0147	31.03.2010	0.0033	−0.0386
17.03.2010	0.0062	−0.0014			

5.4 Scale Shift Hypothesis with Ties

To test a scale hypothesis, we 100 times generated five samples containing ten real numbers that follow the distributions N(0, 1), N(0, 2), N(0, 3), N(0, 4), and N(0, 5) and contain a single tie, and then compute p-statistics by the scheme one-vs-others. If the p-statistics between these samples was greater than 0.95, we concluded that they are homogeneous; otherwise, the *k* samples were considered as not homogeneous. The average p-statistics was equal to 0.777. The confidence interval constructed using 3s-rule is (0.555, 0.907). The interval does not contain 0.95; thus, the samples in whole may be considered as heterogeneous. The results provided in Table 16–20 show that the Kruskal–Wallis and Friedman tests failed.

5.3 Two-Sample Tests for Dell Stock Index

Let us apply the proposed test to data from [47] (Table 21). The authors of this paper investigated the accuracy of the ARIMA predictive model and artificial neural networks ANN using the Dell stock index collected over 23 days from the New York Stock Exchange during the period from August 17 (1988) to February 25 (2011) and containing 5680 observations. Here ARIMA(*p*,*d*,*q*) denotes the model of the autoregressive integrated moving average, where *p* is the number of time lags, *d* is the number of times the data have had past values subtracted, and *q* is the order of the moving-average model. Hereinafter, Var1–Var5 denotes ARIMA(1,0,0)–ARIMA(5,0,0) models.

Comparing the accuracy indicators, the authors of the paper concluded that ANN neural networks are more accurate than the autoregressive integrated moving average model in terms of relative forecast errors. Using the Klyushin-Petunin test, the Kruskal-Wallis test, and the Friedman test, we can draw different conclusions. According to the Klyushin-Petunin test, the errors of the two considered models are not statistically different since the upper bound of the confidence interval for the p-statistic (0.96) is greater than 0.95. Therefore, these models can be considered statistically equivalent, opposite to Kruskal-Wallis and Friedman tests (Tables 22–23).

TABLE 22
SUMMARY OF THE KRUSKAL–WALLIS TEST FOR DELL STOCK INDEX

Kruskal–Wallis statistics (Observed value)	6.607
Kruskal–Wallis statistics (Critical value)	3.841
Degree of freedom	1
p-value (one-tailed)	0.010
Significance level	0.05

TABLE 23
SUMMARY OF THE FRIEDMAN TEST FOR DELL STOCK INDEX

Friedman statistics (Observed value)	7.348
Friedman statistics (Critical value)	3.841
Number of degrees of freedom	1
p-value (one-tailed)	0.007
Significance level	0.05

TABLE 24
UPPER BOUNDS OF THE CONFIDENCE INTERVALS OF THE PAIRWISE P-STATICTICS FOR MODELS ARIMA(1,0,0), ARIMA(2,0,0), ARIMA(3,0,0), ARIMA(4,0,0), AND ARIMA(5,0,0)

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.953	0.929	0.927	0.921
Var2	–	1.000	0.899	0.892	0.902
Var3	–	–	1.000	0.905	0.899
Var4	–	–		1.000	0.939
Var5	–	–	–	–	1

TABLE 25
SUMMARY OF THE KRUSKAL–WALLIS TEST FOR MODELS ARIMA(1,0,0), ARIMA(2,0,0), ARIMA(3,0,0), ARIMA(4,0,0), AND ARIMA(5,0,0)

Kruskal–Wallis statistics (Observed value)	1.000
Kruskal–Wallis statistics (Critical value)	9.488
Degree of freedom	4
p-value (one-tailed)	0.910
Significance level	0.05

5.4 k-Sample Tests for Dell Stock Index

To extend this experiment on the *k*-sample case, let us consider 5 ARIMA models predicting the Dell Stock Index: ARIMA(1,0,0), ARIMA(2,0,0), ARIMA(3,0,0), ARIMA(4,0,0), and ARIMA(5,0,0) using the training sample published in [47] The results are provided in Tables 24–28. According to the Klyushin-Petunin test, the errors of the five considered ARIMA models are not statistically different in total (homogeneous) since the upper bound of the confidence interval for the

ISSN 2395-8618

p-statistic (0.99) is greater than 0.95. Thus, these models can be considered statistically equivalent. Meantime, the Kruskal-Wallis (Table 25, 26) and Friedman tests (Table 27, 28) lead to opposite conclusions. Pairwise comparisons show that the Klyushin–Petunin test recognizes different samples in almost all the cases.

TABLE 26
PAIRWISE P-VALUES OF THE KRUSKAL–WALLIS TEST FOR MODELS ARIMA(1,0,0), ARIMA(2,0,0), ARIMA(3,0,0), ARIMA(4,0,0), AND ARIMA(5,0,0)

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.155	1.000	0.700	0.961
Var2	0.155	1.000	0.213	0.984	0.996
Var3	1.000	0.213	1.000	0.700	0.974
Var4	0.700	0.984	0.700	1.000	1.000
Var5	0.961	0.996	0.974	1.000	1.000

TABLE 27
SUMMARY OF THE FRIEDMAN TEST FOR MODELS ARIMA(1,0,0), ARIMA(2,0,0), ARIMA(3,0,0), ARIMA(4,0,0), AND ARIMA(5,0,0)

Friedman statistics (Observed value)	18.400
Friedman statistics (Critical value)	9.488
Number of degrees of freedom	4
p-value (one-tailed)	0.001
Significance level	0.05

TABLE 28
PAIRWISE P-VALUES OF THE FRIEDMAN TEST FOR MODELS ARIMA(1,0,0), ARIMA(2,0,0), ARIMA(3,0,0), ARIMA(4,0,0), AND ARIMA(5,0,0)

	Var1	Var2	Var3	Var4	Var5
Var1	1.000	0.744	0.744	0.629	0.044
Var2	0.744	1.000	0.112	1.000	0.508
Var3	0.744	0.112	1.000	0.072	0.001
Var4	0.629	1.000	0.072	1.000	0.629
Var5	0.044	0.508	0.001	0.629	1.000

6. CONCLUSION

Direct comparison of forecasts using accuracy indicators without taking into account their stochastic nature is incorrect. Before comparing the accuracy, it is necessary to test the hypothesis about the identity of the distribution functions of different prediction models. To solve this problem, non-parametric methods are widely used, in particular, the Kruskal–Wallis and Friedman tests. We have proposed a new test that is effective for comparing predictive models. The level of asymptotic significance of this test does not exceed 0.05. The Klyushin-Petunin test is more universal than the Kruskal–Wallis and Friedman tests; it allows ordering pairs of samples and is easy to calculate. The practical usefulness of the proposed test is illustrated by examples. Further work will focus on the study of the theoretical properties of the proposed criterion.

REFERENCES

[1] N.V. Smirnov, "Estimate of difference between empirical distribution curves in two independent samples," *Byull. Mosk. Gos. Univ*., vol. 2, no. 2, pp. 3–14, 1939.

[2] N.H. Kuiper, "Tests concerning random points on a circle," *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A*. vol. 63, pp. 38–47, 1960.

[3] W.G. Dixon, "A criterion for testing the hypothesis that two samples are from the same population," *Ann. Math. Stat*., vol. 11, pp. 199–204, 1940.

[4] A. Wald, J. Wolfowitz, "On a test whether two samples ate from the same population," *Ann. Math. Stat.*, vol. 11, pp. 147–162, 1940.

[5] H.C. Mathisen, "A method of testing the hypothesis that two samples are from the same population," *Ann. Math. Stat.*, vol. 14. pp. 188–194, 1943.

[6] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrika*, vol. 1. pp. 80–83, 1945.

[7] H.B. Mann, D.R. Whitney, "On a test of whether one of the random variables is stochastically larger than other," *Ann. Math. Stat.*, vol. 18. pp. 50–60, 1947.

[8] S.S. Wilks, "A combinatorial test for the problem of two samples from continuous distributions," *Proc. Fourth Berkeley Symp. on Math. Stat. and Prob.*, vol. 1, pp 707–717, 1961.

[9] E.J.G. Pitman, "Significance tests which may be applied to samples from any populations," *J. Royal. Stat. Soc. Ser. A*, vol. 4, pp.119–130, 1937.

[10] E.L. Lehmann, "Consistency and unbiasedness of certain non-parametric tests," *Ann. Math. Stat.*, vol. 22, pp. 165–179, 1947.

[11] M. Rosenblatt, "Limit theorems associated with variants of the von Mises statistic," *Ann. Math. Stat.*, vol. 23, pp. 617–623, 1952.

[12] M. Dwass, "Modified randomization tests for non-parametric hypotheses," *Ann. Math. Stat.*, vol. 28, pp. 181–187, 1957.

[13] M. Fisz, "On a result be M. Rosenblatt concerning the Mises–Smirnov test," *Ann. Math. Stat.*, vol. 31, pp. 427–429, 1960.

[14] G.A. Barnard, "Comment on "The spectral analysis of point processes" by M.S. Bartlett," *J. Royal. Stat. Soc. Ser. B*, vol. 25, p. 294, 1963.

[15] Z.W. Birnbaum, "Computers and unconventional test-statistics," *Reliability and Biometry*, F. Prochan and R.J. Serfling (eds), Philadelphia, PA: SIAM, pp 441–458, 1974.

[16] K.-H. Jockel, "Finite sample properties and asymptotic efficiency of Monte Carlo tests," *Ann. Stat.*, vol. 14, pp. 336–347, 1986.

[17] D.L. Allen, "Hypothesis testing using L1-distance bootstrap," *Am. Stat.*, vol. 51, pp. 145–150, 1997.

[18] B. Efron, R.J. Tibshirani, "An Introduction to the Bootstrap". *Monographs on Statistics and Applied Probability*, Vol. 57, New York: Chapman-Hall, 1993.

[19] J.-M. Dufour, A. Farhat, "Exact non-parametric two-sample homogeneity tests for possibly discrete distributions". Preprint 2001-23, Center for Interuniversity research in Quantitative Economics (CIREQ), California Press, pp. 707–717, 2001.

[20] J. Kiefer, "k-Sample analogues of the Kolmogorov–Smirnov, Cramér–V. Mises test," *Ann. Math. Stat*., vol. 30, pp. 420–447, 1959.

[21] F.W. Scholz, M.A. Stephens, "k-Samples Anderson–Darling test," J. *Am. Stat. Assoc.*, vol. 82, pp. 918–924, 1987.

[22] J. Zhang, Y. Wu, "k-Sample test based on the likelihood ratio," *Comput. Stat. Data. Anal.*, vol. 51, pp. 4682–4691, 2007.

[23] B. Jiang, Y. Chao, J.S. Liu, "Nonparametric K-Sample Tests via Dynamic Slicing," *J. Amer. Stat. Assoc.*, vol. 110, no. 510, pp. 642-653, 2015.

[24] Y. Sang, X. Dang, Z. Yichuan, "Jackknife Empirical Likelihood Approach for K-sample Tests," arXiv:1908.00477v1, 2019.

[25] G.J. Sźekely, M.L. Rizzo, "Energy statistics: A class of statistics based on distances," *J. Stat. Plan. Infer.*, vol. 143, pp. 1249-1272, 2013/

[26] G.J. Sźekely, M.L. Rizzo, "The energy of data," *Ann. Rev. Stat. Appl.*, vol. 4, no. 1, pp. 447-479, 2017.

ISSN 2395-8618

[27] V.A. Fernández, M.D. Jiménez-Gamero, J. Garcìa, "A test for the two-sample problem based on empirical characteristic functions," *Comput. Stat. Data Anal.*, vol. 52, pp. 3730–3748, 2008.
[28] M. Hušková, S.G. Meintanis, "Tests for the multivariate k-sample problem based on the empirical characteristic function," J. Nonparametr. Stat.vol. 20, pp. 263–277, 2008.
[29] P. Camblor, J. de Uña-Alvarez, N. Corral, "k-Sample test based on the common area of kernel density estimator," J. Stat. Plan. Inference, vol. 138, pp. 4006–4020, 2008.
[30] P. Martínez-Camblor, J. de Uña-Alvarez, "Non-parametric k-sample tests: density functions vs. distribution functions," *Comput. Stat. Data Anal.*, vol. 53, pp. 3344–3357, 2009.
[31] S. Chen, "A new distribution-free k-sample test: Analysis of kernel density functionals," *Canadian J. Stat.*, vol. 48, No. 2, pp. 167-186, 2020.
[32] T.P. Hettmansperger, "Nonparametric Rank Tests". *International Encyclopedia of Statistical Science*, M. Lovric (eds) , Berlin, Heidelberg: Springer, 2011.
[33] W.H. Kruskal, W.A. Wallis, Use of ranks in one-criterion variance analysis," *J. Amer. Stat. Assoc*., vol. 47, no. 260, pp. 583–621, 1952.
[34] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Amer. Stat. Assoc.*, vol. 32, no. 200, pp. 675–701, 1937.
[35] M. Friedman, "A correction: The use of ranks to avoid the assumption of normality implicit in the analysis of variance", *J. Amer. Stat. Assoc.*. vol. 34, no. 205: p. 109, 1939.
[36] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," Ann. Math. Stat., vol. 11, no. 1, pp. 86–92, 1940.
[37] K.M. Wittkowski, "Friedman-Type statistics and consistent multiple comparisons for unbalanced designs with missing data," J. Amer. Stat. Assoc, vol. 83, no. 404, pp. 1163–1170, 1988.
[38] F.X. Diebold, R.S. Mariano, "Comparing predictive accuracy," J. Bus. Econ. Stat., vol. 20, no. 1, pp. 134–144, 2002.
[39] B.M. Hill, "Posterior distribution of percentiles: Bayes' theorem for sampling from a population," *J. Am. Stat. Assoc.*, vol. 63, pp. 677–691, 1968.
[40] D.A. Klyushin, Yu.I. Petunin, "A Nonparametric Test for the Equivalence of Populations Based on a Measure of Proximity of Samples," *Ukrainian Math. J*., vol. 55, no. 2, pp. 181–198, 2003.
[41] R.I. Andrushkiw, D.A. Klyushin, Yu.I. Petunin, "Proximity measure between samples with repetition factor greater than one," CAMS Research Report:0708-12, New Jersey Institute of Technology, 2008.
[42] D.F. Vysochanskij, YI Petunin "Justification of the 3σ rule for unimodal distributions," *Theory of Probability and Mathematical Statistics*, vol. 21, pp. 25–36, 1980.
[43] F. Pukelsheim The Three Sigma Rule. *The American Statistician*, vol. 48(2), pp. 88-91, 1994.
[44] S.A. Matveichuk, Yu.I. Petunin, "Generalization of Bernoulli schemes that arise in order statistics, I," *Ukrainian. Math. J.,* vol. 42, no. 4, pp. 459–466, 1990.
[45] S.A. Matveichuk, Yu.I. Petunin, "Generalization of Bernoulli schemes that arise in order statistics. II," *Ukrainian Math. J.*, vol. 43, no. 6, pp. 728–734, 1991.
[46] N. Johnson, S. Kotz, "Some generalizations of Bernoulli and Polya-Eggenberger contagion models," *Statist. Paper*, vol. 32, pp. 1–17, 1991.
[47] Adebiyi A, Adewumi A, Ayo C (2014) Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction. J App Math 2014:1-7.

ISSN 2395-8618

Processing Natural Language Queries via a Natural Language Interface to Databases with Design Anomalies

Rodolfo A. Pazos, José A. Martínez, and Alan G. Aguirre

***Abstract*—Natural language interfaces to databases (NLIDBs) have proven to be very promising tools when trying to obtain information from a relational database since they require the end user to have very little training and knowledge about databases to use them. However, their development has not been easy due to problems related to natural language processing. In addition to this, most authors overlook an important factor in developing these tools, which is the quality of the design of the database to be queried by the NLIDB. The problem arises because there can be many alternatives for the design of databases, and some contain design anomalies. Many NLIDBs would not work correctly for these databases since they were designed under the assumption that they would be used with databases without anomalies. This article describes an improvement to the processing performed by a domain-independent interface to treat databases with design anomalies and for the interface to be able to correctly process queries involving such anomalies. The literature on NLIDBs has not mentioned this problem and much less addressed it.**

***Index Terms*—Natural language interfaces, relational databases, user interfaces**

1. Introduction

NOWADAYS, information plays a very important role in business. Most of the information is, in many cases, stored in databases. However, for a user to obtain information from a database (DB), he must have knowledge of a query language for databases (such as SQL).

Due to this situation, access to DBs by inexperienced users is very limited. On many occasions, the information must be accessed by inexperienced users; therefore, it is necessary that they use easy-to-use software that does not require knowledge about DBs. To this end, a large number of tools have been developed, whose main characteristic is to show the DB schema and offer methods for obtaining information based on visual schemas. Although the mentioned tools allow building SQL queries from visual schemas, they are not very easy to use, as they require a certain degree of SQL knowledge.

NLIDBs are tools that allow inexperienced users to compose SQL queries using a natural language (NL) expression [1]. These interfaces are very easy to use for inexperienced users; however, their development has been delayed due to problems related to natural language processing (NLP) and the semantic content of DBs.

Nowadays, it is very common to find DBs that have design anomalies [2][3][4]. This is because sometimes developers include some anomalies in their DB designs to satisfy the needs of the applications for which they were designed (e.g., surrogate keys). However, this affects the performance of NLIDBs, since they are designed to work with correctly designed DBs, i.e., DBs without design anomalies.

In summary, the problem addressed in this article consists of developing a method that allows an NLIDB to be used for querying DBs that have design anomalies and answering correctly as if the DBs had no anomalies.

This paper describes methods for improving an NLIDB, allowing it to process queries formulated on DBs that have design anomalies. The anomalies dealt with are the following: the absence of primary and foreign keys, use of surrogate keys, use of columns for storing aggregate function calculations, and use of repeated columns in two or more tables. Additionally, to demonstrate the efficiency of these methods, we performed experimental tests of our NLIDB using DBs with design anomalies. Additionally, to demonstrate the efficiency of these methods, we performed experimental tests of our NLIDB using DBs with design anomalies.

2. State of the Art

Several NLIDBs have been developed since the 1970s, such as LUNAR [5], RENDEZVOUS [6], LADDER [7], among others. These interfaces could also be configured to work with other DBs despite having been designed for a DB in particular; however, this task was very difficult due to technical limitations at the time.

In [8] some NLIDBs are considered relevant to this work. It is important to remark that for most NLIDBs, there is no software that can be used for testing. These NLIDBs are relevant for two reasons: First, there is a prototype (C-Phrase) or commercial software (ELF) of the NLIDB, which can be tested and, thus, demonstrate if they have any mechanism to treat design anomalies. Second, the other NLIDBs have been recently developed.

Manuscript received on October 12, 2020, accepted for publication on November 14, 2020, published on December 30, 2020.

The author is with the Instituto Tecnológico de Cd. Madero, Tecnológico Nacional de México, Cd. Madero, Tamaulipas 89440, Mexico (e-mail: r_pazos_r@yahoo.com, jose.mtz@gmail.com, li.aguirre.lam@hotmail.com).

ISSN 2395-8618

TABLE 1
STATE OF THE ART NLIDBS

NLIDB	Year	Design Anomaly Treatment
NADAQ	2019	No
Cross-domain NLI	2019	No
TEMPLAR	2019	No
nQuery	2017	No
Aneesah	2015	No
NL2CM	2015	No
NaLIR	2014	No
NLWIDB	2013	No
C-Phrase	2010	Partially
ELF	2004	Partially

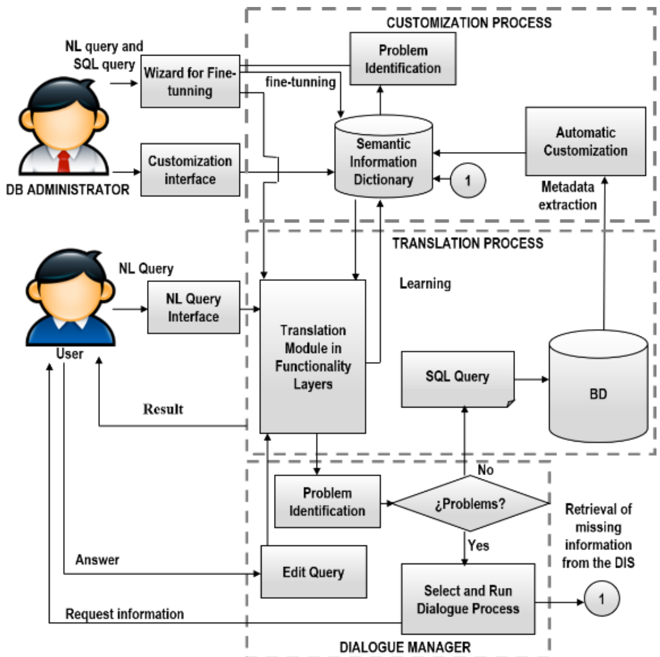

Fig. 1. NLIDB architecture.

In this article, the criteria mentioned are used to choose the relevant NLIDBs for the state of the art.

It is worth mentioning that there are also software tools for design and analysis of DBs. Some examples of these are Visual Paradigm, Vertabelo, DbSchema, Toad Data Modeler, among others.

The aforementioned tools, in addition to allowing users to easily design DBs from entity-relationship diagrams, also have analysis tools on the structure of the designed DB. However, most of the analysis and error correction tools that these software tools have been for correcting errors related to dirty DBs.

Additionally, the design anomalies considered in this article require semantic information that only the user can provide. Unfortunately, none of the mentioned software considers this kind of problem.

Although no work mentioned so far has addressed the problem of design anomalies in DBs, the only readily available interfaces that can be used for testing are included in this analysis of the state of the art (Table 1), such as C-PHRASE [9] and ELF [10]. Others that have been recently developed are also included, such as NADAQ [11], Cross-domain NLI [12], TEMPLAR [13], nQuery [14], Aneesah [15], NL2CM [16], NaLIR [17], NLWIDB [18]. In the former, the performance they have with DBs with design anomalies is evaluated.

It is important to mention that in this analysis, domain-specific NLIDBs are omitted since they are designed to work with the design anomalies of the DB for which they were designed.

Both the more recently developed NLIDBs and commercial NLIDBs have not considered design anomalies in DBs as a problem that must be solved in order to correctly translate NL queries to SQL queries. Therefore, it cannot be guaranteed that the operation of most of the NLIDBs listed will be correct when using DBs with design anomalies.

3. OUR NLIDB

The NLIDB used in this work is a domain independent prototype interface for the Spanish language [19].

A crucial component of this interface is the Semantic Information Dictionary (SID), shown in Figure 1. The SID is a DB that stores metadata of the DB in use (DB schema, columns, tables) and useful semantic information (nominal, verbal and prepositional descriptors, among others) to relate the query in NL to the elements of the SQL statement that the interface needs to build.

In addition to the SID, this interface consists of two main processes: the customization process and the translation process shown in Figure 1.

The customization process is carried out by the DB administrator (DBA), where the DBA configures the SID by entering semantic information regarding the tables, columns and relationships that exist in the schema of the DB in use.

The process of translating a query consists of three sub-processes: lexical analysis, syntactic analysis and semantic analysis (Figure 2).

Each sub-process is carried out by functional layers, where each layer deals with a problem that must be solved so that the interface can obtain a correct transla-tion of the NL query.

Lexical analysis. It performs a lexical tagging process, obtaining the syntactic category (part of speech) of each word in the query from a lexicon stored in a DB. In this layer, lexical errors are corrected, syntactic ambiguity and homography problems are resolved. The result obtained consists of a tagged query.

Syntactic analysis. The tagged query is used to build a syntactic tree, where syntactic errors are corrected, syntactic ellipsis is resolved, and anaphora problems are detected.

ISSN 2395-8618

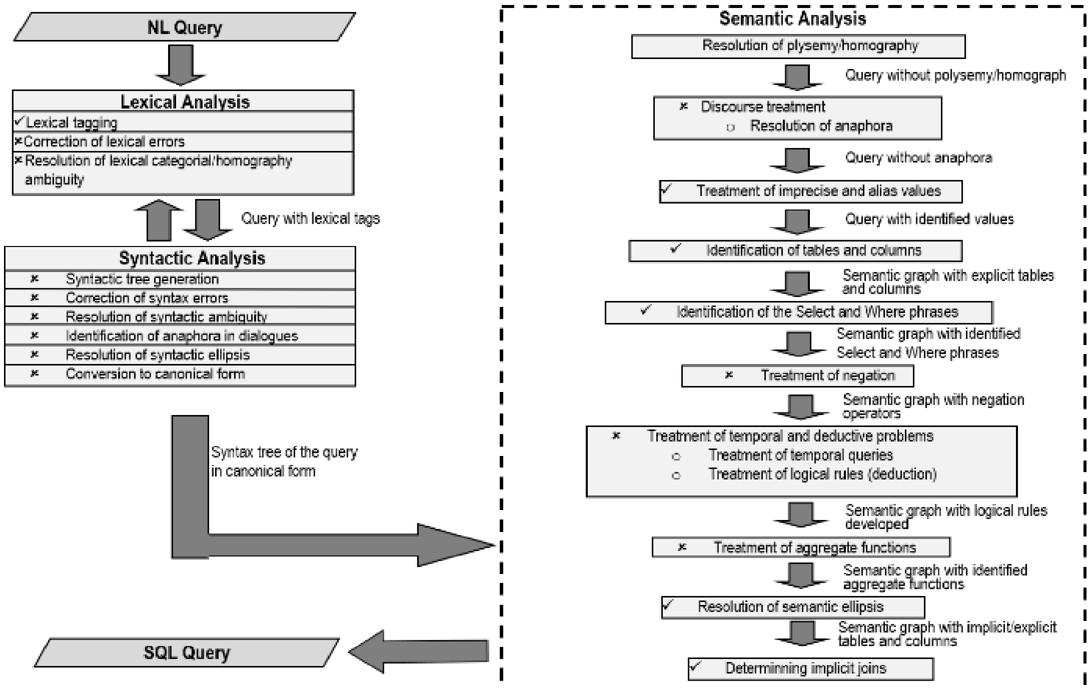

Fig. 2. Functionality layers of the translation module.

Semantic analysis. A representation of the meaning of the tagged query is constructed, which can be used to translate it into SQL. This layer is the most complex since most of the problems are due to the interpretation of the meaning of the query. This layer is divided into the following sub-layers:

1. Treatment of polysemy and homography.
2. Anaphora treatment.
3. Treatment of imprecise values and aliases.
4. Identification of tables and columns.
5. Identification of Select and Where phrases.
6. Treatment of negative queries.
7. Treatment of temporary and deductive problems.
8. Treatment of aggregation and grouping functions.
9. Resolution of semantic ellipsis.
10. Determination of implicit joins.

It is important to mention that, currently, only the most necessary layers (with a checkmark ✓) have been implemented (Figure 2).

Useful information is collected when processing the NL query via the functional layers. With this information, the NLIDB constructs a SQL query whose result contains the information requested by the user.

4. Processing NL Queries Formulated in DBs with Design Anomalies

The NLIDB [19] is designed on the assumption that the DB in use has no design anomalies; therefore, it does not perform well when translating queries to DBs with this problem.

Four design anomalies are considered in this article:

1. Absence of primary and foreign keys.
2. Use of surrogate keys.
3. Columns for storing aggregate function (AF) calculations.
4. Repeated columns in two or more tables.

Before the end user can use the NLIDB to answer queries, an initial setup process must be performed. This process is carried out semi-automatically by the NLIDB with the help of the DBA. In principle, the DBA must choose the DB with which he will work, and the NLIDB obtains the metadata from the DB and uses it to configure the SID. It is important to mention that the metadata obtained includes the design anomalies contained in the DB; therefore, the SID configuration has a representation of the DB with design anomalies.

4.1 Queries that Involve Absence of Primary and Foreign Keys

The absence of foreign keys directly affects the performance of the NLIDB. This is because the NLIDB, by means of the foreign keys defined in the DB schema, saves in the SID the existing relationships between the tables that have foreign keys defined and the tables to which they refer. This process is used in the sublayer Determination of Implicit Joins, where a semantic graph is created to determine the joins between tables when two or more tables are involved in the query.

Without the foreign keys properly defined in the DB, the NLIDB constructs a semantic graph where some tables are not connected as they should be. Additionally, the NLIDB is not

ISSN 2395-8618

able to deduce the existing relationships between tables and, therefore, it is not able to build a SQL query with the joins required by the NL query.

To solve the aforementioned problem, a configuration module was implemented that allows the DB administrator to specify relationships between tables whose foreign keys are not defined in the DB. In this way, the NLIDB can use these relationships to build a semantic graph, which contains enough information to define the joins in the SQL query in case they are required.

For example, consider the following NL query for the Geobase DB [20]:

¿Qué ríos pasan por el estado de Alaska?
Which rivers run through the state of Alaska?

Assuming that the DB does not have foreign keys defined, the SQL query built by the NLIDB would be the following:

```
SELECT river.river_name FROM river, state
WHERE state.state_name LIKE 'Alaska';
```

As can be seen, the NLIDB only detects the tables directly involved in the query; however, there is an intermediate table (*riverstate*), which is not included in the query. Furthermore, relationships between tables are not defined in the SID; therefore, the NLIDB does not com-pose the joins between them.

Once the relationships between tables have been defined in the SID using the configuration module, the following query is obtained in SQL, which contains the intermediate tables and joins necessary in the query:

```
SELECT river.river_name
FROM river, state, riverstate
WHERE state.state_name LIKE 'Alaska'
AND state.abbreviation = riverstate.state_abbreviation
AND riverstate.river_id = river.river_id;
```

4.2 Queries that Involve Surrogate Keys

The use of surrogate keys affects the performance of the NLIDB, because this anomaly can cause data redundancy and lack of relationships between tables. Data redundancy affects the results obtained by the NLIDB. In these cases, some of the rows returned by the NLIDB can be confusing for the end user. The lack of relationships affects the creation of joins in the SQL query and, therefore, the result.

To solve this problem, a configuration interface for the SID was developed to define surrogate keys and relationships through foreign keys between the tables that have a surrogate key and other base tables in the SID [21]. Additionally, an algorithm was implemented to improve the processing of the interface so that it may be able to use the relationships defined by the configuration interface and ignore the relationships defined in the DB schema related to surrogate keys.

For example, consider the NL query:

¿En cuál estado se encuentra el lago Ontario?
In which state is Lake Ontario?

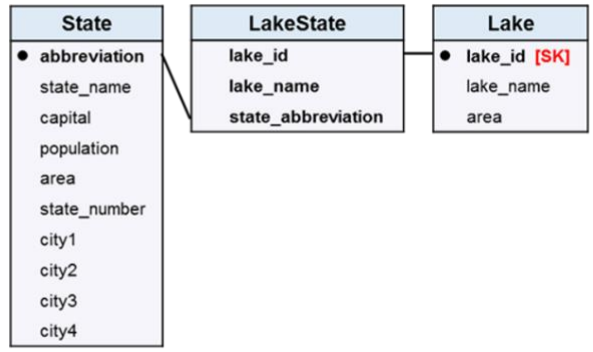

Fig. 3. Geobase DB fragment with a surrogate key.

The tables involved in the query are *State*, *LakeState*, and *Lake*. In Figure 3 a fragment of the DB schema with the design anomaly is presented, where *Lake.lake_id* is a surrogate key, *LakeState.lake_id* is a foreign key that connects with the surrogate key, and there could be a natural relationship between *Lake.lake_name* and *LakeState.lake_name*. However, due to the use of the surrogate key, no such relationship was defined in the DB schema.

The surrogate key *Lake.lake_id* and the foreign key (*LakeState.lake_name* REFERENCES *Lake.lake_name*) are defined in the SID by means of the configuration module. When defining the foreign key, the NLIDB marks the foreign key related to the surrogate key (*LakeState.lake_id* REFERENCES *Lake.lake_id*) as null so that it is ignored in the NL query translation process.

Processing the query using the NLIDB without treat-ing the design anomaly would result in the following SQL query:

```
SELECT State.state_name
FROM State, LakeState, Lake
WHERE Lake.lake_name LIKE 'Ontario'
AND State.abbreviation=LakeState.state_abbreviation
AND LakeState.lake_id = Lake.lake_id;
```

While the query obtained when treating the design anomaly through the configuration interface is as fol-lows:

```
SELECT State.state_name
FROM State, LakeState, Lake
WHERE Lake.lake_name LIKE 'Ontario'
AND State.abbreviation = LakeState.state_abbreviation
AND LakeState.lake_name = Lake.lake_name;
```

The first query includes joins and the surrogate key, while the second query includes a natural join using the columns *LakeState.lake_name* and *Lake.lake_name*.

4.3 Queries that Involve Columns for Storing Aggregate Function Calculations

The columns that can be calculated with AFs cause the NLIDB to obtain erroneous results since the values ob-tained can also be calculated by means of an AF applied to a column of another table.

ISSN 2395-8618

Algorithm 1 Pseudo-code for processing columns for storing AF calculations

```
1:  procedure AFColumnsProcess(Q,n)
2:      for i=0 to n-1 do
3:          if isAFColumn(Q_i) then
4:              AFColVal = getAF(Q_i) + "("+getAFColumn(Q_i)+")"
5:              setFinalTag(Q_i, AFColVal)
6:          endif
7:      endfor
```

Fig. 4. Algorithm 1.

For example, consider the following NL query:

¿Cuál es la población del estado de Mississippi?
What is the population of the state of Mississippi?

The population of a state could be calculated in two ways: using the column *State.population* or by adding the populations of the cities of the state. The DBA must decide whether to use the anomalous column or to use an AF instead.

For this reason, a configuration module was devel-oped that allows the DBA to define the columns that can be calculated with AF in the SID and assign them an AF associated with a column so that it can be used instead of the anomalous column.

An algorithm that modifies the internal functioning of the NLIDB in such a way that it can use the AFs associ-ated with this type of columns was also developed.

Algorithm 1 shows the pseudocode that describes the processing of columns for storing AF calculations. In the pseudocode, Q is the NL query entered by the user, Q_i is a token (word of NL query) of query Q, n is the total number of tokens in Q, and *AFColVal* is a variable used to store the final label that the referred token will be assigned. In line 3, for each token of query Q, it is verified if the column to which the token Q_i refers to is a column that stores an AF calculation. In this case, in line 4, an expression (character string) is stored in *AFColVal*. The expression consists of the AF (avg, count, etc.) stored in the SID for the mentioned column and the name of the column to which the AF will be applied. Lastly, the final label of token Q_i is updated with the information from *AFColVal*.

Regarding the example, when processing the NL query without treating the design anomaly, the NLIDB builds a query in SQL with the column State.population in the Select clause as follows:

```
SELECT State.population
FROM State
WHERE State.name LIKE 'Mississippi';
```

On the other hand, by defining the column *State.population* as a column that stores AF calculations in the SID and assigning the AF SUM(*City.population*) instead, the NLIDB will detect that there are two tables involved in the query: *City* in the Select clause and *State* in the Where clause; therefore, a query will be generated with their respective join:

```
SELECT SUM(City.population)
FROM State, City
WHERE State.name LIKE 'Mississippi' AND
State.abbreviation = City.state_abbreviation;
```

It is important to mention that the DBA can configure the NLIDB for using the column with the design anomaly or using the AF in query processing.

4.4 Queries that Involve Repeated Columns in Two or More Tables

The presence of repeated columns in multiple tables is an anomaly that creates conflict in the NLIDB processing when deciding to choose the right column from the DB that has several occurrences in different tables.

The SID contains information to match the words of an NL query with the columns of a DB. However, when in a DB there are two or more tables with repeated columns, the NLIDB is not able to know which is the right column that contains the information referred to by the user. Most of the time, columns that suffer from this problem can contain incorrect information due to insertion errors. Therefore, when the NLIDB refers to such information, it will be erroneous.

To solve the problems related to this anomaly, the DBA must identify the column that contains the correct information (usually a column that corresponds to an attribute of a strong entity). Afterwards, the DBA must indicate the columns that are repeated through the con-figuration module implemented for this purpose. Once this is done, the SID will have the necessary information to carry out the processing of queries that involve this type of columns.

In addition to the above, an algorithm was developed to allow the NLIDB to identify these columns and, thus, be able to correctly process queries.

In line 3 of Algorithm 2, for each token in query Q, the tokens that refer to a column are identified (the meanings of Q, Q_i and n are the same as those for Algorithm1). Subsequently, in lines 4 and 5, the columns *repCols* that are repeated in the DB with their respective labels *repColsTags* (the columns with the most reliable information) are obtained. This column information is obtained from the SID. In lines 6 to 10 a verification is carried out to identify the token that refers to a repeated column; in case of identifying one, in line 8 the token is labeled with the correct column.

ISSN 2395-8618

TABLE 2
EXPERIMENTAL TESTS RESULTS

Design Anomaly	Our NLIDB		ELF	
	ATIS	Geobase	ATIS	Geobase
Abs. of PKs and FKs	5	5	0	0
Use of Surrogate keys	5	5	0	0
Columns for storing AF calculations	5	5	0	2
Repeated columns	5	5	1	1
Total	20	20	1	3

Algorithm 2. Pseudocode for Processing repeated columns

```
procedure TreatmentOfRepColumns(Q,n)
  for i=0 to n-1 do
    if isColumn(Q_i) then
      repCols = getRepeatedCols()
      repColsTags = getRepeatedColsTags()
      for j=0 to sizeOf(repCols)
        if getColumnTag(Q_i) == repCols_j
          setFinalTag(Q_i, repColsTags_j)
        endif
      endfor
    endif
  endfor
```

Fig. 5. Algorithm 2.

To exemplify the operation of the mentioned pseudocode, consider the query:

¿Qué montañas están en el estado de Alaska?
Which mountains are in the state of Alaska?

When processing the previous query by the NLIDB, it detects the *Mountain.mountain_name* column for the Select clause and the *Mountain.state_name* column for the Where clause with its search value Alaska. The *state_name* column is found in 5 tables of the DB (*City*, *State*, *Mountain*, *High-Low*, *Border*). However, the column that has the most reliable information is *State.state_name*, and the other columns contain duplicate and unreliable information. When running Algorithm 2, the NLIDB avoids the use of the *Mountain.state_name* column as it does not contain reliable information and uses the *State.state_name* column instead. Once this is done, the NLIDB proceeds to identify the implicit joins between tables, and finally obtains the following SQL query:

```
SELECT Mountain.mountain_name
FROM State, Mountain
WHERE State.state_name LIKE 'Mississippi' AND
  State.abbreviation = Mountain.state_abbreviation;
```

5. EXPERIMENTAL RESULTS

In the experimental tests, a comparison of our NLIDB and the ELF NLIDB was made [10]. The ATIS [20] and Geoquery880 [20] corpora were used for the tests. For each design anomaly, five queries from the ATIS corpus and five queries from the Geoquery corpus were selected, giving a total of 40 queries. Each of the test cases has the following characteristics: it considers an NL query that involves a fragment of the DB, a design anomaly (created or existing) was considered in any of the tables involved in the query, and the SQL query resulting from NLIDB processing without/with anomaly treatment.

It is important to mention that for the purposes of this project, some queries of the two aforementioned corpora were modified since to perform tests on some design anomalies, columns that did not exist in the tables had to be introduced to simulate the design anomalies.

Table 2 shows the results of the comparative tests described. Our NLIDB correctly answers 20 queries from the ATIS corpus and 20 queries from the Geobase corpus. This is due to the handling of design anomalies since, without this mechanism, the NLIDB would not answer any query correctly.

To test the absence of foreign keys, they were removed from the DB and from the SID. To carry out the test with ELF, only foreign keys were eliminated from the DB schema, and an express configuration was used. ELF was not able to correctly answer any of these queries because, when configured, it uses the foreign keys defined in the DB schema to store them in its dictionary and, thus, to be able to build the required joins in the SQL queries. However, as the DB schema does not have foreign keys defined, it is not possible for ELF to define the necessary joins. Furthermore, ELF does not offer a tool to specify foreign keys in a DB without modifying the structure of its schema.

In the test for the use of surrogate keys, some columns with identifiers were included to simulate surrogate keys, and foreign keys were created with other tables referring to the defined surrogate key. For testing our NLIDB, DBs with the appropriate characteristics were created, and the SID was configured to include the surrogate keys, as well as their relationships with other tables. For ELF, only the DB schema was modified, and an express configuration was used. In this type of query, our NLIDB correctly answered because, through the configuration module for design anomalies, the surrogate keys were defined, and the existing relationships were modified using natural foreign keys. In contrast, ELF was unable to answer correctly, as it was repeatedly unable to interpret the

ISSN 2395-8618

NL query. Other times it misinterprets the query by including Boolean values in queries about flight fares.

Concerning columns used for storing AF calculations, in our NLIDB the test was carried out by creating columns with this characteristic in the DB and additionally specifying these columns in the SID. Subsequently, this anomaly was treated using the configuration module for design anomalies, and finally, the NLIDB used the processing algorithm to detect and process these columns with anomalies. Regarding ELF, the columns with the anomaly were also defined in the DB schema. However, for ATIS DB queries, ELF built erroneous queries be-cause it used tables and columns that were not required in the query, and it also ignored the use of AFs to process the queries. For Geobase queries, ELF was able to correctly answer two queries, as the structure of these queries was easier to understand, and the Geobase schema is much smaller than that of ATIS. Due to the above, ELF was able in these two cases to ignore the column used to store AF calculations and apply the necessary AF.

A test for the use of repeated columns in multiple tables was carried out in our NLIDB by repeating columns that are not foreign keys in different tables and whose information is already defined in another column of another table. In this test, the NLIDB had no problem when using the columns with the most reliable information because, when it found a repeated column, it was already defined in the SID, and the column with the most reliable information was associated with it. On the other hand, ELF obtained a correct query from the ATIS corpus and a correct query from the Geobase corpus. The main reason why ELF got only two successful queries with this type of anomaly is that, when it finds the repeated column in a table closest to the column associated with a search value, it always uses this column; otherwise, ELF can ignore it and build the query using the column that has the most reliable information.

6. Conclusions and Discussion

There are many NLIDBs. Most of these tools use approaches that aim to solve only the problems related to NL processing, leaving aside the problems inherent to the structure of the DB in use. One of the most important aspects of using a DB in an NLIDB is its design. On many occasions, this design may have anomalies; therefore, most NLIDBs would not work properly with a large number of DBs that suffer from this problem.

In this work, a mechanism to treat design anomalies in DBs was presented, which was implemented in a do-main-independent NLIDB [19]. To treat these design anomalies, a configuration module was implemented to introduce information about the anomalies of a DB in the SID. In this way, it is possible to present to the NLIDB a representation of the DB without design anomalies so that it works correctly. The above is important since a tool of this type should not modify the schema of the DB in use since it is often used by various applications, which would also have to be modified. Consequently, algorithms were implemented to process queries involving DB fragments that have design anomalies. These algorithms, in conjunction with additional information on the anomalies contained in the SID, allow the NLIDB to work with a DB with design anomalies as if it were a DB without anomalies.

The absence of foreign keys affects only the creation of joins when processing queries with this anomaly. However, if an NLIDB has no way of specifying relationships without modifying the DB schema, it is very difficult for it to be able to correctly construct SQL queries.

In the tests presented in Section 5, ELF is able to correctly construct only four queries, of which three were from Geobase and one from ATIS. Of these queries, two involved columns for storing AF calculations. This was possible because ELF ignored the use of columns that had this anomaly. However, for anomalies that required more specific processing, such as missing foreign keys and the use of surrogate primary keys, ELF could not answer any queries correctly.

Two more anomalies have been detected in the literature on NLIDBs: the absence of the second normal form of some DB table and the absence of the third normal form. A method for dealing with these two anomalies will be developed in the near future.

Acknowledgment

PhD student Alan Gabriel Aguirre Lam acknowledges the scholarship (Grantee No. 510415) by the Consejo Nacional de Ciencia y Tecnología, Mexico.

References

[1] I. Androutsopoulos, G. Ritchie, and P. Thanisch, “Natural Language Interface to Databases: An Introduction,” Natural Language Engineering, vol. 1, no. 1, pp. 29–81, 1995.

[2] O. Pivert and H. Prade, “Handling Dirty Databases: From User Warning to Data Cleaning Towards an Interactive Approach,” Proc. Fourth International Conference on Scalable Uncertainty Management, France, 2010.

[3] M.L. Pedro-de-Jesus and P.M.A. Sousa, “Selection of Reverse Engineering Methods for Relational Databases,” Proc. European Conference on Software Maintenance and Reengineering, IEEE, 1999.

[4] N. Mfourga, “Extracting Entity-Relationship Schemas from Relational Databases: A Form-Driven Approach,” Proc. Fourth Working Conference on Reverse Engineering, IEEE, 1997.

[5] W. Woods, R. Kaplan and B. Nash-Webber, “The Lunar Sciences Natural Language Information System: Final Report,” BBN Report 2378, Bolt Beranek and Newman Inc., 1972.

[6] E.F. Codd, “Seven Steps to Rendezvous with the Casual User,” Proc. IFIP Working Conference Data Base Management, pp. 179–200, 1974.

[7] G. Hendrix, E. Sacerdoti, D. Sagalowicz, and J. S. Locum, “Developing a Natural Language Interface to Complex Data,” ACM Transactions on Database Systems, vol. 3, no. 2, pp. 105–147, 1978.

[8] R.A. Pazos, J.A. Martínez, A.G. Aguirre, and M.A. Aguirre, “Issues in Querying Databases with Design Anomalies Using Natural Language Interfaces,” Fuzzy Logic Augmentation of Neural and Optimization Algorithms: Theoretical Aspects and Real Applications, Studies in Computational Intelligence, vol. 749, pp. 461–473, 2018.

[9] M. Minock, “C-phrase: A system for Building Robust Natural Language Interfaces to Databases,” Data Knowl. Eng., vol. 69, no. 3 290–302, 2010.

[10] S. Conlon, J. Conlon, and T. James, “The Economics of Natural Language Inter-

ISSN 2395-8618

faces: Natural Language Processing Technology as a Scarce Resource," Decis. Support Syst., vol. 38, no. 1, pp. 141–159, 2004.

[11] X. Boyan, C. Ruichu, Z. Zhenjie, Y. Xiaoyan, H. Zhifeng, L. Zijian, and L. Zhihao, "NADAQ: Natural Language Database Querying Based on Deep Learning," IEEE Access, vol. 7, pp. 35012–35017, 2019.

[12] W. Wang, "A Cross-Domain Natural Language Interface to Databases Using Adversarial Text Method," Proc. VLDB 2019 PhD Workshop, 2019.

[13] C. Baik, H.V. Jagadish, and Y. Li, "Bridging the Semantic Gap with SQL Query Logs in Natural Language Interfaces to Databases," Proc. IEEE International Conference on Data Engineering (ICDE), 2019.

[14] N. Sukthankar, S. Maharnawar, P. Deshmukh, Y. Haribhakta, and V. Kamble, "nQuery - A Natural Language Statement to SQL Query Generator," Proc. 55th Annual Meeting of the Association for Computational Linguistics Student Research Workshop, pp. 17–23, 2017.

[15] J. D. O'Shea, K. Shabaz, and K.A. Crockett, "Aneesah: A Conversational Natural Language Interface to Databases," Proc. World Congress on Engineering 2015, vol. 1, 2015.

[16] Y. Amsterdamer, A. Kukliansky, and T. Milo, "A Natural Language Interface for Querying General and Individual Knowledge," in Proc. VLDB Endow., 2015.

[17] F. Li and H.V. Jagadish, "Constructing an Interactive Natural Language Interface for Relational Databases," Proc. VLDB Endow., 2014.

[18] R. Alexander, P. Ruksha, and S. Mahesan, "Natural Language Web Interface for Database (NLWIDB)," Proc. Third International Symposium, SEUSL, Sri Lanka, 2013.

[19] R.A. Pazos R., M.A. Aguirre, J.J. González B., J.A. Martínez, J. Pérez, and A.A. Verástegui, "Comparative Study on the Customization of Natural Language Interfaces to Databases," SpringerPlus vol. 5, 553, 2016.

[20] Geobase880 Query corpus, https://www.cs.utexas.edu/users/ml/ nldata/geoquery.html, 2020.

[21] G. Sidorov, R.A. Pazos, J.A. Martínez, J.M. Carpio, and A.G. Aguirre, "Configuration Module for Treating Design Anomalies in Databases for a Natural Language Interface to Databases," Intuitionistic and Type-2 Fuzzy Logic Enhancements in Neural and Optimization Algorithms: Theory and Applications, vol. 862, Springer, 2020.

ISSN 2395-8618

Multicriteria Analysis for IoT Selection in a Telemetry System

Hermann M. Klusmann, Renzo M. Carnero

***Abstract*—The speed of technology evolution offers us more and better alternatives to solve problems and to increase the efficiency of processes. Such is the case of Communications networks and especially the Internet. Nowadays, Internet connection has become, to some extent, a necessity within societies; all kinds of devices can be connected through the Internet (IoT). In this sense, using the Internet for data transmissions at industrial levels (Industry 4.0) brings technical and economic advantages that allow improving the processes efficiency. It is possible to obtain good quality data in real-time and at a low cost using the Internet in Telemetry processes. This paper proposes a multicriteria analysis to find the most suitable IoT technology for a telemetry network of water meters in the city of Huacho, Peru. First, the types of IoT available in the area are established, then a balance is made accordingly to technical, social, and economic criteria; and finally, the most appropriate IoT technology is obtained for the case study.**

***Index Terms*—Internet, information technology and systems, network protocols, secision tables.**

1. Introduction

THERE is no doubt that technological evolution conditions the way we see the world and interact with it. This evolution is increasingly accelerated. The techniques and knowledge that arise in response to everyday problems acquire new applications in a short time. Such is the case with telemetry and the Internet.

The appearance of concepts such as the Internet of Things (IoT), which defines the possibility of connecting computers through the Internet, has resulted in the evolution of applications that explore the possibility of connecting industrial systems remotely through this communication network. It is at this point where telemetry systems acquire a new dimension.

This paper proposes a multicriteria analysis to find the most suitable IoT technology for a telemetry network of water meters in the city of Huacho, Peru. First, the available IoT platforms are described showing their most important characteristics. Then, in the third section, the methodology used for the multicriteria analysis is explained. This section presents the entire procedure and the equations applied to reach an objective decision. The fourth section contains the analysis carried out for the case study. Finally, in the fifth section, the result obtained in the previous one is discussed, and possible future applications of the methodology used are proposed.

Manuscript received on September 7, 2020, accepted for publication on November 20, 2020, published on December 30, 2020.

H. M. Klusmann is with the Peruvian University of Applied Sciences, Lima 15023, Peru (e-mail: hermann.klusmann@upc.edu.pe).

R. M. Carnero is with Static Control Components, Inc., Sanford, NC 27332, USA (e-mail: renzoc@scc-inc.com).

2. IoT Technologies

The term IoT encompasses everything connected to the Internet, but it is increasingly being used to define objects that “talk” to each other [1]. This technology has been being applied in different industrial areas, such as smart agriculture, smart health care, or smart manufacturing [2]. The evolution of technology has brought low-power consumption devices, more efficient communication protocols, and several link platforms. These important advantages for distance connection between devices make the application of the Internet in Telemetry systems almost an obligation.

IoT platforms have a central role within communications systems and allow the implementation of different applications accordingly to their own characteristics [3]. In this sense, to select an IoT platform for a Telemetry system, there are several criteria that should be established, then weighted accordingly to their importance, and finally evaluated.

There are several IoT technologies available around the world; they will be briefly described below.

2.1 DASH7

The DASH7 Alliance (D7A) is an open-source active RFID standard for WSAN protocol. D7A complies with the ISO/IEC 18000-7 standard. ISO/IEC 18000-7 is an open standard for the license-free 433 MHz ISM band air-interface for wireless communications. The 433 MHz frequency provides D7A with a long propagation distance and better penetration. A full OSI stack (7 OSI layers) known as D7A protocol (D7AP) is specified. It provides a long-range (up to 2 Km) and low latency with multi-year battery life to connect moving objects [4].

2.2 Nb-IoT

Narrow Band Internet of Things was set up by 3GPP in Cellular systems in support of ultra-low complexity and low throughput Internet of Things. It defines a new radio access technology that can be integrated into the LTE standard. NB-IoT is built from existing LTE functions, but many features have been removed to keep this standard as simple as possible to reduce device cost and minimize battery consumption. This optimization includes removing handover, carrier aggregation, measurements to monitor the channel quality, and dual connectivity. NB-IoT operates on the same licensed frequencies used by LTE and employs QPSK and BPSK modulations [4].

ISSN 2395-8618

2.3 LoRaWan

LoRaWAN is an open standard architecture developed by LoRa Alliance. LoRa is a physical layer technology that enables long-range, low data rate, and low power wireless communication. It is an unlicensed band technology that modulates the signals in the sub–GHz ISM band using the spread spectrum technique [4].

Like Sigfox, GPRS, and NB-IoT, the LoRaWAN protocol is based on a star protocol where each device communicates with a base station which relays the information to and from a central server via an IP-based protocol [5].

2.4 GPRS

In the past, most of the applications that required low data rates for a long-range were using cellular networks. This type of network provides the users with many services. Before the emergence of LPWAN technologies, cellular networks had been offering the GSM, GPRS, EDGE, 3G, and 4G technologies. Today, 3G/4G technologies aim to provide users with minimum latency and high data rates for multimedia applications. For this purpose, most of IoT applications were used in the GPRS networks. GPRS is a 2.5G mobile communication that provides a data rate of 56 to 114 kbps with a range up to 26 Km. The primary disadvantages of the GPRS network are the power consumption and high maintenance cost [4].

The GPRS systems have been deployed for many years and serve as the reference for LPWA technology in many markets today. GPRS is the packet radio service built on top of GSM and uses GMSK modulation. It requires a frequency reuse scheme of up to 12, providing an inefficient spectral density. GPRS and NB-IoT operate in the licensed bands and are therefore not restricted by duty cycle or listen before talking limitations [5].

2.5 Sigfox

SigFox uses Ultra-Narrow Band (UNB) modulation with Differential Binary Phase-Shift Keying (DBPSK) at 100 bps. In SigFox, the device initiates a transmission by sending three uplink packages in sequence on three random carrier frequencies. The base station will successfully receive the package even if two of the transmissions are lost due to, e.g., collision with other devices or interference from other systems using the same frequency. The duty cycle restrictions of the utilized sub-band in the 868 MHz EU ISM band is 1 %. Therefore, a SigFox device may only transmit 36 seconds per hour. The time on-air is 6 sec per package, and thus the maximum is six messages per hour with a payload of 4, 8, or 12 bytes. [5]

2.6 Weightless

Weightless is managed by the Weightless-Special Interest Group. Three standards have been proposed by the group, namely Weightless-N, Weightless-W, and Weightless-P. This section will focus on the most recent standard: Weightless-P.

Weightless-P is a non-proprietary physical layer technology. It uses GMSK and QPSK for modulating the signal. These modulating schemes are very well known and are used in various commercial products; hence the end devices do not require a proprietary chipset. Weightless-P divides the Sub-GHz ISM spectrum into 12.5KHz narrow channels, and each channel offers a data rate of 200 bps to 100 Kbps. The firmware of these devices can be upgraded by air using its own wireless link because the bidirectional communication is fully supported. [6].

2.7 RPMA

Formerly known as On-Ramp Wireless, it came up with Random Phase Multiple Access (RPMA), which is a spread-spectrum technology operating on the 2.4GHz ISM band instead of the sub 1GHz bands and leverages more relaxed regulations on the spectrum across different regions.

Like in LoRaWAN, a base station in RPMA's is also capable of receiving transmissions on all the spreading factors. Also, like LoRaWAN adaptive data rate (ADR) technique is employed by the devices, where devices can select optimum spreading factors based on the downlink signal strength.

RPMA uses a form of Viterbi algorithm that allows guaranteed message arrival at the base station even with the Packet Error Rate (PER) as high as 50%, and security is improved using encryption [6].

2.8 NB-Fi

NB-Fi is an LPWAN protocol that supports secure bidirectional communication for IoT, machine-to-machine (M2M), Smart Grid, Smart Utilities, Smart City, and industrial applications. NB-Fi is a protocol that was developed by WAVIoT and designed for secure wireless transmission of small volumes of data over long distances with low energy consumption. NB-Fi is an open standard with the disclosed format of NB-Fi messages and relevant technical data required for manufacturers to produce compatible end-devices.

NB-Fi standard supports up to 4.3 billion devices in a single network with a 32-bit ID for each device. NB-Fi does not use IP addressing (IPv4, IPv6) to optimize the size of the payload. IoT devices such as sensors and gauges can transmit tiny data packages, only a few bytes. As the minimum size of the IP header is 20 bytes, the Non-IP Data Delivery (NIDD) approach allows developing simpler and cheaper devices. Data exchange between devices and third-party applications is possible via the WAVIoT IoT platform's API [7].

2.9 LTE Cat-M1

As one of the advanced wide area network technologies, LTE is regarded as a promising candidate for accommodating a large amount of MTC devices. However, the current infrastructure of LTE networks is built mainly for broadband communication used by smartphones. To support MTC devices, 3GPP has been working on study items for MTC since 2011 [2] and has already undergone several amendments to include new MTC-oriented features based on the existing LTE

ISSN 2395-8618

architecture. By reusing most of the existing LTE technology, it requires only a software upgrade to provide MTC service. One critical enhancement is LTEeMTC (enhanced MTC) introduced in Release 13, which is mainly for low-cost, low-power, low-rate, and delay-insensitive MTC devices. Specifically, Release 13 defines a new user equipment (UE) category, namely CategoryM1 (Cat-M1), to achieve enhanced coverage in reduced bandwidth as low as 1.4MHz. Moreover, to further reduce the cost and complexity of these eMTC devices, Cat-M1 also supports operation with only one receive antenna. The need for low-complexity hardware design also makes coverage enhancement more challenging [8].

For the multicriteria analysis to select the most suitable IoT technology to be implemented in the Telemetry System, only four platforms were included: LoRaWAN, Sigfox, Gprs, Nb-IoT. The other platforms are not available in the country.

3. Methodology

As mentioned before, there are many IoT technologies in the market with different technical characteristics. The selection of one of them for project implementation is not a simple task. There are several factors that should be analyzed, and each factor has its own importance depending on the project's nature. Regarding this, a multicriteria analysis is carried out to select the best alternative.

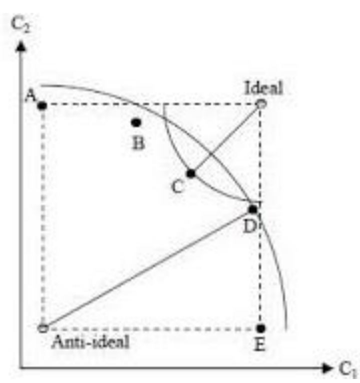

Fig 1. Ideal and Anti-Ideal points in TOPSIS method

Multicriteria analysis (MCA) provides a systematic methodology to integrate heterogeneous and uncertain information with cost-benefit information and stakeholders' views in an understandable framework to rank project alternatives. MCA is highly useful as a tool for project evaluation during the developed phase when decision makers do not have sufficient knowledge regarding details, but the importance of making the right decision is considerable [9].

A vast number of multicriteria decision-making methods have been developed to deal with the problem of ranking a set of alternatives evaluated in a multicriteria fashion. Very often, these methods assume that the evaluation among criteria is statistically independent. However, in actual problems, the observed data may comprise dependent criteria, which, among other problems, may result in biased rankings [10].

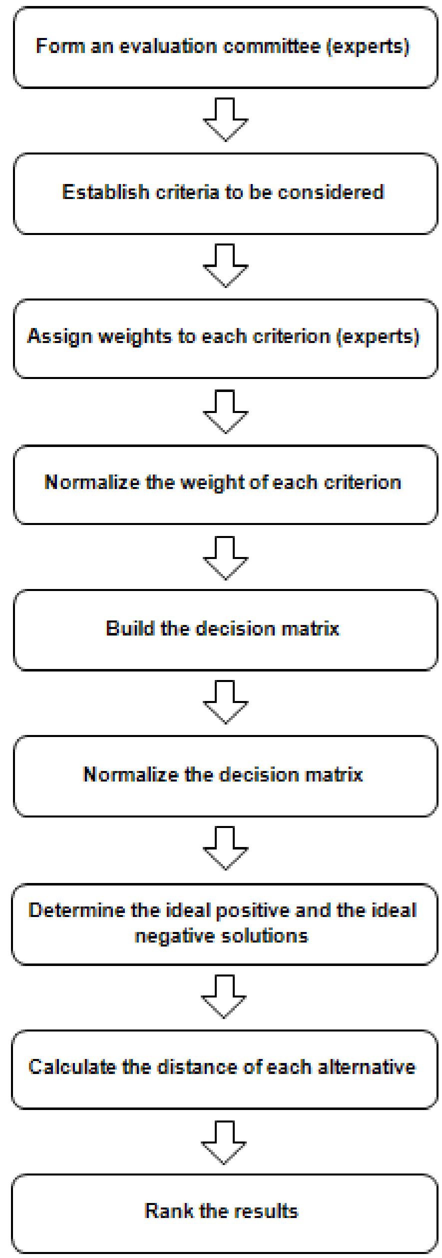

Fig. 2. TOPSIS method

To evaluate IoT technologies and choose the best option, a multicriteria analysis based on TOPSIS was used.
TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) is among the most popular MCDM (Multiple Criteria Decision Making) methods. Decision making is the process of selecting a possible course of action from all the available alternatives. In almost all such problems, the multiplicity of criteria for judging the alternatives is pervasive. That is, for many such problems, the decision maker wants to attain more than one objective or goal in selecting the course of action while satisfying the constraints dictated by environment, processes, and resources [11]. TOPSIS provides a broader principle of compromise for solving this kind of prob-

ISSN 2395-8618

lem [12]; it was originally developed by Hwang and Yoon [13]. Traditionally, this method is applied to ranking problems, where alternatives are evaluated based on Euclidean distances from an ideal and a non-ideal solution [14].

The basic concept of this method is that the selected alternative should have the shortest distance to the positive ideal solution as well as the farthest distance from the negative ideal solution [15].

As an example, Fig. 1 shows five alternatives, A, B, C, D, and E, with a choice of 2 criteria; it also shows the ideal and anti-ideal points. It is obvious that if the usual Euclidean distance (p = 2) is applied with equal weights, point C is the closest to the ideal, and D is the furthest. TOPSIS solves this dilemma in the choice between the ideal and the anti-ideal [16].

For applying the TOPSIS Method, steps could be resume as shown in Fig. 2.

To apply the multicriteria analysis based on TOPSIS, the criteria under which each alternative will be evaluated were first established. According to the expert's opinion, ten criteria were selected and divided into three groups, as shown in Table 1:

TABLE 1
CRITERIA

Number	Groups	Description
1	Technical	Availability. There are three values: "2" if this technology is available; 1 if the technology can be implemented; and 0 if it is not available
2	Technical	Maximum transmission speed (kbps)
3	Technical	Large of data frame per transmission (bytes)
4	Technical	Power consumption in years considering one transmission per day
5	Technical	Nodes per gateway (x1000)
6	Technical	Maximum transmission distance (km, with sightline)
7	Economic	Device cost (US$)
8	Economic	Service/Maintenance cost (US$/transmission/device)
9	Social	Local support (number of suppliers)
10	Social	Government permission (need government license 1, license not needed 0)

Each criterion has a value according to the IoT platform characteristics, as shown in Table 2.

Four experts were asked to assign weights to each criterion, according to their knowledge and experience in this kind of system. They were asked to consider the importance that each criterion has in Telemetry systems for the case study, especially the location. To set the appropriate weight, the Likert Scale was applied.

The Likert scale is widely used in social work research and is commonly constructed with four to seven points, but it can be increased to eleven, a common metric that ranges from 0 to 10. Also, it can be treated as a continuous measure, and hence arithmetic operations can be used [17].

In this research, a scale from 1 to 7 was used, where "7" represents the greatest weight in importance for the criterion. The result was a matrix $W_{4 \times 10}$ with w_{ij} elements, where i represents each expert and j represents the index of the criterion.

A unique weight for each criterion is obtained using the geometric mean [15]:

$$\mathrm{Gj} = \sqrt[4]{\mathrm{w}_{1\mathrm{j}} \times \mathrm{w}_{2\mathrm{j}} \times \mathrm{w}_{3\mathrm{j}} \times \mathrm{w}_{4\mathrm{j}}} \quad (1)$$

Then, the weights should be normalized in order to be compared with each other [15]. For this operation, the sum of all the weights is needed:

$$\mathrm{S} = \sum_{\mathrm{j}=1}^{10} \mathrm{G}_{\mathrm{j}} \quad (2)$$

The normalized weights of each criterion is calculated by:

$$Y_j = \frac{G_j}{S} \quad (3)$$

The idea of TOPSIS can be expressed in a series of steps [18]. The first task of the TOPSIS algorithm consists of creating a decision matrix. At this point, once the value of each criterion (j) corresponding to each alternative (k) is assigned, the result is the matrix X with elements x_{kj}. This is shown in Table 2.

To calculate de normalized decision matrix A, the normalized value a_{kj} is calculated as:

$$A = [a_{kj}]_{m \times n}$$

$$a_{kj} = \frac{x_{kj}}{\sqrt{\sum_{k=1}^{n} (x_{kj})^2}} \quad (4)$$

$$k = 1, 2, 3\ldots, m; \; j = 1, 2, 3, \ldots, n,$$

where x_{kj} represents the x value of alternative k corresponding to criterion j.

Then the Ideal Positive Value (a^+) and the Ideal Negative Value (a^-) should be calculated for each criterion. This is done with these expressions:

$$a^+_j = \max(\mathrm{a}_{\mathrm{kj}}) \text{ where } k = 1,2,3\ldots m \quad (5)$$

$$a^-_j = \min(\mathrm{a}_{\mathrm{kj}}) \text{ where } k = 1,2,3\ldots m \quad (6)$$

ISSN 2395-8618

TABLE 2
ASSIGNED VALUES FOR EACH IOT ALTERNATIVE

Ideal	**max**	**max**	**max**	**max**	**max**	**max**	**min**	**min**	**max**	**min**
Alternative\Criteria	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
LoraWan	1	50	243	15	100	5	2.38	0.39	1	0
GPRS	2	10000	10000	5	200	100	3	2	3	1
Sigfox	0	0.1	12	20	100	10	5.94	1.6	1	0
NB-IoT	0	1000	1600	10	200	1	23.8	0.6	0	1

TABLE 3
ASSIGNED WEIGHTS BY EXPERTS

E\C	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	7	5	5	4	5	3	4	6	2	4
2	7	6	1	3	2	2	3	6	3	2
3	7	7	2	6	4	5	3	6	4	3
4	7	5	1	5	2	4	4	5	3	1
G.M	7	5.692	1.778	4.356	2.991	3.31	3.464	5.733	2.913	2.213
N.W.	0.177	0.144	0.045	0.11	0.076	0.084	0.088	0.145	0.074	0.056

TABLE 4
NORMALIZED DECISION MATRIX

Ideal	**max**	**max**	**max**	**max**	**max**	**max**	**min**	**min**	**max**	**min**
Alternative\Criteria	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
LoraWan	0.45	0.00	0.02	0.55	0.32	0.05	0.10	0.15	0.30	0.00
GPRS	0.89	1.00	0.99	0.18	0.63	0.99	0.12	0.75	0.90	0.71
Sigfox	0.00	0.00	0.00	0.73	0.32	0.10	0.24	0.60	0.30	0.00
NB-IoT	0.00	0.10	0.16	0.37	0.63	0.01	0.96	0.23	0.00	0.71

Calculating the separation measures using the *n*-dimensional Euclidean distance. The separation of each alternative from the positive ideal solution is given as:

$$pa_j = Y_j \,.\, (a^+{}_j - a_j)^2 \quad (7)$$

$$D^+{}_k = \sqrt{\sum_{j=1}^{m}(pa_j)} \quad (8)$$

Similarly, the separation from the negative ideal solution is given as:

$$na_j = Y_j \,.\, (a^-{}_j - a_j)^2 \quad (9)$$

$$D^-{}_k = \sqrt{\sum_{j=1}^{m}(na_j)} \quad (10)$$

For each “distance,” D^+ and D^-, a_j represents the normalized value "*a*" for each criterion "*j*."

For each alternative, calculate the ratio R_k as:

$$R_k = D^-{}_k \,/\, (D^-{}_k + D^+{}_k) \quad (11)$$

Finally, alternatives should be ranked in increasing order according to the radio R_k.

4. RESULTS

For the case study, four experts assigned importance values (weights) for each criterion according to a 1-7 Likert Scale. The results are shown in Table 3.

Each row represents the expert's assigned values, and each column represents each criterion. GM row is the Geometric Mean, and NW row corresponds to the Normalized Weight according to equation 3.

The assigned values for each IoT alternative were shown before in Table 2.

Table 4 shows the normalized decision matrix A, according to equation 4 for each element of the matrix.

Table 5 shows the Positive Ideal and the Negative Ideal for each criterion.

To obtain D^+, the corresponding factors were calculated and are shown in Table 6:

In this sense, D^+ for each alternative is shown in Table 8:

ISSN 2395-8618

TABLE 5
POSITIVE IDEAL AND NEGATIVE IDEAL FOR EACH CRITERION

Ideal	**max**	**max**	**max**	**max**	**max**	**max**	**min**	**min**	**max**	**min**
Criteria	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Positive Ideal	0.89	1.00	0.99	0.73	0.63	0.99	0.10	0.15	0.90	0.00
Negative Ideal	0.00	0.00	0.00	0.18	0.32	0.01	0.96	0.75	0.00	0.71

TABLE 6
D+ FACTORS

Ideal	**max**	**max**	**max**	**max**	**max**	**max**	**min**	**min**	**max**	**min**
Alternative\Criteria	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
LoraWan	0.04	0.14	0.04	0.00	0.01	0.07	0.00	0.00	0.03	0.00
GPRS	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.05	0.00	0.03
Sigfox	0.14	0.14	0.04	0.00	0.01	0.07	0.00	0.03	0.03	0.00
NB-IoT	0.14	0.12	0.03	0.01	0.00	0.08	0.07	0.00	0.06	0.03

TABLE 7
D- FACTORS

Ideal	**max**	**max**	**max**	**max**	**max**	**max**	**min**	**min**	**max**	**min**
Alternative\Criteria	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
LoraWan	0.04	0.00	0.00	0.01	0.00	0.00	0.07	0.05	0.01	0.03
GPRS	0.14	0.14	0.04	0.00	0.01	0.08	0.06	0.00	0.06	0.00
Sigfox	0.00	0.00	0.00	0.03	0.00	0.00	0.05	0.00	0.01	0.03
NB-IoT	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.04	0.00	0.00

TABLE 8
D+ FOR EACH IOT ALTERNATIVE

Alternative	D^+
LoraWan	0.5759
GPRS	0.3384
Sigfox	0.6798
NB-IoT	0.7344

TABLE 9
D- FOR EACH IOT ALTERNATIVE

Alternative	D^-
LoraWan	0.451396
GPRS	0.734482
Sigfox	0.342476
NB-IoT	0.232531

D^- is obtained in the same way. Table 7 shows the corresponding factors that were calculated according to equation 9.

Table 9 shows D^- for each IoT alternative.

Finally, R is calculated according to equation 11, and the results are shown in Table 10.

As shown, applying the TOPSIS method, the best alternative for the Telemetry System is GPRS which is the one that has more suppliers in Peru and is more developed around the country. The second one is LoraWan, which low-cost and low power consumption equipment make this platform a good alternative for smaller systems.

TABLE 10
R FOR EACH IOT ALTERNATIVE

Alternative	R	Rank
LoraWan	0.439414	2
GPRS	**0.684608**	**1**
Sigfox	0.335029	3
NB-IoT	0.240487	4

On the other hand, the worst alternative is NB-IoT, which despite its technical characteristics, it is still being developed in the country.

5. CONCLUSIONS

The selection of an IoT technology for a Telemetry System depends on many factors. The most important element is platform availability at the location of the system. Other important factors are the involved costs and the government permissions for communication networks.

A multicriteria analysis was done considering not only the criteria that could have an important impact on the system implementation but also the opinion of experts with experience in the case study (similar systems and locations).

Accordingly to the analysis, the suitable IoT technology for a Telemetry System in Huacho, Peru, is GPRS. This result is consistent with the "experts" previous thoughts, which selection was the same. They considered network availability, costs, and the number of suppliers to make their decision

ISSN 2395-8618

(supported by their experience). However, the TOPSIS method offers an objective analysis that supports this selection.

As shown before, one of the most important factors of this method is the “experts” opinion. It is necessary to build an “evaluation committee” with enough knowledge and experience to achieve good results.

It is demonstrated that this analytical method works and could be scalable to another kind of system or to solve decision problems considering multiple criteria. For future research, as soon as more IoT alternatives are available at the location, they could be evaluated using the same methodology. Furthermore, this kind of analysis could be done to other system problems like equipment maintenance, choice of suppliers, etc.

References

[1] Matt Burgess, "What is the Internet of Things?" WIRED explains, http://www.wired.co.uk/article/internet-of-things-what-is-explained-iot

[2] H. Hejazi, H. Rajab, T. Cinkler and L. Lengyel, "Survey of platforms for massive IoT," 2018 IEEE International Conference on Future IoT Technologies (Future IoT), Eger, 2018, pp. 1-8, doi.org/10.1109/FIOT.2018.8325598

[3] A. A. Ismail, H. S. Hamza and A. M. Kotb, "Performance Evaluation of Open Source IoT Platforms," 2018 IEEE Global Conference on Internet of Things (GCIoT), Alexandria, Egypt, 2018, pp. 1-5, doi.org/10.1109/GCIoT.2018.8620130.

[4] W. Ayoub, A. E. Samhat, F. Nouvel, M. Mroue and J. Prévotet, "Internet of Mobile Things: Overview of LoRaWAN, DASH7, and NB-IoT in LPWANs Standards and Supported Mobility," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1561-1581, Second quarter 2019, doi.org/10.1109/COMST.2018.2877382.

[5] B. Vejlgaard, M. Lauridsen, H. Nguyen, I. Z. Kovacs, P. Mogensen and M. Sorensen, "Coverage and Capacity Analysis of Sigfox, LoRa, GPRS, and NB-IoT," 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, NSW, Australia, 2017, pp. 1-5, doi.org/10.1109/VTCSpring.2017.8108666.

[6] S. Aggarwal and A. Nasipuri, "Survey and Performance Study of Emerging LPWAN Technologies for IoT Applications," 2019 IEEE 16th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT and AI (HONET-ICT), Charlotte, NC, USA, 2019, pp. 069-073, doi.org/10.1109/HONET.2019.890811

[7] Satya N. Gupta; Alice Maltseva, Keshav Sharma. Collaborating for Self-Reliance–Co-Creating IoT Connectivity Solutions Using NB. Fi.

[8] P. Hsieh, Y. Jia, D. Parra and P. Aithal, "An Experimental Study on Coverage Enhancement of LTE Cat-M1 for Machine-Type Communication," 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, 2018, pp. 1-5, doi.org/10.1109/ICC.2018.8422888

[9] Karleuša B, Hajdinger A, Tadić L. The Application of Multi-Criteria Analysis Methods for the Determination of Priorities in the Implementation of Irrigation Plans. Water. 2019; 11(3):501. doi.org/10.3390/w11030501

[10] Guilherme Dean Pelegrina, Leonardo Tomazeli Duarte, João Marcos Travassos Romano, "Application of independent component analysis and TOPSIS to deal with dependent criteria in multicriteria decision problems", Expert Systems with Applications, Volume 122, 2019, Pages 262-280, ISSN 0957-4174, doi.org/10.1016/j.eswa.2019.01.008

[11] Lei, F., Wei, G., Gao, H. et al. TOPSIS Method for Developing Supplier Selection with Probabilistic Linguistic Information. Int. J. Fuzzy Syst. 22, 749–759 (2020). doi.org/10.1007/s40815-019-00797-6

[12] K. Paul Yoon, Won Kyung Kim, "The behavioral TOPSIS", Expert Systems with Applications, Volume 89, 2017, Pages 266-272, ISSN 0957-4174, doi.org/10.1016/j.eswa.2017.07.045.

[13] Hwang CL., Yoon K. (1981) Methods for Multiple Attribute Decision Making. In: Multiple Attribute Decision Making. Lecture Notes in Economics and Mathematical Systems, vol 186. Springer, Berlin, Heidelberg. doi.org/10.1007/978-3-642-48318-9_3

[14] Diogo Ferreira de Lima Silva, Adiel Teixeira de Almeida Filho, "Sorting with TOPSIS through boundary and characteristic profiles", Computers & Industrial Engineering, Volume 141, 2020, 106328, ISSN 0360-8352, doi.org/10.1016/j.cie.2020.106328.

[15] Husam Jasim Mohammed, Maznah Mat Kasim, and Izwan Nizal Mohd Shaharanee, "Selection of suitable e-learning approach using TOPSIS technique with best ranked criteria weights", AIP Conference Proceedings 1905, 040019 (2017); doi.org/10.1063/1.5012207

[16] YILMAZ, Baris; HARMANCIOGLU, N. Multi-criteria decision making for water resource management: a case study of the Gediz River Basin, Turkey. Water SA, 2010, vol. 36, no 5.

[17] Huiping Wu & Shing-On Leung (2017) Can Likert Scales be Treated as Interval Scales?—A Simulation Study, Journal of Social Service Research, 43:4, 527-532, doi.org/10.1080/01488376.2017.1329775

[18] Lifeng Ren, Yanqiong Zhang, Yiren Wang, Zhenqiu Sun, Comparative Analysis of a Novel M-TOPSIS Method and TOPSIS, Applied Mathematics Research eXpress, Volume 2007, 2007, abm005, doi.org/10.1093/amrx/abm005

ISSN 2395-8618

Forecasting the Demand of Parts in an Assembly Plant Warehouse Using Time-Series Models

G. Rivera, R. Florencia-Juárez, J. P. Sánchez-Solís, V. García, and C. D. Luna

***Abstract*—Knowing the demand for products in advance would be ideal for companies that strategically relocate products in their warehouses to facilitate the picking process, which is the most expensive activity in a warehouse. In this sense, an assembly plant from Ciudad Juárez, Chihuahua, Mexico, that handles approximately 10,477 parts in its inventory, periodically relocates them in warehouse zones to facilitate the picking process. However, relocation is done empirically based on the total number of outbound inventory movements of each of the parts made in a given time. This paper describes the implementation of time-series models to forecast the demand for parts that could improve the relocation process. For this purpose, different Holt-Winters Seasonal and SARIMA models were implemented. For the implementation of the SARIMA models, the Box-Jenkins methodology was followed. The AIC and BIC metrics were used to identify the best Holt-Winters Seasonal model and the best SARIMA model. Tests were performed on the residual series to check that model is fit to the data. The RMSE and MAPE metrics were used to evaluate the performance of Holt-Winters Seasonal and SARIMA models. The results of the evaluation carried out indicate that the SARIMA model outperforms to Holt-Winters Seasonal model.**

***Index Terms*—Forecast demand, Holt-Winters Seasonal model, Seasonal Auto-Regressive Integrated Moving Average model, Time Series, Box-Jenkins methodology.**

1. Introduction

CIUDAD Juárez is a city whose economy is strongly based on the assembly plant industry. It has the largest number of assembly plants operating under the IMMEX (*Industria Manufacturera, Maquiladora y de Servicios de Exportación*) scheme in the entire state of Chihuahua with 330 assembly plants (65%), followed by the city of Chihuahua with 109 (22%) and the other municipalities with 66 (13%). At the national level, it is the second city with 330 assembly plants under this scheme. Among the different sectors, the Automotive, the Electronics, the Medical, the Plastics Metals, the Call Center, and the Packaging predominate [1].

In an assembly plant of this city dedicated to the manufacture of smoke detectors, fire alarms, among other products, the raw material used is stored in a warehouse within the same plant.

Currently, the warehouse is divided into four zones: A, B, C, and D. These zones are strategically organized to minimize the time for picking parts. The parts with the highest demand are stored in bins located in zone A, close to the production area. On the contrary, the parts with the lowest demand, either because they are obsolete or because they were never used, are located in zone D. It is important to note that, based on the parts demand, these are periodically relocated to the warehouse zones.

The criteria used by the assembly plant to locate each of the parts in the warehouse zones is based on the cumulative distribution of the demand they present. The demand is defined according to the outbound inventory movement of each of the parts. The parts that represent a cumulative distribution of 70% of the demand are located in zone A. The parts that are located in zone B represent 25% of the demand. In zone C are located the parts that represent 5% of the demand. Finally, in zone D, the parts that did not have movement are located.

Although the parts with the highest demand are identified in this empirical way, the trend of movements is not considered when assigning their location in the warehouse. For example, following this criterion, the part with the highest number of movements should be located in zone A. However, if the trend of its movements shows a considerable decrease in recent months, perhaps that part should not be located in zone A but zone B.

Failure to properly locate the parts in warehouse zones that minimize the picking time could cause the delivery time of these to the production area to be long; even the production lines could be stopped. For this reason, knowing the future demand of the parts could help improve the relocation process.

In this sense, time series models have been successfully applied in various sectors of the supply chain to forecast demand. In [2], different ARIMA models were implemented following the Box-Jenkins methodology to model and forecast the demand in a food company. Using the Akaike, Schwarz Bayesian, maximum likelihood, and standard error criteria, they identified the best model was an ARIMA (1,0,1). The authors note that forecasting future demand affects the supply chain and provides reliable guidelines for decision making.

In [3], different time series models applied to forecasting

Manuscript received on October 10, 2020, accepted for publication on November 22, 2020, published on December 30, 2020.

G. Rivera, R. Florencia-Juárez, J. P. Sánchez-Solís, and V. García are with the Department of Electrical and Computer Engineering, Autonomous University of Cd. Juárez, Cd. Juárez, 32310, Mexico (e-mail: {gilberto.rivera, rogelio.florencia, julia.sanchez, vicente.jimenez}@uacj.mx).

C. D. Luna is with the Applied Computer Science Graduate Program at Autonomous University of Cd. Juárez, Cd. Juárez, 32310, Mexico (e-mail: al199185@alumnos.uacj.mx).

ISSN 2395-8618

the workload in a zone picking warehouse were analyzed. A real case study is presented in which they demonstrate the value of applying time series forecasting models to forecast the daily number of order lines. They mention that improving workload planning can contribute to an effective and efficient order-picking process.

In [4], the demand for an inventory of historical data of Panadol 650mg was optimized using forecasting techniques. The forecast models single moving average, single exponential smoothing, double moving average, double exponential smoothing, regression, Holt-Winters additive, Seasonal additive, Holt-Winters multiplicative, seasonal multiplicative, and ARIMA were compared using the metrics RMSE, MSE, MAD, and MAPE. The authors report that the best forecasting technique was regression analysis.

In [5], the combination of Bootstrap aggregating and Holt-Winters methods was proposed to forecast demand in the air transport industry. They compared their proposal against the SARIMA, Holt-Winters, ETS, Bagged.BLD.MBB.ETS and Seasonal Naive models using Symmetric Mean Absolute Percentage Error (sMAPE). The authors report that their proposal provided more accurate forecasts.

This paper describes the implementation of time-series models to forecast the demand for inventory parts, which could improve the relocation process in warehouse zones and facilitate the picking process. The characteristics of the time series were analyzed, and the Augmented Dickey-Fuller and Kwaitowski-Phillips-Schmidt-Shin tests were applied to determine whether the series is stationary. The Holt-Winters Seasonal and a modification of ARIMA models, called SARIMA, were selected because of the characteristics of the time series. The Box-Cox transformation was applied to the time series to fit SARIMA models. Additionally, the time series was differentiated to transform it into a stationary series. The SARIMA model that best fits the time series was selected based on the Akaike Information Criteria and Bayesian Information Criteria metrics. Tests were applied to the SARIMA model to see if the residual series is white noise.

The paper is organized in the following way: Section 2 provides a background of the concepts related to time series models. Section 3 presents the preprocessing performed on the data provided by the warehouse to be used by the time-series models. Section 4 shows the empirical and statistical analysis applied to the data for studying its features. Section 5 shows the selection process of the time series model. Section 6 shows the results and discussions. Finally, Section 7 presents the conclusions of this work.

2. Time Series Background

According to Chris Chatfield, "*a time series is a set of observations measured sequentially through time*" [6]. Analyzing time series is usually very useful to find relationships or trend patterns in the variable to be analyzed. Applying time-series models allows forecasting the future behavior of the variable to be analyzed based on its historical data.

A time series consists of four components [7], [8]:

1. *Trend*. It refers to the behavior of the time series. Its direction and slope can remain constant or change throughout the series. It can be linear, stochastic, or deterministic. When it remains stagnant, it means there is no trend.
2. *Seasonality*. It is when the series presents regular seasonal patterns, that is, periodic variations, such as annual, monthly, quarter changes, etc. These patterns are generally of a fixed nature; that is, they are associated with some calendar aspect (for example, days, months, years).
3. *Cycles*. It represents a pattern that repeats itself over time. It is similar to the seasonal component; however, the cyclical effect patterns are not of fixed duration; that is, their duration varies from one cycle to another.
4. *Irregular variation*. It represents random noise, that is, the residual variations of the series after being divided into the three previous components. When this component is completely random (i.e., not autocorrelated), it is known as white noise, which is analyzed to estimate the adequacy of a time-series model.

A time series can be either stationary or non-stationary. It is stationary when it does not have a trend or seasonal effects and its mean or variance are consistent over time. It is worth mentioning that the statistical modeling methods require that the time series be stationary since they are easier to model. Non-stationary series exhibit seasonal patterns and trends, and their mean and variance change over time. Statistical modeling methods require making the data stationary, removing these components.

There are different time-series models such as (a) Simple Exponential Smoothing for data without trend or seasonality, (b) Holt's Linear Trend model for data with a trend but no seasonality, (c) Holt-Winters Seasonal model for data with trend and/or seasonality, (d) Autoregressive Integrated Moving Average (ARIMA) models (also known as Box–Jenkins models), among others. As can be imagined, choosing a time-series model depends on the features of the time series, which is crucial for achieving the expected results. Of the aforementioned models, Holt-Winters Seasonal and a modification of ARIMA, called SARIMA, were implemented in this paper.

3. Data Preprocessing

The database provided by the assembly plant contains the inventory movements generated from the year 2017 to September 2020 of the parts handled by the warehouse. When a part is used to fulfill a customer service order, an outbound inventory movement is recorded with the date the movement was made, thus forming a time series for each part.

ISSN 2395-8618

The records in this database are made up of the attributes: *material name*, *description*, *type of movement*, *plant*, *location*, *quantity*, and *PostingDate*. Of these attributes, PostingDate was chosen as a period for the time-series model because it represents the date on which the movements were made. Additionally, outbound inventory movements of each part were counted to represent the value of the independent variable.

Due to the way the warehouse operates, the count of outbound inventory movements was grouped monthly to form the periods for each of the parts. The monthly movements of each of the parts were stored in different .csv files. It is worth mentioning that so far, the time-series models described in this paper have only been implemented in the part with the highest demand in the historical data. Figure 1 shows the histogram of movements of the part with the highest demand in the historical data.

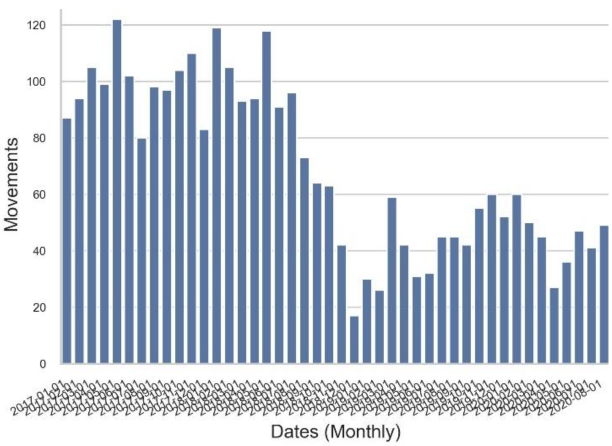

Fig. 1. Histogram of the part with the highest demand in the historical data.

4. DATA ANALYSIS

Before selecting the time-series model, it was determined whether the time series data exhibits trend and seasonal effects. To determine if the data exhibits trend and seasonal effects, the following analyzes were performed: (a) empirical visual analysis and (b) statistical tests.

4.1 Empirical visual analysis

Figure 2 shows the time series corresponding to the part with the highest number of movements in the warehouse inventory. At first glance, it is difficult to determine whether the data is seasonal.

In order to visually perform an exploratory analysis of the time series, it was decomposed into the trend (T_t), seasonality (S_t), and residual (E_t) components, shown in Figures 3 to 5, respectively. This decomposition allows better identification of the underlying patterns of the time series. A classical additive decomposition was used because the seasonal variation has the same magnitude over time where each value (Y_t) of the time series is the sum of these three components, as shown in Equation (1):

$$Y_t = T_t + S_t + E_t \quad (1)$$

In Figure 3 is observed that the trend increases or decreases based on time; that is, the mean changes over time. In Figure 4, a seasonal effect is perceived; that is, a repeated pattern is shown. Based on the above, it is suspected that the series could be non-stationary. Also, the residuals displayed in Figure 5 do not show a stable pattern and are not close to 0.

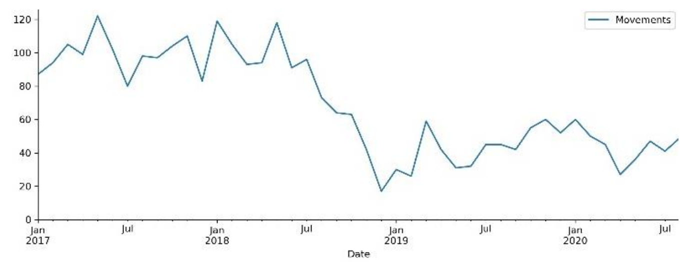

Fig. 2. Time series of the part with the highest number of movements.

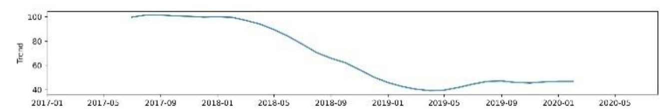

Fig. 3. Trend component.

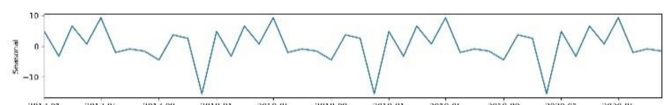

Fig. 4. Seasonal component.

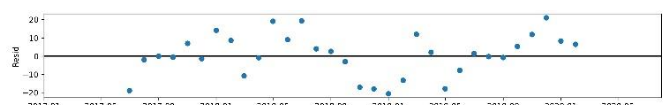

Fig. 5. Residual component.

4.2 Statistical tests to determine seasonality

A non-stationary time series shows seasonal effects and trends. If a time series is non-stationary, it means that the mean and the variance are not constant; that is, they change over time. Some forecasting models are susceptible to non-stationary time series. For this reason, before applying a time-series model to a series, it is necessary to know if the time series is non-stationary.

To reinforce the empirical analysis presented in Section 4.1, the Augmented Dickey-Fuller (ADF) test [9] was used to determine if a series is non-stationary. Additionally, the Kwai-towski-Phillips-Schmidt-Shin (KPSS) test [10] was also applied to determine if the series is stationary.

ADF is a statistical test known as the unit root test. The purpose of the test is to determine the strength with which a time series is defined by a trend. The null hypothesis of the

ISSN 2395-8618

test is that there is a unit root in the time series, that is, that the series is not stationary. The alternative hypothesis of the test is that there is no unit root in the time series, that is, that the series is stationary.

If the *p-value* ≥ 0.05, the test fails to reject the null hypothesis (the series has a unit root, that is, it is non-stationary). If the *p-value* < 0.05, the test rejects the null hypothesis (the series does not have a unit root, that is, it is stationary).

After applying the ADF test, the computed *p-value* was 0.650, so fail to reject the null hypothesis, and it is confirmed that the time series is non-stationary.

On the other hand, the KPSS test is a statistical test used to analyze the stationarity of a series based on a deterministic trend. The null hypothesis of the test is that the time series is stationary. The alternative hypothesis of the test is that the time series is not stationary.

If the *p-value* ≥ 0.05, fail to reject the null hypothesis (the series is stationary). If the *p-value* < 0.05, reject the null hypothesis (the series is non-stationary).

After applying the KPSS test, the computed *p-value* was 0.069, so it fails to reject the null hypothesis, and it is confirmed that the time series is stationary.

Because the ADF test confirms that the series is non-stationary and the KPSS test indicates that it is stationary, the series had to be transformed as described in Section 5.2.

5. Time series model selection

Because the time series exhibits a pattern in the seasonal component, shown in Figure 4, forecasting models that consider seasonality were considered. The models considered were *Holt-Winters Seasonal* and *Seasonal ARIMA* (SARIMA). These models were implemented through the *StatsModels* library in the Python language. Statsmodels is a suite of high-level statistical models that allows for statistical tests and data exploration [11].

To determine the best forecasting model, the metrics commonly used in classical time series, *Akaike Information Criteria* (AIC) [12] and *Bayesian Information Criteria* (BIC) [13], were considered.

5.1 Holt-Winters Seasonal model

The Holt-Winters model [14] that extends the Holt model is used to predict time series data that contain trends and seasonal components. It contains three smoothing equations for the: *level* of the series, *trend*, and *seasonal* components.

There are two versions of this model described in [7], [15], [16], whose difference lies in how the seasonality is modeled, *additive* and *multiplicative* versions. The additive version is suitable when variations are independent of the level and the multiplicative version when seasonal variations are changing proportionally to the level of the series.

Analyzing the seasonal fluctuations shown in Figure 2, it was not observed that these increase in magnitude with the level of the series. For this reason, the additive version was used. Equation (2) is applied to smooth the level of the series. Equation (3) to smooth the trend. Equation (4) to smooth the seasonal components. Equation (5) to forecast the next *h* values of the time series.

$$L_t = \alpha(Y_t - S_{t-M}) + (1-\alpha)(L_{t-1} + T_{t-1}) \quad (2)$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \quad (3)$$

$$S_t = \gamma(Y_t - L_t) + (1-\gamma)S_{t-M} \quad (4)$$

$$Y_{t+h} = L_t + (h)T_t + S_{t-M-h} \quad (5)$$

where:

- Y_{t+1} is the forecast value at time $t+1$.
- L_t is the level at time t.
- T_t is the trend value at time t.
- S_t is the seasonal value at time t.
- α, β, γ are smoothing parameters.
- h is the forecast horizon.
- M is the number of seasonal periods;

$0 \le \alpha \le 1;\ 0 \le \beta \le 1;\ 0 \le \gamma \le 1.$

Different combinations of values in the *smoothing_level* (SL), *smoothing_slope* (SSL), *smoothing_seasonal* (SSE), and *seasonal_period* (SP) parameters were tested using brute force to identify the best model. The values used in each parameter were 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. The time used in this experimentation was 343.89 seconds using a computer with an Intel Corte i7-8650U @ 1.90GHz 2.11GHz processor, 16 GB RAM, and Windows 10 pro 64-bits operating system. Table 1 shows the three best models that obtained the lowest values in the AIC and BIC metrics. Of these, the best model was Holt-Winters Seasonal $(0.8, 0.1, 0.0)_{12}$ with AIC=212.880 and BIC=238.217 whose parameters were SL=0.8, SSL=0.1, SSE=0.0 and SP=12.

TABLE 1
Best Holt-Winters Seasonal Additive models

Model $(SL,SSL,SSE)_{SP}$	AIC	BIC
$(0.8, 0.1, 0.0)_{12}$	**212.880**	**238.217**
$(0.9, 0.1, 0.0)_{12}$	212.899	238.235
$(0.7, 0.2, 0.0)_{12}$	212.968	238.304

5.2 ARIMA models

In addition to smoothing techniques, ARIMA models were also used.

ARIMA (Auto-Regressive Integrated Moving Average) models are formed by three components: *AR* for the autoregressive component, *I* for the integrated component, and *MA* for the moving average component. It is usually expressed with the notation: ARIMA (*p*, *d*, *q*) where *p* represents the number of correlated lags to be included in the AR component, *d* the number of times that the raw observations were differenced, and *q* the size of the moving average window [8]. The parameters *p*, *d*, and *q* are non-negative integers that

ISSN 2395-8618

indicate the order of the different components of the model. The AR (p) and MA (q) components are predictors that explain the autocorrelation, and the I component indicates the order of differentiation that has been applied to the time series to leave the series stationary, since before including AR or MA terms, the series must be stationary.

SARIMA models are a modification of ARIMA to support a seasonal time series. They are usually expressed as ARIMA (p, d, q)×(P, D, Q)s where p, d, and q are the non-seasonal parameters, P, D, Q the seasonal component of the time series, and s is the seasonal periodicity. SARIMA models are defined from Equations (6) to (10).

$$\varphi_p(B)\phi_P(B^s)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \quad (6)$$

$$\varphi_p(B) = 1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p \quad (7)$$

$$\varphi_p(B^s) = 1-\phi_s B^s-\phi_{2s} B^{2s}-\cdots-\phi_{Ps} B^{Ps} \quad (8)$$

$$\theta_q(B) = 1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q \quad (9)$$

$$\Theta_Q(B^s) = 1-\Theta_s B^s-\Theta_{2s} B^{2s}-\cdots-\Theta_{Qs} B^{Qs} \quad (10)$$

where:

$\varphi_p(B)$: Polynomial of order p that represents the non-seasonal autoregressive component.

$\phi_P(B^s)^d$: Polynomial of order P that represents the seasonal autoregressive component.

$\theta_q(B)$: Polynomial of order q that represents the non-seasonal moving average component.

$\Theta_Q(B^s)$: Polynomial of order Q that represents the seasonal moving average component.

ε_t: An independently distributed random variable (white noise).

$(1-B^s)^D$: D^{th} seasonal difference of season s.

To implement SARIMA models, the four steps of the Box-Jenkins methodology described in [17] were followed.

***Step 1*. Stationarity of the data**

Because the ADF test indicated that the series is not stationary and the KPSS test indicated the opposite, the Box-Cox transformation was applied to stabilize the variance of the time series. Subsequently, the series was differentiated to eliminate the trend and seasonality, resulting in a stationary time series, shown in Figure 6. Compared with Figure 2, the variance decreased, and the trend is constant; however, irregular fluctuations are observed between the months of November 2018 and March 2019.

When both tests were reapplied, the *p-value* of the ADF test was 0.000, so the null hypothesis was rejected, indicating that the series is stationary. The *p-value* of the KPSS test was 0.100, so the null hypothesis was not rejected, indicating that the series is stationary.

***Step 2*. Identification of the model**

Subsequently, the correlogram of the series was generated to obtain the autocorrelation function (ACF) and the partial autocorrelation function (PACF). ACF and PACF plots are used to determine the appropriate values of p and q and their seasonal equivalent P and Q of the possible candidate models since more than one model could be considered from these values. It should be mentioned that in this step, subjectivity is presented when selecting the best model since the selection is based on the interpretation of the ACF and PACF plots. Figures 7 and 8 show the ACF and PACF plots of the series, respectively.

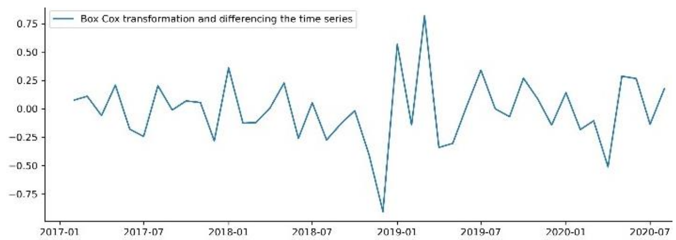

Fig. 6. Box-Cox transformation and differencing the time series.

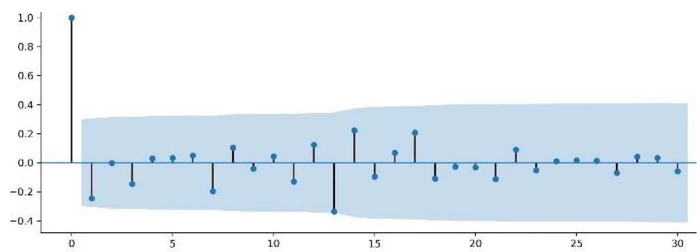

Fig. 7. Autocorrelation function plot.

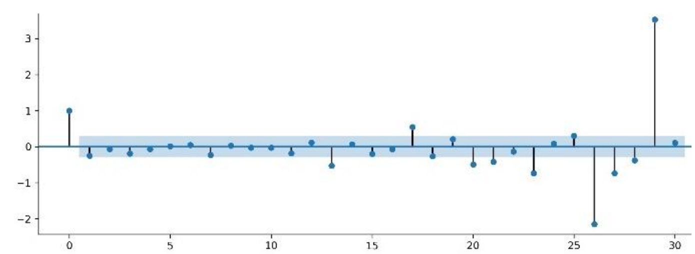

Fig. 8. Partial autocorrelation function plot.

It can be seen in the ACF plot shown in Figure 7 that the first lag is significant, so MA (1) component is suggested, that is, set q = 1. In the PACF plot shown in Figure 8, it can be observed that the first lag is significant, so AR (1) component is suggested, that is, set p = 1. Since the series was differentiated only once, set d = 1.

To define the seasonal component, it can be observed in the ACF plot shown in Figure 7 that lags 0 and 13 are significant, so set Q = 1 is suggested. In the PACF plot shown in Figure 8, it can be observed that only lags 1, 13, and 26 are significant, so set P = 1 is suggested. Finally, set D = 0 and set the seasonal period s = 12 or 13. Therefore, SARIMA $(1,1,1)x(1,0,1)_{12}$ or SARIMA $(1,1,1)x(1,0,1)_{13}$ was identified as an initial model.

ISSN 2395-8618

```
                                  SARIMAX Results
==============================================================================
Dep. Variable:                                  y   No. Observations:                   44
Model:             SARIMAX(0, 1, 0)x(0, 0, [1], 13)   Log Likelihood                  -5.130
Date:                            Sun, 29 Nov 2020   AIC                             14.260
Time:                                    16:29:57   BIC                             17.782
Sample:                                01-01-2017   HQIC                            15.559
                                     - 08-01-2020
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.S.L13      -0.3779      0.242     -1.563      0.118      -0.852       0.096
sigma2         0.0709      0.012      5.925      0.000       0.047       0.094
===================================================================================
Ljung-Box (Q):                       23.53   Jarque-Bera (JB):                 8.33
Prob(Q):                              0.98   Prob(JB):                         0.02
Heteroskedasticity (H):               1.27   Skew:                            -0.10
Prob(H) (two-sided):                  0.66   Kurtosis:                         5.15
===================================================================================
```

Fig. 9. Summary of the SARIMA model (0,1,0) × $(0,0,1)_{13}$.

Based on the values determined from the ACF and PACF plots, different SARIMA models were tested to find the best fitting model by varying the values of their parameters between 0 and 2 for seasonal periods s = 12 and 13. This was because calculating all the possible combinations of values of the parameters of a SARIMA model by brute force can be computationally costly.

Table 2 shows the three best SARIMA models that obtained the lowest values in the AIC and BIC metrics. Of these, the best model was SARIMA (0,1,0) x $(0,0,1)_{13}$ with an AIC = 14.260 and BIC = 17.782. This model does not contain non-seasonal components, nor does it contain seasonal autoregressive components, so it only consists of one seasonal moving average component.

TABLE 2
BEST SARIMA MODELS

Model	AIC	BIC
(0,1,0) × $(0,0,1)_{13}$	**14.260**	**17.782**
(0,1,0) × $(1,0,0)_{13}$	14.487	18.010
(0,1,1) × $(0,0,1)_{13}$	14.644	19.928

Step 3. Parameter estimation

In this step, the coefficient of the seasonal moving average component of the SARIMA model (0,1,0) x $(0,0,1)_{13}$ was estimated. The estimation was performed with the *Summary* function of the trained model, which is an object of the *SARIMAX* class from the *statsmodels.tsa.statespace.sarimax* library. The parameter was estimated using the *log-likelihood* for the *maximum likelihood* estimation. The calculated coefficient was $\Theta_1 = -0.3779$. Figure 9 presents the summary of the model.

Step 4: Diagnostic Checking

After adjusting the SARIMA model, tests were applied to identify if the residual series is white noise, that is, that the residual errors do not show autocorrelation with time and that they are normally distributed. This gives evidence that the fitted model is suitable for forecasting. If the tests gave evidence that the residual series was not white noise, it would indicate the need to improve the model further.

For this purpose, the *plot_diagnostics* function was used, which presents four graphs: a) standardized residual, b) histogram, c) normal Q-Q, and d) correlogram. Figures 10 to 13 show the plots generated by this function.

The standardized residual plot in Figure 10 indicates the absence of trend and constant variance with a mean approaching zero. The histogram plot in Figure 11 shows that the residuals follow the bell curve distribution (normal distribution). The normal Q-Q plot in Figure 12 indicates that most residuals fit a straight line which suggests that they are normally distributed. Finally, the ACF plot in Figure 13 indicates that the residuals are not autocorrelated with time.

Additionally, the Ljung-Box Q [18] and Heteroskedasticity test were applied. The null hypothesis of the Ljung-Box Q test is that the data is distributed independently. The alternative hypothesis is that the data is not distributed independently. After applying the Ljung-Box test, the computed *p-value* was 0.98, so fail to reject the null hypothesis, and it is confirmed that the data is distributed independently. The null hypothesis of the Heteroskedasticity test is that the variance of the residual series is constant, that is, not heteroskedasticity. The alternative hypothesis is that heteroskedasticity exists. After applying the Heteroskedasticity test, the computed *p-value* was 0.66, so fail to reject the null hypothesis, and it is confirmed no heteroskedasticity.

ISSN 2395-8618

Fig. 10. Standarized residual plot of the SARIMA model (0,1,0) × (0,0,1)$_{13}$.

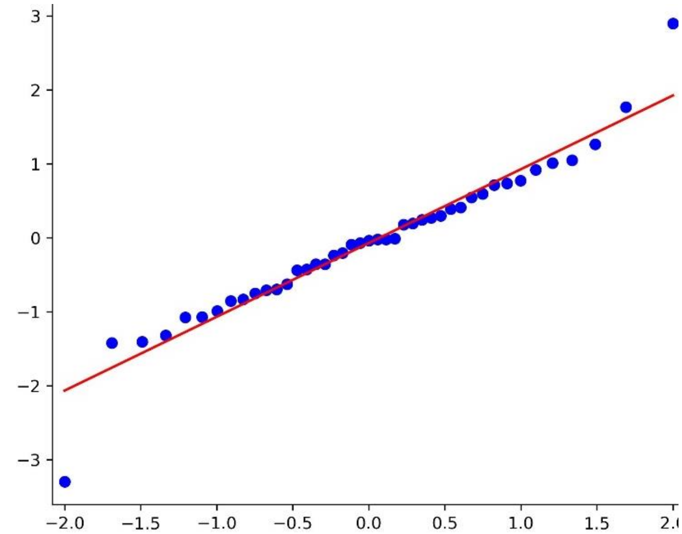

Fig. 12. Normal QQ plot of the SARIMA model (0,1,0) × (0,0,1)$_{13}$.

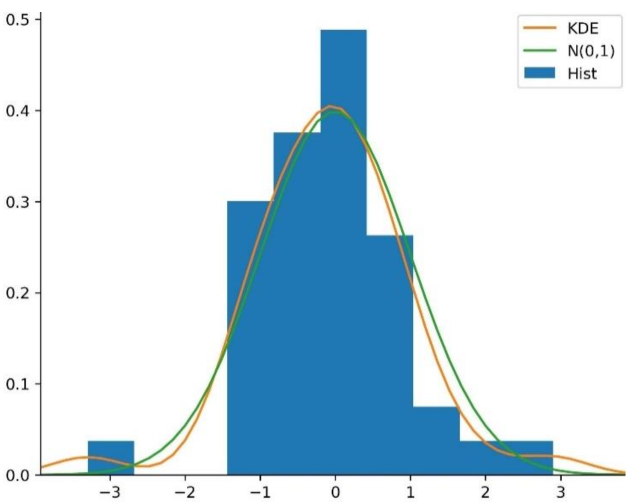

Fig. 11. Histogram plot of the SARIMA model (0,1,0) × (0,0,1)$_{13}$.

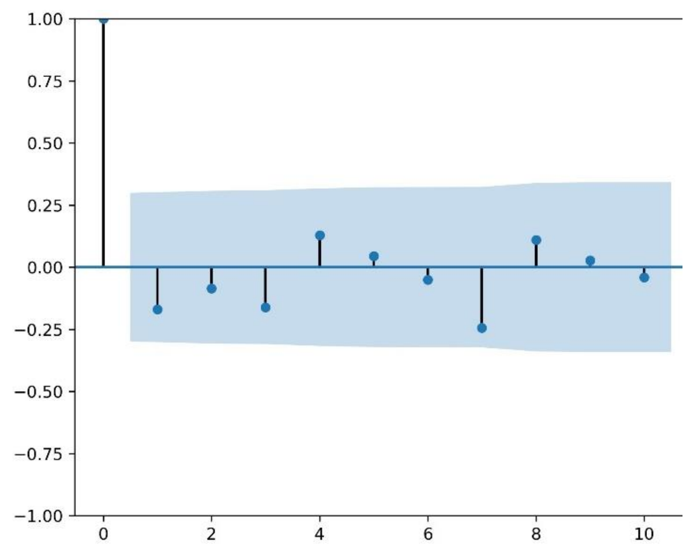

Fig. 13. Correlogram plot of the SARIMA model (0,1,0) × (0,0,1)$_{13}$.

6. Results and Discussion

The forecast values of the Holt-Winters Seasonal (0.8, 0.1, 0.0)$_{12}$ and SARIMA (0,1,0) × (0,0,1)$_{13}$ models were compared. To compare performance, the error metrics *Mean Absolute Percentage Error* (MAPE) and *Root Mean Square Error* (RMSE) were used, expressed in Equations (11) and (12), respectively.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i-\widehat{y_i}}{y_i}\right|, \quad (11)$$

$$RSME = \sqrt{\frac{\sum_{i=1}^{N}(y_i-\widehat{y_i})^2}{N}}, \quad (12)$$

where:

y_i is the actual value.
$\widehat{y_i}$ is the forecast value.
N is the number of fitted points.

The dataset was divided into 80% for training and 20% for testing the models. The forecast values estimated by each model were compared against the real values of the test set, and the RMSE and MAPE metrics were used to measure the error. Table 3 shows the comparison of the results of the RMSE and MAPE metrics. Figures 14 and 15 show the forecast values generated by the Holt-Winters Seasonal and SARIMA models, respectively.

TABLE 3
COMPARATIVE RESULTS

Model	RMSE	MAPE	AIC	BIC
SARIMA (0,1,0) × (0,0,1)$_{13}$	**7.55**	**15.09**	**14.260**	**17.782**
Holt-Winters (0.8,0.1,0.0)$_{12}$	40.03	95.88	212.880	238.217

ISSN 2395-8618

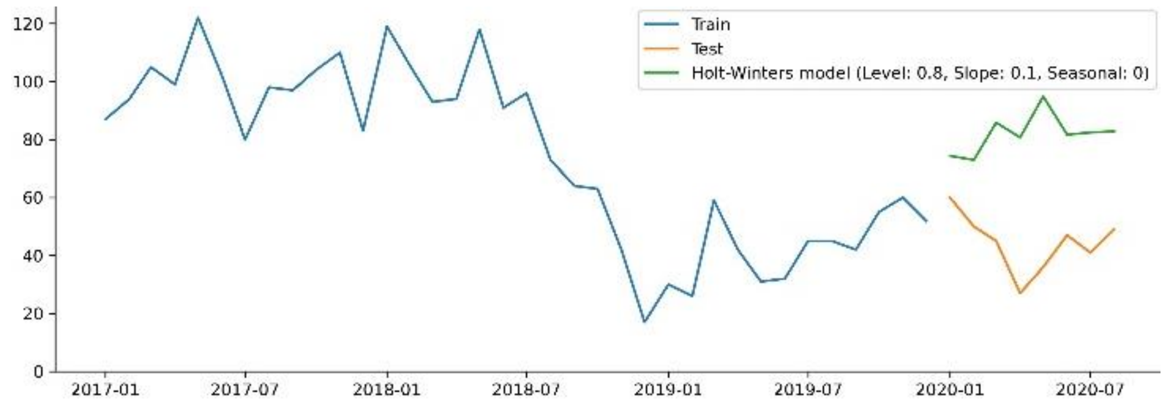

Fig. 14. Forecast values by Holt-Winters Seasonal $(0.8,0.1,0.0)_{12}$ model.

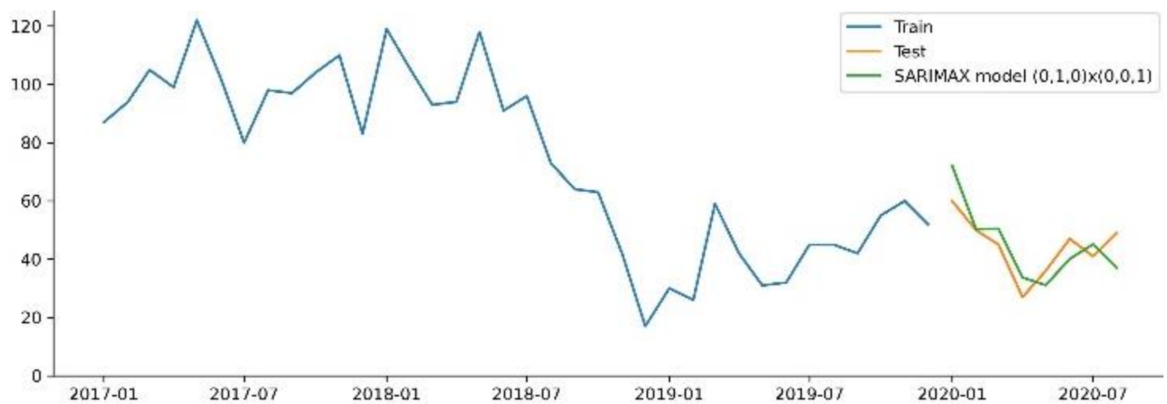

Fig. 15. Forecast values by SARIMAX $(0,1,0) \times (0,0,1)_{13}$ model.

As can be seen in Table 3, the SARIMA $(0,1,0) \times (0,0,1)_{13}$ model had the best performance since it obtained the lowest values in the RMSE and MAPE metrics. Additionally, it is worth mentioning that this model also obtained the lowest values in the AIC and BIC metrics with 14.260 and 17.782, respectively, against 212.880 and 238.217 obtained by the Holt-Winters Seasonal model.

Because the historical data provided by the warehouse was grouped monthly as mentioned above, the dataset corresponding to the time series used was generated with 45 observations. Despite being few observations, it can be seen from the results that the SARIMA models shown in Table 2 are better than those obtained by the Holt-Winter Seasonal models shown in Table 1. The possible cause of the Holt-Winter Seasonal models could not achieve a better performance could be that they could not adequately capture the seasonality of the time series due to the few observations in the dataset, which could indicate that the SSE value was set to 0.

7. Conclusions

In this work, the case of an assembly plant located in Cd. Juárez, Chihuahua, to forecast the future demand of one of the parts of the inventory was presented. Historical data between January 2017 and September 2020 was used.

Different Holt-Winters Seasonal and SARIMA models were implemented to forecast demand since the data exhibits trend and seasonal effects.

For the implementation of the SARIMA models, the Box-Jenkins methodology was used. As selection criteria for the best Holt-Winters Seasonal and SARIMA model, the AIC and BIC metrics were used. The selected models were Holt-Winters Seasonal $(0.8,0.1,0.0)_{12}$ and SARIMA $(0,1,0) \times (0,0,1)_{13}$.

RMSE and MAPE metrics were used to determine the model that makes the most accurate forecasts.

The metrics indicate that the SARIMA model has a better performance than the Holt-Winters Seasonal model since it manages to better fit the predicted data of the time-series of the used part.

The results provide evidence the SARIMA model fits the data well when forecasting the demand for inventory parts.

Based on the above, applying the SARIMA model to all the parts to know their future demand could lead to a better strategic relocation in the warehouse and, thus, facilitate the picking process.

ISSN 2395-8618

REFERENCES

[1] “Index Juárez Asociación de Maquiladoras, A.C.,” *Información Estadística Mensual*. https://indexjuarez.com/ (accessed Dec. 05, 2020).

[2] J. Fattah, L. Ezzine, Z. Aman, H. El Moussami, and A. Lachhab, "Forecasting of demand using ARIMA model," *Int. J. Eng. Bus. Manag.*, vol. 10, pp. 1–9, 2018, doi: 10.1177/1847979018808673.

[3] T. van Gils, K. Ramaekers, A. Caris, and M. Cools, "The use of time series forecasting in zone order picking systems to predict order pickers' workload," *Int. J. Prod. Res.*, vol. 55, no. 21, pp. 6380–6393, Nov. 2017, doi: 10.1080/00207543.2016.1216659.

[4] A. T. Bon and T. K. Ng, "An Optimization of Inventory Demand Forecasting in University Healthcare Centre," in *IOP Conference Series: Materials Science and Engineering*, Feb. 2017, vol. 166, no. 1, p. 012035, doi: 10.1088/1757-899X/166/1/012035.

[5] T. Mendes Dantas, F. L. Cyrino Oliveira, and H. M. Varela Repolho, "Air transportation demand forecast through Bagging Holt Winters methods," *J. Air Transp. Manag.*, vol. 59, pp. 116–123, Mar. 2017, doi: 10.1016/j.jairtraman.2016.12.006.

[6] C. Chatfield, *Time-Series Forecasting*. CRC press, 2000.

[7] R. a Yaffee and M. McGee, "An introduction to time series analysis and forecasting: with applications of SAS® and SPSS®." Elsevier, 2000.

[8] A. T. Jebb, L. Tay, W. Wang, and Q. Huang, "Time series analysis for psychological research: examining and forecasting change," *Front. Psychol.*, vol. 6, no. JUN, p. 727, Jun. 2015, doi: 10.3389/fpsyg.2015.00727.

[9] D. A. Dickey and W. A. Fuller, "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *J. Am. Stat. Assoc.*, vol. 74, no. 366a, pp. 427–431, Jun. 1979, doi: 10.1080/01621459.1979.10482531.

[10] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *J. Econom.*, vol. 54, pp. 159–178, 1992.

[11] “StatsModels.” https://www.statsmodels.org/stable/index.html (accessed Dec. 05, 2020).

[12] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, pp. 716–723, 1974, doi: 10.1109/TAC.1974.1100705.

[13] G. Schwarz, "Estimating the Dimension of a Model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, Mar. 1978, doi: 10.1214/aos/1176344136.

[14] P. R. Winters, "Forecasting Sales by Exponentially Weighted Moving Averages," *Manage. Sci.*, vol. 6, no. 3, pp. 324–342, Apr. 1960, doi: 10.1287/mnsc.6.3.324.

[15] H. A. Engelbrecht and M. Van Greunen, "Forecasting methods for cloud hosted resources, a comparison," in *Proceedings of the 11th International Conference on Network and Service Management, CNSM 2015*, Dec. 2015, pp. 29–35, doi: 10.1109/CNSM.2015.7367335.

[16] M. Markovska, A. Buchkovska, and D. Taskovski, "Comparative study of ARIMA and Holt-Winters statistical models for prediction of energy consumption," 2016.

[17] S. Makridakis and M. Hibon, "ARMA Models and the Box-Jenkins Methodology," *J. Forecast.*, vol. 16, no. 3, pp. 147–163, May 1997, doi: 10.1002/(SICI)1099-131X(199705)16:3<147::AID-FOR652>3.0.CO;2-X.

[18] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, Aug. 1978, doi: 10.1093/biomet/65.2.297.

ISSN 2395-8618

School Bus Routing Problem with Fuzzy Walking Distance

Eduardo Sánchez-Ansola, Ana Camila Pérez-Pérez, and Alejandro Rosete

***Abstract*—The School Bus Routing Problem is a type of Vehicle Routing Problem that aims to optimize the planning of bus routes for a school. This problem has received increased interest in the last decade. One of the aspects that stand out most in the progress of optimization problems is making them as close to reality as possible. In this sense, fuzzy optimization is a suitable way to do this by considering certain levels of uncertainty. Although the fuzzy approach has been applied to the Vehicle Routing Problem, it has not been so with the School Bus Routing Problem. Therefore, the objective of this paper is to introduce a fuzzy model for the School Bus Routing Problem, particularly with the maximum student walking distance as a fuzzy element. This fuzzy version of the School Bus Routing Problem allows obtaining a set of solutions with different trade-offs between cost and relaxation of the original conditions. The results obtained in 31 instances by using the parametric approach are analyzed, taking into account three characteristics of the problem: number of bus stops, number of students, and walking distance. It is shown that the introduced fuzzy version is useful for decision-makers by providing relaxed alternative solutions with significant cost savings.**

***Index Terms*—Fuzzy optimization, parametric approach, School Bus Routing Problem.**

1. Introduction

THE SCHOOL Bus Routing Problem (SBRP) aims to create, in an optimal way, a set of routes for the transportation of students to their schools from different locations [1]. This problem has been extensively studied since the publication of [2].

The SBRP is considered a type of Vehicle Routing Problem (VRP) [3]. From the point of view of optimization, in the creation of the routes, several objectives may be taken into account, e.g., minimizing the total distance traveled [4] or minimizing the number of buses to be used [5]. Also, a set of restrictions must be met, e.g., the capacity of buses [6] or the entry and exit times to schools [7]. The SBRP has been applied in different settings, e.g., in [8], workers are transported to their places of work instead of students.

On the other hand, the management of uncertainty is one of the ways to obtain models and solutions to situations more similar to reality. Two outstanding techniques for uncertainty management are stochastic approximations [9] (where some of the elements of the problem present a random behavior) and fuzzy optimization [10], where some elements involve certain levels of subjectivity or ambiguity.

In the case of fuzzy optimization, there are several studies on its application in VRP, its variants, and in other optimization routing problems. For example, in [11], several models were developed for variants of the VRP with imprecise travel times modeled as fuzzy triangular numbers. The authors of [12] present a fuzzy multi-objective optimization problem to model single frequency routes bus timetabling and solve the model by a kind of genetic algorithm. In [13], a model for the Truck and Trailer Routing Problem (TTRP) with imprecise capacity restrictions and fuzzy treatment was proposed. More recently, [14] presented an integrated production inventory routing problem (IRP) with fuzzy approximation in the demand of retailers. Likewise, in [15], the Fuzzy Green Vehicle Routing Problem (FGVRP) is considered for the design of a supply chain, where customer demands are considered fuzzy. Finally, the authors of [16] introduced a Fuzzy Electric Vehicle Routing Problem with time windows and recharging stations (FEVRPTW), where fuzzy numbers were used to treat the uncertainty of service times, battery energy consumption and travel times.

As it has been shown, in these examples of VRP models and their variants, various aspects have been considered fuzzy, such as travel time, service times, customer demand, or vehicle capacity. However, to the best of our knowledge, similar techniques have not been applied to SBRP. Particularly in the SBRP, one of the elements that can be treated as fuzzy is the student's walking distance to reach the bus stops. The distance has been treated as a fuzzy element in other optimization problems, such as the MCLP [17] from the point of view of coverage, with interesting results.

When modeling the student walking distance as a fuzzy element, the decision-maker interested in the solution of an SBRP would have the opportunity to obtain a diverse set of solutions with different trade-offs between the relaxation of the original constraints and the optimization objectives. With this set of solutions, the decision-maker could select among them the one that most satisfies a particular interest. For example, assume an SBRP instance where the objective is to minimize the total distance traveled by buses and the student's walking distance is restricted to 300 meters. In this situation, the best solution, which is the sum of all routes, may have a cost of 15 kilometers. However, if the walking distance is considered a fuzzy constraint, it may be treated in a relaxed way. For example, if the walking distance is increased to 350 meters, the best solution may decrease to 13.5 kilometers, while if the limit is increased to 400 meters, the best solution may be 13 kilometers. Hence, the decision-maker would have three options.

In the previous example, a linguistic value can be used to classify the distance between students and stops. This linguistic value can be "admissible" or "not admissible." When the student's

Manuscript received on September 12, 2020, accepted for publication on November 21, 2020, published on December 30, 2020.

The author are with the Department of Artificial Intelligence and Informatics Systems Infrastructure, Universidad Tecnológica de La Habana José Antonio Echeverría (CUJAE), La Habana, Cuba (e-mail: esancheza@ceis.cujae.edu.cu, acperez@ceis.cujae.edu.cu, rosete@ceis.cujae.edu.cu).

ISSN 2395-8618

walking distance is treated as fuzzy, this distance can be "admissible" with different degrees of satisfaction in [0,1].

Taking these antecedents into account, the objective of this work is to introduce a fuzzy model of the SBRP and thus to demonstrate its impact on solving this type of problem. In particular, in the proposal, the restriction that limits the maximum student's walking distance becomes fuzzy.

The rest of the document is organized as follows. Section 2 addresses the general characteristics of the SBRP and a study of the literature on the subject. In Section 3, the mathematical model and its fuzzy approach for the SBRP are presented. Section 4 presents and discusses the experimental results. Finally, the conclusions and future work are presented in Section 5.

2. School Bus Routing Problem

The School Bus Routing Problem (SBRP) is defined to ensure an optimal transportation policy for the students of a school or school district. It was identified as a problem in [2] more than 60 years ago. However, it is not until the last decades that contributions to the modeling and solution of this problem have increased.

According to [18] and [19], the SBRP can be divided into five sub-problems: 1) Preparation of data, 2) Selection of bus stops, 3) Generation of routes, 4) Route calendar, and 5) Adjustment to school bell time. On the other hand, [20] suggests that the preparation of the data (1) can be part of each sub-problem, and therefore treats it as such and not as a sub-problem. Likewise, it describes a new sub-problem proposed in [21] called strategic transportation policies (6).

There are several approaches for modeling and solving SBRP (see [19] and [20]). In this way, the SBRP sub-problems can be classified according to:

1. Number of schools: one or multiple schools.
2. Service environment: urban or rural.
3. Load type: distinguishes whether a bus can carry students to more than one school. Mixed loads generally refer to allowing a bus to carry students from multiple schools.
4. Fleet mix: identifies whether the buses under consideration have the same capacity (homogeneous) or varying capacities (heterogeneous).
5. Objective: generally, it is the minimization of one or more of the following aspects:
    a. Number of buses (N).
    b. Total bus traveled distance or time (TBD).
    c. The total student riding distance or time (TSB).
    d. Total student walking distance (SWD).
    e. Maximum route length (MRL).
6. Constraints: normally, in each SBRP model, it is proposed to meet one or more of the following conditions:
    a. Vehicle capacity (C).
    b. Maximum riding time (MRT).
    c. School Time Window (TW).
    d. Maximum walking time or distance (MWT).

From the solution point of view, multiple approaches have been used, including some exact, heuristic, and metaheuristic algorithms. Before the study presented in [19], only a few solutions based on metaheuristics could be found. However in the last 10 years, the interest in SBRP has substantially increased and especially the use of metaheuristic algorithms [20], such as Genetic Algorithms (e.g. [22], [23], [24]), Ant Colonies (e.g. [25], [26], [27]), Simulated Annealing (e.g. [28]), Tabu Search (e.g. [29], [30]) and GRASP (e.g. [31], [32], [33]). The use of metaheuristics in the SBRP solution is mainly due to the increase in size and complexity of the instances to be solved. Furthermore, these algorithms have been shown to provide good solutions to other combinatorial optimization problems similar to the SBRP.

3. Mathematical Model

In [34], a metaheuristic solution for SBRP with bus stops selection and homogeneous fleet was presented. The objective was to minimize the total distance traveled by the buses, with constraints related to the bus capacity and the students walking distance.

The fuzzy model presented here is a fuzzy extension of the model presented in [34]. The main difference is that the model presented here allows a relaxation of the student's walking distance. In general terms, the new model is described in terms of the variables used in [34].

3.1 Variables

Input variables

c: Capacity of each bus.
b: Number of buses.
d: Maximum students walking distance.
P: Set of possible stops.
E: Set of students.
C^p: A set of vectors with coordinate pairs of the possible stops.
C^e: A set of vectors with pairs of coordinates of the house of each student.

Auxiliary Variables

C_{ij}: Cost matrix between each pair of stops (i, j).
D: A distance that indicates the cost between a pair of stops or between a student and a bus stop. Euclidian distance is used.
$c^p{}_i$: Are the coordinates of the stop located on the i index of C^p.

$$C_{ij} = \begin{cases} D\left(c^p{}_i, c^p{}_j\right), & i \neq j \\ 0, & i = j \end{cases} \quad \forall i, \forall j \in P \tag{1}$$

S_{pe}: Binary matrix. 1 if the student j can reach the stop i, and 0 otherwise.
$c^e{}_j$: Are the coordinates of the student located on the j index of C^e.
p_0: This is the first element of P, which means the location of the school.
$\leq^f$: Indicates the imprecision of the constraint (2).

$$S_{pe} = \begin{cases} 1, D\left(c^p{}_i, c^e{}_j\right) \leq^f d \\ 0, D\left(c^p{}_i, c^e{}_j\right) >^f d \end{cases} j \in E, i \in P - \{p_0\} \tag{2}$$

Decision Variables

R_{km}: Indicates the stop that is visited by the bus k in the order m.
Z_e: Indicates the stop where the student e is picked up

ISSN 2395-8618

3.2 Objective Function and Constrains

Objective Function

$$\text{Min} \sum_{k=1}^{b} \sum_{m=1}^{|P|-1} C_{[R_{km}][R_{km+1}]} \quad (3)$$

Constrains

$$|\{R_{km} | R_{km} = p\}| \leq 1$$
$$\forall p \in P - \{p_0\}, \forall k \in \{1, \dots, b\}, \forall m \in \{1, \dots, |P|\} \quad (4)$$
$$\{(e,p) | Z_e = p\} \subseteq \{(e,p) | S_{pe} = 1\}$$
$$\forall e \in E, \forall p \in P \quad (5)$$
$$|\{e | \exists m \; R_{km} = Z_e\}| \leq c$$
$$m \in \{0, \dots, |P|\}, \forall k \in \{1, \dots, b\}, \forall e \in E \quad (6)$$
$$|\{R_{km} | Z_e = R_{km}\}| = 1$$
$$\forall e \in E \quad (7)$$

The objective function, equation (3), minimizes the total distance traveled by the entire bus fleet. Equations (4), (5), (6) and (7) represent the restrictions that must be met for the solution to be feasible. Equation (4) guarantees that each stop is visited at most once, except for stop p_0, which represents the school, the final destination of all buses. Equation (5) ensures that each student can reach their assigned bus stop. Equation (6) takes into account that the capacity of each bus is not exceeded on the route. And, finally, with Equation (7), it is guaranteed that each stop to which at least one student is assigned is visited by a bus.

Concerning the model introduced in [34], the difference is that the model in [34] wasn't fuzzy. Particularly, the fuzzy nature of the model presented here is related to equation (2), where the fuzzy operators ($>^f, \leq^f$) are used, replacing the crisp operators ($>$, $\leq$) used in [34].

3.3 Fuzzy model for SBRP

As can be seen in the previous model, restriction (5) depends on the value of the auxiliary variable S_{pe}. This is the focus of the proposed fuzzy model. The new way of posing this constraint implies that the feasibility of a student reaching a stop becomes fuzzy (i.e., not crisp) and therefore has different degrees of membership.

Taking this element into account, if the maximum walking distance for students is 200 meters, then a student that walks 190 meters satisfies it with a grade of 1. On the other hand, if a student walks 210 meters, the degree of membership (degree of satisfaction of this constraint) may be less than 1, but higher than if the stop is 250 meters away. Conversely, a stop that is located 500 meters away could be considered unreachable. All these values (e.g., 200, 500) will depend on the admissible conditions and the allowed tolerance. These values will imply that solutions will have different degrees of compliance with the restrictions. From the decision-making point of view, this relaxation allows a small increase in the distance that the students could walk to find a relaxed solution with a better value of the objective function (reduced cost).

To model this situation, it is necessary to define a tolerance H, which determines the maximum admissible distance that a student could walk. Fig. 1 shows the function to measure the degree of compliance with the restriction taking into account the distance d and the tolerance H.

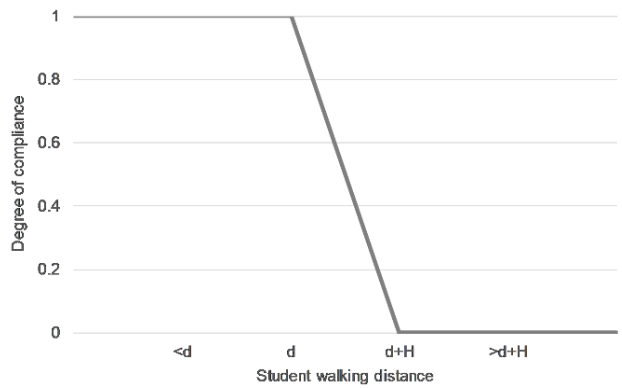

Fig. 1. Membership function of the compliance of the constraint associated with the student walking distance.

To understand this function, having d as the original maximum student walking distance and H as the maximum admissible tolerance, a student located at any distance less or equal to d has a degree of compliance of 1. On the other hand, if the student is located at any distance between d and d+H, it has a degree of compliance in the interval [0,1]. Finally, if the student location is at any distance greater than d+H to a stop, the degree of compliance is 0, and then it is assumed that the student can't reach the stop.

This is a linear function, and the parametric approximation method [10] can be applied based on the principles of parametric linear programming and the concept of alpha-cuts [10]. The concept of alpha-cut applied to this case implies that different sets of feasible solutions are associated with a particular value of alpha, i.e., those solutions with a degree of feasibility (the accomplishment of the original conditions) equal or greater than alpha. Consequently, with smaller values of alpha, some relaxed solutions are considered feasible. Then, instead of using the previous expression (2) now the expression (8) is used to obtain S_{pe}:

$$S_{pe} = \begin{cases} 1, D({c^p}_i, {c^e}_j) \leq d + H(1-\alpha) \\ 0, D({c^p}_i, {c^e}_j) > d + H(1-\alpha) \end{cases} \quad (8)$$

With this change, when $\alpha = 1$ the problem remains crisp, and then students can only reach those bus stops that are at the maximum original walking distance, i.e., it is the most restrictive case. On the other hand, when $\alpha = 0$, students are allowed to reach those bus stops that are at the maximum original distance plus the maximum tolerance, i.e., it is the greatest relaxation.

3.4 Solution approach of the fuzzy SBRP model

To illustrate the solution of the new fuzzy SBRP model, we use the parametric approach. In Fig. 2, a diagram gives a general picture of this approach, where a fuzzy problem ($P_\sim$) is transformed into a set of crisp problems ($\cup_\propto P_\propto$), where each problem corresponds to a different value of alpha ($\propto$). A solution is found for each of these crisp problems ($S_\propto$) and finally, the set of these solutions ($\cup_\propto S_\propto$), each associated with each alpha value, forms the solution of the original fuzzy problem ($S_\sim$).

ISSN 2395-8618

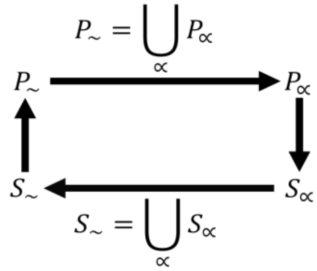

Fig. 2. Descriptive diagram of the parametric approach [35]. The cycle starts in the upper left corner.

To solve each crisp instance, any of the available solution methods for SBRP may be used. Here, we use the metaheuristic method presented in [34] to solve each crisp problem. This method consists of a metaheuristic that combines two heuristic algorithms for the construction of the initial solution with a local search strategy that has a probabilistic-based selection mechanism for mutation operators.

The first heuristic is in charge of assigning each student to a bus stop, while the second is in charge of building the routes for each bus that start from the school and go through each stop where there is at least one assigned student, to finally go back to school. This last heuristic has a greedy approach.

Four mutation operators are available for the selection mechanism for local search: Swap, Two-opt, Section swap, and Reorder.

In [34] this method was validated in the solution of the test instances previously used in [31]. These previous results demonstrate the efficacy of this metaheuristic approach that can reach the best-reported solutions in several instances and good solutions in all cases.

4. Results and Discussion

To validate the proposed fuzzy model for SBRP, as well as the algorithms used in its solution, 31 instances of problems were selected from the set studied in [34]. The characteristics of these instances can be seen in Table 1. In the case of the column Maximum Walking Distance for students, the unit for the values is “unit” in a Euclidean plain because the instances are artificial and a generic approach was applied. In a practical situation, each distance may be expressed in terms of meters or minutes.

Following what was stated in [36], these instances can be grouped according to the number of bus stops, a key element of SBRP. Particularly, there are six instances with 5, 10, and 40 bus stops, eight instances with 20 bus stops, and five instances with 80 bus stops.

Taking into account that the instances in Table 1 are the original crisp instances, each one induces an instance of the proposed fuzzy model. For the creation of these fuzzy instances, the tolerance was set to 20% of the original walking distance, i.e., H = 0.2 * d. Following the parametric approach, five values of alpha allow different relaxation for each fuzzy instance, thus allowing different degrees of membership to the crisp case, $\alpha \in \{0, 0.25, 0.5, 0.75, \text{and } 1\}$. By combining these alpha values, the solution of the 31 fuzzy instances derived in the solution of 31 × 5 = 155 crisp instances.

TABLE 1
Characteristics of the 31 Instances

Id	Instance	Stops count	Students count	Maximum walking distance for students	Capacity of the buses
1	Inst6	5	25	20	50
2	Inst9	5	50	5	25
3	Inst11	5	50	10	25
4	Inst15	5	50	40	25
5	Inst21	5	100	20	25
6	Inst24	5	100	40	50
7	Inst27	10	50	10	25
8	Inst32	10	50	40	50
9	Inst33	10	100	5	25
10	Inst37	10	100	20	25
11	Inst40	10	100	40	50
12	Inst42	10	200	5	50
13	Inst54	20	100	20	50
14	Inst55	20	100	40	25
15	Inst56	20	100	40	50
16	Inst57	20	200	5	25
17	Inst60	20	200	10	50
18	Inst64	20	200	40	50
19	Inst70	20	400	20	50
20	Inst72	20	400	40	50
21	Inst78	40	200	20	50
22	Inst79	40	200	40	25
23	Inst80	40	200	40	50
24	Inst84	40	400	10	50
25	Inst90	40	800	5	50
26	Inst95	40	800	40	25
27	Inst97	80	400	5	25
28	Inst99	80	400	10	25
29	Inst102	80	400	20	50
30	Inst108	80	800	10	50
31	Inst112	80	800	40	50

Table 2 shows the best solution value obtained for each instance from 30 executions of the metaheuristics described in [34] with 10,000 evaluations of the objective function. Each row constitutes the fuzzy solution for each instance of the fuzzy problem.

From these results, it can be observed that in 6 instances were obtained five different solutions, one for each level of relaxation or alpha-cut. In other instances, a reduced number of different solutions are obtained because some more relaxed cases do not imply an improvement in the objective function. From the point of view of a decision-maker, there is no rationale to allow a greater relaxation of a constraint if it does not imply an improvement in the cost. Thus, the solutions of interest are highlighted in bold in Table 2 because they are relevant to the decision-maker. For example, in Inst6, only two values are highlighted in bold, alpha=1 and alpha=0. That means that with the other alpha-cuts (0.75, 0.5, 0.25), the solution is not better than the solution with alpha=1; thus, these solutions should not be taken into account.

ISSN 2395-8618

TABLE 2
RESULTS OBTAINED FOR THE 31 FUZZY INSTANCES

Instance	Alpha				
	1	0.75	0.5	0.25	0
Inst6	**110.058**	110.058	110.058	110.058	**97.773**
Inst9	**286.681**	286.681	286.681	286.681	286.681
Inst11	**193.551**	193.551	193.551	**175.911**	175.911
Inst15	**13.794**	**11.072**	**9.069**	9.069	9.069
Inst21	**159.909**	159.909	159.909	159.909	159.909
Inst24	**39.807**	39.807	**33.036**	**24.167**	**12.033**
Inst27	**266.064**	266.064	266.064	266.064	**254.194**
Inst32	**56.882**	**32.797**	32.797	**25.791**	25.791
Inst33	**403.178**	403.178	403.178	403.178	403.178
Inst37	**220.359**	220.359	**198.066**	**173.156**	173.156
Inst40	**38.360**	38.360	**36.690**	**24.545**	**21.231**
Inst42	**506.060**	506.060	506.060	506.060	506.060
Inst54	**216.126**	**176.256**	**156.128**	**148.115**	148.115
Inst55	**57.540**	**43.106**	**35.732**	**31.920**	**18.805**
Inst56	**28.327**	**18.872**	18.872	**8.346**	**4.172**
Inst57	**932.522**	**928.684**	928.684	928.684	928.684
Inst60	**488.853**	**488.622**	**488.232**	**477.221**	**468.663**
Inst64	**55.197**	55.197	**32.686**	**30.206**	**28.894**
Inst70	**340.791**	**330.395**	330.395	330.395	**328.709**
Inst72	**95.540**	**79.565**	**63.200**	63.200	63.200
Inst78	**354.306**	**333.627**	**326.656**	**284.662**	**247.585**
Inst79	**95.226**	**89.373**	**63.636**	63.636	63.636
Inst80	**72.364**	**50.971**	**34.724**	**24.971**	**21.628**
Inst84	**840.655**	840.655	**800.118**	800.118	800.118
Inst90	**1376.405**	1376.405	1376.405	**1364.521**	1364.521
Inst95	**418.702**	**404.616**	**393.256**	393.256	393.256
Inst97	**1686.818**	1686.818	1686.818	1686.818	1686.818
Inst99	**1401.222**	1401.222	**1354.376**	1354.376	**1300.822**
Inst102	**566.535**	**513.113**	**474.049**	**439.875**	**427.293**
Inst108	**1459.557**	**1393.490**	1393.490	1393.490	**1368.123**
Inst112	**142.167**	**131.811**	**122.653**	**108.354**	**103.563**

Of the instances with five interesting solutions, instances Inst55 and Inst80 stand out because their greatest possible relaxation ($\alpha = 0$) implies a cost reduction of almost 70%. Fig. 3 illustrates the interesting trade-off between cost (y-axis) and relaxation (alpha value in x-axis) of the fuzzy solution of the instance Inst80 with five solutions of interest. In this case, the solution value with alpha 0 was 21.6284, which represents a 0.3 fraction of the solution value with alpha 1, 72.3635. Therefore, 70% of cost reduction can be obtained in this instance when the maximum relaxation is used.

In five of the instances, four solutions of interest were obtained; in 9 instances, three possible solutions; and six instances offer two different solutions. On the other hand, only in five instances (Inst9, Inst21, Inst33, Inst42, Inst97), the relaxation does not contribute any improvement to the original solution, i.e., the most relaxed solution ($\alpha = 0$) does not imply any improvement. The fact that a better solution than the original one can't be found in these instances is mainly because the change in the constraints of the instance does not allow a significant change in the alternative stops that are available for each student.

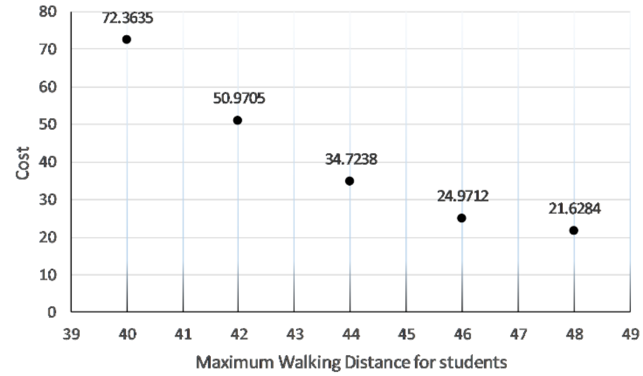

Fig. 3. Instance Inst80 solution in all membership values.

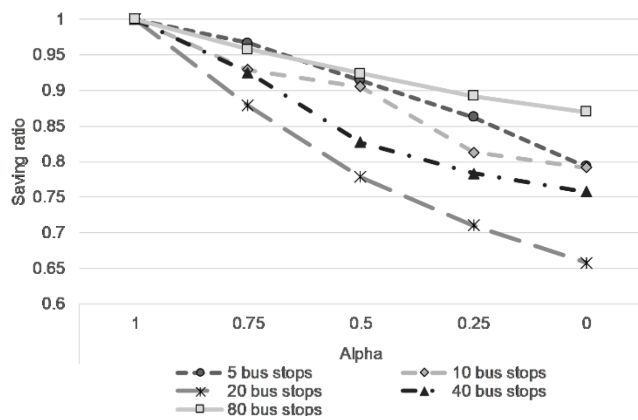

Fig. 4. Saving ratios in the instances grouped by the number of bus stops.

These results allow us to affirm that, by increasing the students walking distance by at most 20%, considerable savings can be achieved in the total distance traveled by the routes and therefore in the fuel used. This implies that the proposed fuzzy model may be meaningful for a decision-maker. For example, in the fuzzy solution of the instance Inst80 presented in Fig. 3 it can be appreciated the possibility that at the cost of increasing the walking distance from 40 ($\alpha = 1$, crisp) to 48 ($\alpha = 0$, the most relaxed), a cost-saving of more than 70% is achieved. Other intermediate values of relaxation, values of α, and cost may also be interesting.

Another analysis that can be made is to compare the archived results with solution values presented in [31]. In this comparison, the relaxation at any level brings a better solution in 18 instances. The average cost-saving in all instances is about 14%. On the other hand, if only the 18 instances with the best solutions are compared, the average cost-savings is around 25%.

Fig. 4 shows a graph in which the average savings obtained in the instances can be observed according to the number of stops, based on the different values of α (0: the greatest relaxation; 1: no relaxation, i.e., the original problem).

In Fig. 4, it can be seen that, on average, in all types of instances, some savings between 4%, when the relaxation is the smallest one (alpha=0.75), and 35%, when the relaxation is the highest one (alpha=0), are achieved. In this way, the instances with 20 stops stand out because they allow saving almost 35% when the maximum relaxation is achieved. On the contrary, the

ISSN 2395-8618

instances with 80 stops only allow savings of 10% with the maximum relaxation. According to this analysis, the relaxation of the walking distance seems to be less important when there are many stops due to the existence of multiple options in the same interval.

From another point of view, Fig. 5 shows a graph of the average proportional savings obtained for the different values of α in the instances, grouped by the maximum amount that a student can walk.

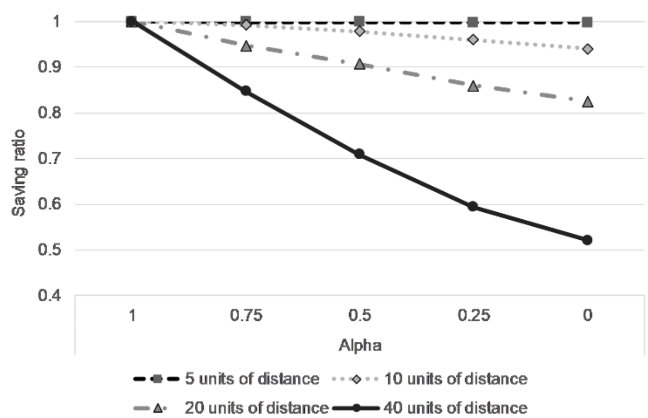

Fig. 5. Saving ratios in the instances grouped by the maximum students walking distance.

In this case, it can be seen how the greatest savings are achieved in the instances with the greatest walking distance (40), which is between 15% and 48% (greatest relaxation). On the other hand, it can be observed that in the instances with the shortest walking distance (5), practical savings are not achieved.

Likewise, the results can be analyzed, taking into account the number of students. In Fig. 6, the proportion of savings achieved for the different values of α is shown. In this case, the instances are grouped according to the number of students. Savings are achieved in all types of instances, the instances with 100 students being the ones with the most remarkable savings with values between 10% and 40% for the maximum relaxation.

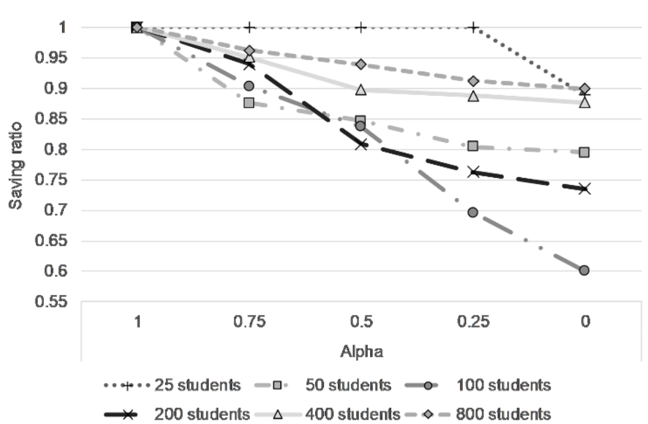

Fig. 6. Saving ratios in the instances are grouped by the number of students.

In general, the results obtained in these instances indicate that intermediate values of the number of bus stops and students and large values of walking distances in the instances tend to allow a greater reduction in the cost.

The proposed approach has the main advantage with respect to the solutions found in the literature. Our proposal allows the decision-makers to have more than one solution of interest to be considered; thus, they can evaluate different trade-offs between cost-savings and compliance with the original conditions.

In addition, from the point of view of computational cost, the parametric approach can be applied in a very efficient way. The solution found without any relaxation can be used to search for the solution with the lower allowed relaxation, and then this new solution can be used as the starting solution for the next level of relaxation and so on.

5. Conclusions

This paper introduced a fuzzy model for the SBRP that allows modeling uncertainty for the constraint associated with the student walking distance. The proposed model and its solution allow obtaining a fuzzy solution as a set of crisp solutions to the SBRP with interesting trade-offs between cost and accomplishment of the original maximum student's walking distance. A main contribution of the proposal is the possibility of offering to the decision-maker the opportunity to analyze multiple solutions with different degrees of membership to the original conditions. Likewise, the results allow us to affirm that, at the cost of slightly increasing the student's walking distance; considerable savings can be achieved in the evaluation of the objective function.

This research allows us to trace a way forward in the application of fuzzy optimization in the SBRP concerning other aspects, e.g., the capacity of buses. Another interesting aspect is to study how to set the alpha values to obtain the most interesting trade-offs between cost and relaxation.

References

[1] A. Bock, E. Grant, J. Könemann, and L. Sanità, "The School Bus Problem on Trees," Berlin, Heidelberg, 2011, pp. 10-19.

[2] R. Newton and W. Thomas, "Design of school bus routes by computer," *Socio-Econ. Plan. Sci.* , vol. 3, pp. 75-85, 1969.

[3] L. Li and Z. Fu, "The school bus routing problem: a case study," *Journal of the Operational Research Society,* vol. 53, pp. 552–558 2002.

[4] H. Caceres, R. Batta, and Q. He, "School bus routing with stochastic demand and duration constraints," *Transportation Science,* 2017.

[5] J. Kinable, F. Spieksma, and G. Vanden, "School bus routing—a column generation approach," *International Transactions in Operational Research,* vol. 21, pp. 453–478, 2014.

[6] D. M. Miranda, R. S. d. Camargo, S. V. Conceio, M. F. Porto, and N. T. R. Nunes, "A multi-loading school bus routing problem," *Expert Systems with Applications,* vol. 101, pp. 228–242, 2018.

[7] F. M. de Souza Lima, D. S. D. Pereira, S. V. da Conceição, and R. S. de Camargo, "A multi-objective capacitated rural school bus routing problem with heterogeneous fleet and mixed loads," *4OR - A Quarterly Journal of Operations Research,* vol. 15, pp. 359-386, 2017/12/01 2017.

[8] S. B. Kiriş and T. Özcan, "Metaheuristics Approaches to Solve the Employee Bus Routing Problem With Clustering-Based Bus Stop Selection," in *Artificial Intelligence and Machine Learning Applications in Civil, Mechanical, and Industrial Engineering*, ed, 2020, pp. 217-239.

ISSN 2395-8618

[9] C. Erbao and L. Mingyong, "The open vehicle routing problem with fuzzy demands," *Expert Systems with Applications*, pp. 2405–2411, 2010.

[10] A. Ebrahimnejad and J. L. Verdegay, *Fuzzy Sets-Based Methods and Techniques for Modern Analytics*: Springer, Cham, 2018.

[11] J. Brito, F. J. Martínez, J. A. Moreno, and J. L. Verdegay, "Fuzzy approach for Vehicle Routing Problems with fuzzy travel time," in *International Conference on Fuzzy Systems*, 2010, pp. 1-8.

[12] S. L. Tilahun and H. C. Ong, "Bus Timetabling as a Fuzzy Multi-objective Optimization Problem Using Preference-based Genetic Algorithm," *Promet – Traffic&Transportation*, vol. 24, pp. 183-191, 2012.

[13] I. Torres, C. Cruz, and J. L. Verdegay, "Solving the Truck and Trailer Routing Problem with Fuzzy Constraints," *International Journal of Computational Intelligence Systems*, vol. 8, pp. 713-724, 2015/07/04 2015.

[14] A. Ghasemkhani, R. Tavakkoli-Moghaddam, S. Shahnejat-Bushehri, S. Momen, and H. Tavakkoli-Moghaddam, "An integrated production inventory routing problem for multi perishable products with fuzzy demands and time windows," *IFAC-PapersOnLine*, vol. 52, pp. 523-528, 2019/01/01/ 2019.

[15] A. Giallanza and G. L. Puma, "Fuzzy green vehicle routing problem for designing a three echelons supply chain," *Journal of Cleaner Production*, vol. 259, p. 120774, 2020/06/20/ 2020.

[16] S. Zhang, M. Chen, W. Zhang, and X. Zhuang, "Fuzzy optimization model for electric vehicle routing problem with time windows and recharging stations," *Expert Systems with Applications*, vol. 145, p. 113123, 2020/05/01/ 2020.

[17] V. C. Guzmán, D. A. Pelta, and J. L. Verdegay, "An approach for solving maximal covering location problems with fuzzy constraints," *International Journal of Computational Intelligence Systems*, vol. 9, pp. 734-744, 2016/07/03 2016.

[18] J. Desrosiers, J. Ferland, J. Rousseau, G. Lapalme, and L. Chapleau, "An overview of a school busing system," in *Scientific Management of Transport Systems*, N. Jaiswal, Ed., ed Amsterdam, North-Holland, 1981, pp. 235-243.

[19] J. Park and B. Kim, "The school bus routing problem. A review.," *European Journal of Operational Research*, vol. 202, pp. 311–319, 2010.

[20] W. Ellegood, S. Solomon, J. North, and J. Campbell, "School bus routing problem: contemporary trends and research directions," *Omega*, 2019.

[21] W. A. Ellegood, J. F. Campbell, and J. North, "Continuous approximation models for mixed load school bus routing," *Transportation Research Part B: Methodological*, vol. 77, pp. 182-198, 2015/07/01/ 2015.

[22] M. Kang, S. Kim, J. Felan, H. Rim, and M. Cho, "Development of a genetic algorithm for the school bus routing problem," *International Journal of Software Engineering and Its Applications*, vol. 9, pp. 107-126, 2015.

[23] L. Sales, C. Sousa, T. Oliveira, and B. Athayde, "Memetic algorithm for the heterogeneous fleet school bus routing problem," *Journal of Urban Planning and Development*, vol. 144, 2018.

[24] O. Unsal and T. Yigit, "Using the genetic algorithm for the optimization of dynamic school bus routing problem," *Broad Research in Artificial Intelligence and Neuroscience*, vol. 9, Mayo 2018.

[25] H. Lingmei, Y. Guifeng, F. Bowen, W. Hongzhou, and G. Weitao, "School bus routing problem based on ant colony optimization algorithm," in *2014 IEEE Conference and Expo Transportation Electrification Asia-Pacific (ITEC Asia-Pacific)*, 2014, pp. 1-5.

[26] B. Yao, Q. Cao, Z. Wang, P. Hu, M. Zhang, and B. Yu, "A two-stage heuristic algorithm for the school bus routing problem with mixed load plan," *Transportation Letters*, vol. 8, pp. 205-219, 2016/08/07 2016.

[27] N.-a. Mokhtari and V. Ghezavati, "Integration of efficient multi-objective ant-colony and a heuristic method to solve a novel multi-objective mixed load school bus routing model," *Applied Soft Computing*, vol. 68, pp. 92-109, 2018/07/01/ 2018.

[28] F. M. de Souza Lima, D. S. Pereira, S. Vieira da Conceição, and N. T. Ramos Nunes, "A mixed load capacitated rural school bus routing problem with heterogeneous fleet: Algorithms for the Brazilian context.," *Expert Systems with Applications*, vol. 56, pp. 320-334, 2016.

[29] J. Pacheco, R. Caballero, M. Laguna, and J. Molina, "Bi-Objective Bus Routing: An Application to School Buses in Rural Areas," *Transportation Science*, vol. 47, pp. 397-411, 2013.

[30] P. Aparicio Ruiz, J. Muñuzuri Sanz, and J. Guadix Martín, "Mixed Trips in the School Bus Routing Problem with Public Subsidy," in *Enhancing Synergies in a Collaborative Environment*, P. Cortés, E. Maeso-González, and A. Escudero-Santana, Eds., ed Cham: Springer International Publishing, 2015, pp. 105-113.

[31] P. Schittekat, J. Kinable, K. Sörensen, M. Sevaux, F. Spieksma, and J. Springael, "A metaheuristic for the school bus routing problem with bus stop selection," *European Journal of Operational Research*, 2013.

[32] M. F. Faraj, J. F. Sarubbi, C. M. Silva, M. F. Porto, and N. T. R. Nunes, "A real geographical application for the School Bus Routing Problem," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2014, pp. 2762-2767.

[33] V. S. de Siqueira, E. N. da Silva, R. V. S. da Silva, and M. L. Rocha, "Implementation of the metaheuristic grasp applied to the school bus routing problem," *International Journal of e-Education, e-Business, e-Management and e-Learning*, vol. 6, pp. 137-145, 2016.

[34] A. C. Pérez-Pérez, E. Sánchez-Ansola, and A. Rosete, "Una solución metaheurística al problema de planificación de rutas de autobuses escolares con flota homogénea y selección de paradas," *Ingeniería*, vol. 26, to be published 2021.

[35] J. Herrera-Franklin, A. Rosete, M. García-Borroto, C. Cruz-Corona, and D. A. Pelta, "On the Impact of Fuzzy Constraints in the Variable Size and Cost Bin Packing Problem," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Cham, 2020, pp. 230-240.

[36] R. Pérez-Rodríguez and A. Hernández-Aguirre, "Probability model to solve the school bus routing problem with stops selection," *International Journal of Combinatorial Optimization Problems and Informatics*, vol. 7, pp. 30-39, 2016.

ISSN 2395-8618

Optimization of Many Objectives with Intervals Applying the MOEA/D Algorithm

Lorena R. Rosas-Solórzano, Claudia G. Gómez-Santillán, Nelson Rangel-Valdez, Laura Cruz-Reyes, Fausto A. Balderas-Jaramillo and José A. Brambila-Hernández

***Abstract*—Project Portfolio Selection (PPS) is a major strategic decision problem faced by any organization. PPS decides how to invest resources into projects subject to a decision process influenced by multiple conflicting criteria. The portfolio's compromise to the organization's well-being has an uncertainty that directly affects a decision maker's preferences (DM). MOEA/D is a well-known approach to tackle multicriteria optimization problems, and it is still open for the development of strategies to handle uncertainty on its search process. This work proposes I-MOEA/D, a new method based on a MOEA/D approach, to deal with DM's uncertainty in costs and benefits of portfolios' projects. The proposed novel features include (a) handling large numbers of objectives; (b) a method to generate the initial population; and (c) handling the uncertainty of resources, costs, and benefits through intervals. An experiment compared I-MOEA/D against the state-of-the-art I-NSGA-II algorithm in instances with two to fifteen objectives. Results demonstrate the competitiveness of I-MOEA/D by improving the quality of solution of I-NSGA-II in most instances.**

***Index Terms*—Decision making, uncertainty, multi-objective optimization, mathematics of intervals, project portfolio problem.**

1. Introduction

ORGANIZATIONS usually address Project Portfolio Selection (PPS) aided by a decision-maker (DM). The DM and the decision analyst often must provide information on portfolio values; however, such information might be incomplete, causing a condition of uncertainty. The PPS has distinct solutions in state-of-the-art works; however, there is still a lack of research that handles uncertainty in the PPS, even fewer using intervals [1][2][3]. An interval is a range used to represent unclear projects' values defined for organizational resources, e.g., benefits, costs, requirements, times, synergies, partial support, among others.

Manuscript received on September 14, 2020, accepted for publication on November 24, 2020, published on December 30, 2020.

L. R. Rosas- Solórzano, C. G. Gómez-Santillán, L. Cruz-Reyes, F. A. Balderas-Jaramillo, and J. A. Brambila-Hernández are with the Postgraduate & Research Division, National Mexican Institute of Technology/Madero Institute of Technology, 89440, Madero, Tamaulipas, Mexico (e-mail: {g13071208, claudia.gomez, lauracruzreyes, austo.balderas}@itcm.edu.mx, g02070820@cdmadero.tecnm.mx).

N. Rangel-Valdez is a Researcher of CONACYT, attached to Postgraduate & Research Division, National Mexican Institute of Technology/Madero Institute of Technology, 89440, Madero, Tamaulipas, Mexico (e-mail: nelson.rangel@itcm.edu.mx).

The DM is responsible for selecting the portfolio that best meets the organizational objectives. However, to carry out this activity, he or she faces difficulties in solving the PPS due to: the exponential complexity of the optimization problem, the number of involved objectives, the lack of information about the exact contribution of the projects to the portfolio, and the imprecise knowledge of the requirements or resources needed to complete the projects and their availability. The improper modeling of the previous difficulties can lead to portfolios that can affect the institutions' interests.

Carazo [4] defines a project as a temporary, unique, and unrepeatable process that pursues a specific set of objectives, which, when combined, will impact the vi-sion and mission of the organizations. A portfolio is a set of projects that, carried out in a given period, share a series of resources, among which there may be rela-tionships of complementarity, incompatibility, and synergies produced by sharing costs and benefits derived from the implementation of more than one project at a time [5].

The proper selection of projects for a portfolio can ben-efit any organization based on the DM's objectives. Therefore, a portfolio decision analysis can help the DMs select a subset of an extensive set of projects through modeling, considering relevant constraints, preferences, and inaccuracies in the information [6].

Currently, some state-of-the-art strategies to tackle with PPS and uncertainty are: fuzzy sets [7][8][9][10][11][12][13], interval analysis [14][15][16][17][18] and probability distributions [19]. Some authors have also introduced modifications in the portfolio to minimize uncertainty and model the DM attitude to risk [20] [21][22].

This paper proposes the analysis of an algorithm based on intervals as the solution to PPS under uncer-tainty. The approach uses an MOEA family algorithm (Multi-Objective Evolutionary Algorithms) that solves problems with many objectives and of interest to the scientific community. Specifically, we worked with the MOEA/D (Multi-Objective Evolutionary Algorithm based on Decomposition) algorithm. The proposed MOEA/D, denoted I-MOEA/D for Interval Multi-Objective Evolutionary Algorithm based on Decomposition, includes novel features as intervals to manage uncertainty and adequate handling of many-objective optimization problems. A developed experi-ment shows the performance of I-MOEA/D against I-NSGA-II, an algorithm from the scientific

ISSN 2395-8618

literature ba-sed on NSGA-II (Non-Dominated Sorting Genetic Algo-rithm II) that also handles uncertainty [10].

This document is structured as follows: Section 2 pro-vides some background on Multi-Objective PPS; also, it presents the proposed I-MOEA/D, initialization fun-ction, and random instance generator. Section 3 descri-bes the experiment conducted to validate I-MOEA/D and the results; besides, it provides the analysis that demonstrates the proposed strategy's advantages. Fina-lly, Section 4 summarizes the main conclusions drawn from the research.

2. Proposed Solution

2.1 Multi-objective Project Portfolio Selection

Until today, Multi-objective Project Portfolio Selection (or just PPS) has distinct approaches that solve it [4] [5]. A solution is a portfolio composed of one or more projects. A project is a series of activities related to each other to reach a specific objective, which consumes resources. A formal definition of PPS with uncertainty (UPPS) is the following.

Let the binary vector $\vec{x} = \langle x_1, x_2, \ldots, x_p \rangle$ of size p be a portfolio, where p is the available projects, $x_i = 1$ or $x_i = 0$ represents whether or not a project i is in the portfolio, respectively. Let $c(\vec{x})$ and $f(\vec{x}) = \{f_1(\vec{x}), f_2(\vec{x}), \ldots, f_m(\vec{x})\}$ be the portfolio cost and fitness. Let B be the budget available to form the portfolio. Finally, let $A = \{A_1, A_2, \ldots, A_a\}$ and $R = \{R_1, R_2, \ldots, R_r\}$ bounds over specific areas and regions of interest that must be satisfied by the portfolio. If $c(.) = [\underline{c}, \overline{c}]$, $f(.) = [\underline{f}, \overline{f}]$, $B = [\underline{B}, \overline{B}]$, $A_i = [\underline{A}, \overline{A}]$, and $R_i = [\underline{R}, \overline{R}]$ are intervals defined by a lower $\underline{l}$ and upper $\overline{u}$ bounds then equations 1 and 2 formalize the definition of UPPS.

$$\max f(\vec{x}) = \{f_1(\vec{x}), f_2(\vec{x}), \ldots, f_n(\vec{x})\} \quad (1)$$

$$\begin{aligned} \text{Subject to:} \quad & c(\vec{x}) \leq B \\ & \underline{A_i} \leq A_i(\vec{x}) \leq \overline{A_i} \\ & \underline{R_i} \leq R_i(\vec{x}) \leq \overline{R_i} \\ & \vec{x} \in \{0,1\}^n \end{aligned} \quad (2)$$

where the basic arithmetic and relational operations follow previously defined computations (cf., [15, 20]), e.g. $f_i(\vec{x})$, $c(\vec{x})$, $A(\vec{x})$, and $R(\vec{x})$ are product of a linear combination of the contribution of each project i in the portfolio in fitness, cost, area, or region, respectively.

2.2 Generate Initial Population with Exchange

This section describes the initialization strategy for the population of I-MOEA/D, in the presence of intervals. The process is simple; it chooses a project i as part of the portfolio whenever a random uniform value v lies under a predefined selection threshold selection, β (set to 0.5 for this research work). Following a trial-and-error ap-proach, the algorithm discards those solutions that be-came infeasible in the process. Algorithm 1 shows the pseudocode of the method.

Algorithm 1. Generation of Initial Population

Input:

- β-Threshold selection
- *m*-Objectives number
- *p*-Total projects
- *a*-Number of areas
- *r*-Number of regions

Output:

- Initial population

```
0.  x⃗ = {1, 1, …,1}
1.  while (! Feasibility (x⃗)) do
2.    x⃗ = {0, 0, …,0}
3.    for each i ∈ {1,2, …, p} do
4.      r=random (0,1)
5.      If (r < β) then
6.        x_i = 1
7.        c (x⃗) += c(i)
8.        for each j ∈ {1, …, m} do
9.          f_i(x⃗) += f_j(i)
10.         end
11.         for each j ∈ {1, …, a} do
12.           A_i(x⃗) += A_j(i)
13.         end
14.         for each j ∈ {1, …, r} do
15.           R_i(x⃗) += R_j(i)
16.         end
17.       end
18.     end
19. end
20. return x⃗
```

Line 1 uses function Feasibility(.) to ensure a feasible solution; it validates the restrictions of the UPPS of budget, area, and region.

The algorithm tests each project for inclusion into the portfolio in Lines 4 and 5. Whenever the condition is satisfied the costs, and values for objectives, areas, and regions are accumulated (Lines 7, 8, 11, and 14, respectively).

Feasibility requires the addition and relational ≤ operations. Given two interval numbers $\boldsymbol{E} = [\underline{E}, \overline{E}]$ and $\boldsymbol{D} = [\underline{D}, \overline{D}]$, the result of $\mathbf{C} = \mathbf{E} + \mathbf{D}$ can be computed as $\mathbf{C} = [\underline{E} + \underline{D}, \overline{E} + \overline{D}]$. In the other hand, the relational operation $\boldsymbol{D} \leq \boldsymbol{E}$ can be estimated using the relational quotient defined by equation 3.

$$p_{ED} = \frac{\overline{E} - \underline{D}}{(\overline{E} - \underline{E}) + (\overline{D} - \underline{D})} \quad (3)$$

Based on the relational quotient, equation 4 defines the possibility measure of $Poss(\boldsymbol{D} \leq \boldsymbol{E})$ used to express the desired relationship between the intervals **D** and **E**. This work establishes that $Poss(\boldsymbol{D} \leq \boldsymbol{E}) \geq 0.5$.

$$Poss(\boldsymbol{D} \leq \boldsymbol{E}) = \begin{cases} 1 & \text{if } p_{ED} > 1, \\ p_{ED} & \text{if } 0 \leq p_{ED} \leq 1, \\ 0 & \text{if } p_{ED} \leq 0 \end{cases} \quad (4)$$

Figure 1 depicts the process performed by Algorithm 1 in the construction of a portfolio. This figure shows an array

ISSN 2395-8618

with 25 cells representing a portfolio and the 25 potential projects. Each cell also represents an iteration and its randomly generated value. Note that the shadow cells correspond to those where the random value lay under the threshold β = 0.5. The process is repeated as many solutions the initial population of I-MOEA/D has.

1	2	**3**	4	5	**6**	7	8
0.89	0.59	**0.16**	0.76	0.68	**0.42**	0.73	0.96

9	10	**11**	12	**13**	14	15	16
0.09	0.53	**0.38**	0.71	**0.12**	0.83	0.88	0.65

17	18	19	**20**	21	22	23	**24**
0.99	0.79	0.55	**0.36**	0.68	0.62	0.92	**0.28**

25
0.64

Selected random value

Fig. 1. Graphic representation of the iterative process of Algorithm 1 to build a portfolio.

While Figure 1 shows the process of selecting projects, Figure 2 shows the binary representation required by I-MOEA/D. The portfolio must be feasible, and all the constraints of costs, areas, and regions must be satisfied.

1	2	3	4	5	6	7	8	9
0	0	1	0	0	1	0	0	1

10	11	12	13	14	15	16	17	18
0	1	0	1	0	0	0	0	0

19	20	21	22	23	24	25
0	1	0	0	0	1	0

Fig. 2. A binary vector representing a solution (or portfolio) used by I-MOEA/D. Here the value 1 means that the project is part of the portfolio and 0 otherwise.

2.3 Multi-objective Evolutionary Algorithm Based on Decomposition with Intervals (I-MOEA/D)

MOEA/D is a technique proposed by Zhang and Hui [23]. This algorithm consists of decomposing a multi-objective optimization problem into several sub-problems that are optimized simultaneously. I-MOEA/D is a variant of MOEA/D that solves PPS with uncertainty; it implements evolutive operators to handle intervals. The intervals represent a mean of expression of uncertainty in objectives values, costs, and resources. Algorithm 2 shows the general pseudocode of the proposed strategy. I-MOEA/D gives an external population (*EP*) containing the non-dominated solutions found during the optimization process.

Algorithm 2. I-MOEA/D

Input:

- MOP= Multi-objective Optimization Problem
- N= Population size
- p=Number of projects
- m= Number of objectives
- T= Neighborhood size of the weight vectors
- MaximumEvaluations= Number of Generations

Output:

EP= External population

0. W = ReadWeightsVector()
1. EP= ϕ
2. Calculate_Euclidean_Distance (V)
3. SortVector ()
4. **Population=GenerationOfInitialPopulation()**
5. Initialize_Z (Population)
6. Generations=0;
7. While (Generations < MaximumEvaluations)
8. *i*=1
9. For each $i \in N$ do
10. [p1, p2]=**SelectionByTournament(Population,B(i), T)**
11. *offspring*= CrossoverOnePoint (p, [p1, p2])
12. *offspring*= GeneMutation (p, *offspring*)
13. *offspring*=**ImprovementGeneMutation**(*offspring*)
14. **UpdateZ** (**$\vec{z}$,** *f(offspring)*)
15. **UpdateEP** (**EP**, *f(offspring)*)
16. end
17. Generations++
18. end

The binary vectors $\vec{x} = \langle x_1, x_2, \ldots, x_p \rangle$ encode one portfolio or solution provided by the algorithm. Such vectors are chromosomes in the evolutive approach, and the vectors' indexes of the array are alleles denoting distinct projects.

The methods of I-MOEA/D that distinguish from implementations of other MOEA/D are five, appearing in bold in Algorithm 2 (Lines 2, 10, 13-15). These methods are the initialization function, the selection operator, the repair/improve operator, and the Z vector and EP set update. The remaining section provides a detailed description of the methods.

Initialization Phase. In the first stage, the I-MOEA/D algorithm makes the weight vector's initial set (Line 0) and initializes EP to empty (Line 1). It also initializes the weight vectors $\{w_1, w_2, \ldots, w_N\}$, computes the Euclidean distance among them and initializes their vectors' neighborhood $B(i)$, $1 \le i \le N$ (Lines 2 to 4). The Algorithm 1 fills the initial solution vector B of size N and associates each solution B_i to a weight vector w_i. Then, the neighborhood $B(i)$ of a weight vector w_i contains the closest weight vectors index by Euclidean distance. Finally, the initialization phase fills the vector Z, the vector of best objective values found in the search process (Line 5). The vector B corresponds to the initial population. The main loop of the I-MOEA/D begins having as stop criterion a maximum number of evaluations previously defined (Line 7).

ISSN 2395-8618

Selection by Tournament. The selection method selects from the population two solutions at random. Then it compares them by cost and assigns the best as first parent p_1, and the other as second parent p_2. Algorithm 3 returns both parents (cf. [24]). This method requires comparing the cost using the interval relational comparison, as shown in section 2.1.

Algorithm 3. Selection by Tournament

Input

- B = Population
- B_i = Neighborhood of each weight vector *i*
- T= neighborhood size of a weight vector

Output

Parents *p1* y *p2*

0. **While (**$k==l$**) do**
1. k = *Random* ()
2. *l* = *Random* ()
3. **end**
4. $x = B_{i,k}$
5. $y = B_{i,l}$
6. **If** (c (B[*x*]) < c (B [y]) **then**
7. p_1=B[*x*]
8. p_2=B[*y*]
9. **else**
10. p_1=B[*y*]
11. p_2=B[*x*]
12. **end**

Crossover one point. The two chosen parents from the tournament selection method combine their chromosomes to produce one new offspring. For this purpose, the method selects a random index in the parents' vector as a cutting point to inherit the genes to the new child from each parent. This is a technique by Holland [25] and implemented in I-MOEA/D to solve UPPS (Algorithm 4). This strategy does not require handling intervals.

Algorithm 4. One Point Crossover

Input

- *p* = number of projects (allels on each parent)
- [*p1*, *p2*] = parents Parent *x*1

Output

y = child

0. cut = *Random* (1, *p*-1)
1. *y* [0 … *cut* – 1] = *p*1[0 … *cut* – *1*]
2. *y* [*cut* … *p* – 1] = *p*2[*cut* … *p* – *1*]
3. return *y*

Algorithm 4 has two phases. First, a *cut* is chosen at random, and it must be between 1 and $p - 1$, where *p* is the number of projects (Line 0). After that, the child is created using parts from the parents p_1, p_2. The first parent will transmit the genes corresponding to alleles in indexes 0 to $cut - 1$ of its corresponding vector (Line 1); this is the best parent of both by cost. The second parent will donate the genes from its vector' indexes from *cut* to $p - 1$ (Line 2). The new child is the offspring that the method returns. Figure 3 shows a graphic depiction of how the parents' genes are inherited to the child using our method.

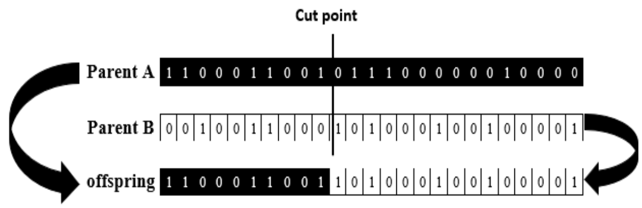

Fig. 3. Parents' gene inheritance to the offspring

Gene mutation. The mutation operator chosen is a simple mutation. This process selects one allele from the solution and changes its value. Given that the solution is a binary vector, the chosen allele will change its value from 1 to 0, or vice versa [26]. Algorithm 5 shows this strategy that also does not deal with intervals.

Algorithm 5. Simple Mutation

Input

- *p* = number of projects
- *y* = Solution to be mutated

Output

y'= Mutated child

0. *r*= Random (0, *p* – 1)
1. *y*' = *y*
2. *y*'[*r*] = *y*'[*r*] + 1) % 2;
3. **return** *y*'

After applying the genetic operators by I-MOEA/D (Lines 10-12, Algorithm 2), the generated solution goes into a repair/improvement process (Line 13, Algorithm 2). The method repairs/improves a solution by making unfeasible solutions feasible. The strategy used randomly takes out projects until the satisfaction of the restrictions. This procedure requires the implementation of interval operations, both arithmetic and relational.

Each iteration of I-MOEA/D updates the vectors $\vec{z}$, B (or Population), and EP with the offspring. The offspring substitutes the ideal objectives' values in $\vec{z}$ if necessary. The EP set must eliminate solutions dominated by offspring and include it if is nob-dominated. The dominance condition uses the interval relation operations previously defined.

Finally, when the stop criterion is met, I-MOEA/D reports the set EP as the approximated region of interest.

2.4 PPS instance generator with intervals

Algorithm 6 shows the pseudocode of the proposed instance generator for PPS with intervals. The user configurable parameters to create an instance are: budget, number of objectives, projects, areas and regions, and limits of costs, and objectives. The outputs are the interval values that define the

ISSN 2395-8618

budget, areas, regions, and for projects their costs, objectives values, and the area and region where they belong.

The generator creates an interval budget in Line 0 based on the input budget *B*. Figure 4 shows an example of the definition of such intervals.

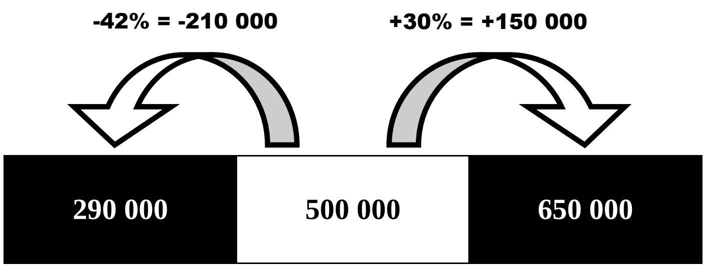

Fig. 4. Budget parameter

From Lines 1 to 6 the generator creates the values for the areas of the instance. Here, it uses the budget *B* to define appropriate limites to areas' values (Lines 1 and 2). After that, it randomly chose values withing those limites as the bounding values of each area (Lines 4 and 5).
From Lines 7 to 12 the generator assigns values to the regions in a similar fashion as done in the areas; i.e., it uses the budget to define appropriate maximum limits, and with them randomly chose values to bound the distinct regions.

The next step in the generator is the definition of values for the projects. Lines 14 to 26 perform this task. First the area and regions are randomly chosen in lines 14 and 15. After that, the cost of the project is created within the limits provided as inputs (Lines 16 to 17). From Lines 18 to 26 the process generates the values for the objectives of a project. It uses two strategies, one based on the costs of the project (Line 20), and the other based on the limits for the objectives established as input (Line 22). With the value *o* the generator creates an interval for the objective using 80% of it as lower bound and 120% as upper bound.

3. Experimentation

This section contains a series of experiments aimed to validate the quality of the I-MOEA/D compared with the I-NSGA-II algorithm [18].

The configuration of the experimental design took into account the following details: a) the size of the set of tested UPPS instances was 7; b) the project set involved was always of cardinality 100; c) the number of objectives involved in the cases varied according to {2,3,4,8,9,13,15}; d) the proposed random generator shown in section 2.4 generated the instances. Concerning the algorithms, the population size was 100, the stop criterion was after 500 generations, and the crossover and mutation operators considered a probability of 100%.

The algorithm test environment was implemented in the Java programming language and ran on a computer with the following features: 2.20 GHz Intel Core i5 CPU, 4 GB RAM, and Windows 10 Operating System.

Algorithm 6. PPS with intervals instance generator

Input:

- $B \leftarrow$ Budget (No interval)
- $\{m, p, a, r\} \leftarrow$ Numbr of objectives, projects, areas, and regions
- $[\underline{c}, \overline{c}] \leftarrow$ Project Costs extreme limits (No Intervals)
- $[\underline{m}, \overline{m}] \leftarrow$ Objectives extreme limits

Output:

- $[\underline{B}, \overline{B}] \leftarrow$ Budget as interval
- $\{[\underline{a_1}, \overline{a_1}], [\underline{a_2}, \overline{a_2}], \ldots, [\underline{a_a}, \overline{a_a}]\} \leftarrow$ Limits of each areas *i*
- $\{[\underline{r_1}, \overline{r_1}], [\underline{r_2}, \overline{r_2}], \ldots, [\underline{r_r}, \overline{r_r}]\} \leftarrow$ Limits of each region *r*
- $\{\{C_1, A_1, R_1\}, \ldots, \{C_p, A_p, R_p\}\} \leftarrow$ Cost, Area and region for each project *p*
- $\{[\underline{f_{1p}}, \overline{f_{1p}}], [\underline{f_{2p}}, \overline{f_{2p}}], \ldots, [\underline{f_{mp}}, \overline{f_{mp}}]\}$ Benefit from the m objectives of each project p (in intervals)

0. $[\underline{B}, \overline{B}]$ = [0.58*B*, 1.3*B*]
1. $[\underline{a_l}, \overline{a_l}]$ = $[(0.7 * B)/(1.7a+0.1a^2), (1.27 * B)/(1.7a+0.1a^2)]$
2. $[\underline{a_u}, \overline{a_u}]$ = $[((2.159 + 0.127a) * B) / a, ((2.635+0.155a) * B)/a]$
3. **for each** $i \in \{1, 2, \ldots, a\}$ **do**
4. $\underline{a_i} = \underline{a_l}$ + Random $(\overline{a_l} - \underline{a_l})$
5. $\overline{a_i} = \underline{a_u}$ + Random $(\overline{a_u} - \underline{a_u})$
6. **end**
7. $[\underline{r_l}, \overline{r_l}]$ = $[(0.8 * B)/(1.7r+0.1r^2), (1.2 * B)/(1.7r+0.1r^2)]$
8. $[\underline{r_u}, \overline{r_u}]$ = $[((1.02+0.06r) * B)/r, ((2.38 + 0.14r) * B)/r]$
9. **for each** $i \in \{1, 2, \ldots, r\}$ **do**
10. $\underline{r_i} = \underline{r_l}$ + Random $(\overline{r_l} - \underline{r_l})$
11. $\overline{r_i} = \underline{r_u}$ + Random $(\overline{r_u} - \underline{r_u})$
12. **end**
13. **for each** $i \in \{1, 2, \ldots, p\}$ **do**
14. A_i= Random(*a*)
15. R_i= Random(*r*)
16. $v = \underline{c}$ + Random $(\overline{c} - \underline{c})$
17. $[\underline{C_i}, \overline{C_i}] = [0.99 * v, 1.2 * v]$
18.   **for each** $j \in \{1,2, \ldots, m\}$ **do**
19.   **if (**Random**.**nextBoolean()**) then**
20.   obj = Random $((v - \underline{c}) / (\overline{c} - \underline{c}))$
21.   **else**
22. $o = \underline{m}$ + Random $(\overline{m} - \underline{m})$
23.   **end**
24. $\underline{f_{ij}}$= 0.8*o
25         $\overline{f_{ij}}$ = 1.1*o
26.   **end**
27. **end**

The *number of non-dominated portfolios*, and the *portfolios' cardinality* are the two quality measurements of interest for this work to assess the algorithms' performance for comparison purposes. Equations (5) to (8) show the indicators formed from the previous measurements, where *EP* is the final non-dominated set of algorithm's solutions after 30 independent

ISSN 2395-8618

runs. Let's point out that larger indicator' values represent better performance in an algorithm.

$$I_1 = |EP| \tag{5}$$

$$I_2 = \frac{\sum_{x \in EP} |x|}{|EP|} \tag{6}$$

$$I_3 = \min_{x \in EP}\{|x|\} \tag{7}$$

$$I_4 = \max_{x \in EP}\{|x|\} \tag{8}$$

Table 1 compares I-MOEA/D and I-NSGA-II. Columns 1 and 2 show the instances' names and algorithms, respectively. Columns 3 to 6 shows the indicators' values. Note that the encoded name *oipj* contains the numbers of objectives *i* and projects *j*.

According to Table 1, I-NSGA-II improves I-MOEA/D in the instance with two objectives. The differences range from 5% to 12% in the indicators' observed values; this is a common condition since NSGA-II generally has a good performance in that number of objectives.

Consequently, I- MOEA/D is the clear winner in the remaining instances; its performance differences vary from 69% to 97%. In conclusion, the overall results shown in Table 2 demonstrates that I-MOEA/D improves I-NSGA-II in all the indicators. These results also indicate poor performance of I-NSGA-II in many-objective problems, a condition previously observed.

TABLE 1
COMPARISON OF I-MOEA/D AND I-NSGA-II BY QUALITY INDICATORS

Instance	Algorithm	I_1	I_2	I_3	I_4
o2p100	I-NSGA-II	63	57	56	58
	I-MOEA/D	55	54	53	55
o3p100	I-NSGA-II	526	39	37	40
	I-MOEA/D	4556	44	40	46
o4p100	I-NSGA-II	139	58	57	59
	I-MOEA/D	449	68	67	68
o8p100	I-NSGA-II	585	38	35	40
	I-MOEA/D	21327	46	41	47
o9p100	I-NSGA-II	579	41	38	42
	I-MOEA/D	27863	45	40	47
o13p100	I-NSGA-II	521	52	49	54
	I-MOEA/D	5123	63	61	64
o15p100	I-NSGA-II	677	34	31	37
	I-MOEA/D	21417	47	41	48

To provide further insights, we compute the relative differences among the indicators measured for I-MOEA/D and I-NSGA-II. For this purpose, equation 9 defines a metric to calculate the percentage of improvement achieved by the winner algorithm for the given indicator I^k_j, where j is the indicator and k=1 if the algorithm is I-MOEA/D or k=2 if it is I-NSGA-II. A winner algorithm has the highest indicator value. Tables 2 and 3 summarizes the results obtained from this metric for instances with objectives 3 to 15.

$$\mathrm{Diff}(I_j^1, I_j^2) = \begin{cases} 100\left(\dfrac{I_j^1 - I_j^2}{I_j^1}\right), & \text{if } I_j^1 > I_j^2 \\ 100\left(\dfrac{I_j^2 - I_j^1}{I_j^2}\right), & \text{otherwise} \end{cases} \tag{9}$$

The results from Tables 2 and 3 shows that I-MOEA/D improves all the indicators measures with respect to I-NSGA-II in percentual ranges that vary in [69, 97], [8, 27], [7, 24], and [10,13] for the indicators I_1, I_2, I_3, and I_4, respectively. These results tell that I-MOEA/D obtains more non-dominated solutions and portfolios with a greater number of projects, which is desirable.

TABLE 2
PERCENTAGE DIFFERENCE OF NON-NOMINATED PORTFOLIOS AND AVERAGE CARDINALITY

Instance	Diff ($I^1{}_1$, $I^2{}_1$)	Diff ($I^1{}_2$, $I^2{}_2$)
o3p100	88%	11%
o4p100	69%	14%
o8p100	97%	17%
o9p100	97%	8%
o13p100	89%	17%
o15p100	96%	27%

TABLE 3
PERCENTAGE DIFFERENCE IN MINIMUM AND MAXIMUM CARDINALITY OF PORTFOLIOS

Instance	Diff ($I^1{}_3$, $I^2{}_3$)	Diff ($I^1{}_4$, $I^2{}_4$)
o3p100	7%	13%
o4p100	14%	13%
o8p100	14%	14%
o9p100	5%	10%
o13p100	19%	15%
o15p100	24%	22%

Finally, Table 4 compares the dominance proportion per algorithm. For this purpose, a set *EP** combines the final sets EP^1 and EP^2; this new set is the final non-dominated front. Then, we calculate the number of solutions of EP^1 and EP^2 that appear in *EP**. Let's note that EP^1 and EP^2 correspond to the final non-dominated fronts EP of I-MOEA/D and I-NSGA-II, respectively. Column 3 contains the number of non-dominated solutions still appearing in EP*. Column 4 reports the number of solutions that became dominated after integration.

ISSN 2395-8618

TABLE 4
DOMINANCE PROPORTION AMONG I-MOEA/D E I-NSGA-II

Instance	Algorithm	Total, non-dominated portfolios	Dominated portfolios
o2p100	I-NSGA-II	0	63
	I-MOEA/D	55	0
o3p100	I-NSGA-II	441	85
	I-MOEA/D	4556	0
o4p100	I-NSGA-II	0	139
	I-MOEA/D	449	0
o8p100	I-NSGA-II	1	584
	I-MOEA/D	21327	0
o9p100	I-NSGA-II	546	33
	I-MOEA/D	27863	0
o13p100	I-NSGA-II	0	521
	I-MOEA/D	5123	0
o15p100	I-NSGA-II	1	676
	I-MOEA/D	21417	0

Considering the information of Table 4, all the solutions of I-MOEA/D remain non-dominated. Moreover, it turns out that they dominate several solutions provided by I-NSGA-II and, in some cases, all of them (instances 2 and 13). These results corroborate the excellent performance of I-MOEA/D to solve UPPS over I-NSGA-II with many objectives.

The experimental design concluded with an analysis of the statistical differences of the observed results. Particularly, a Wilcoxon's test [27] validated the difference in the indicator I_1. The considered sample was the number of non-dominated portfolios of each of the 30 runs of an instance. The test utilized a significance level of 5%. The null hypothesis was "$\mathbf{H_0}$=*The medians of the differences between the two group samples are equal*". Table 5 summarizes the results.

TABLE 5
WILCOXON TEST

Instance	***p value***	**Result**
o2p100	0.57746866	Is accepted H0
o3p100	0.04311445	Is rejected H0
o4p100	0.04311445	Is rejected H0
o8p100	0.04311445	Is rejected H0
o9p100	0.04311445	Is rejected H0
o13p100	0.04311445	Is rejected H0
o15p100	0.04311445	Is rejected H0

The results from Table 5 show that there are significant differences in 6 instances, and based on information from Table 1, the differences favor I-MOEA/D. Let us point out I-NSGA-II performed better than I-MOEA/D only in the instance *"o2p100"*. However, there is no significant statistical difference in their performance. In conclusion, the overall performance of I-MOEA/D improves largely that of I-NSGA-II in the selected instances of UPPS.

4. CONCLUSIONS

This article proposes a new evolutionary strategy called I-MOEA/D. The main features of this algorithm are the use of intervals to express uncertainty and handling many objectives. A comparison in performance between I-MOEA/D and I-NSGA-II (a state-of-the-art approach) assessed the relevance of our approach. In equal experimental conditions under a controlled environment, the results show that I-MOEA/D outperforms I-NSGA-II [18], demonstrating the significance of I-MOEA/D.

The I-MOEA/D requires at least to modify the genetic operators, the repair/improve operator, the update methods of ideal objectives values and population, in order to integrate the use of intervals properly. The strategy required the definition of some interval operators to perform arithmetic, relational and dominance operations. The dominance operator appears with the definition of relational operators for comparison.

The observed results show that I-MOEA/D and I-NSGA-II solve UPPS. However, with increasing objectives, the performance of I-MOEA/D improves that of I-NSGA-II, as expected, particularly in the analyzed instances with number of objectives varying from two to fifteen. The results show that with an increasing number of objectives, I-MOEA/D return solutions with better quality.

Finally, the number and diversity of solutions offered by I-MOEA/D are large. This is a good condition in contrast to I-NSGA-II because it means that I-MOEA/D approximates the Pareto front better. However, it is interesting to ask if the search process of I-MOEA/D can include DM's preferences. If the latter is possible, then, a narrower set of solutions could be delivered to the DM, based on his/her priorities. Hence, the proper incorporation of preferences in the search process of I-MOEA/D represents an attractive research area for future developments.

ACKNOWLEDGMENTS

Authors thanks to CONACYT for supporting the projects from (a) Cátedras CONACYT Program with Number 3058. (b) CONACYT Project with Number A1-S-11012 from Convocatoria de Investigación Científica Básica 2017–2018 and CONACYT Project with Number 312397 from Programa de Apoyo para Actividades Científicas, Tecnológicas y de Innovación (PAACTI), a efecto de participar en la Convocatoria 2020-1 Apoyo para Proyectos de Investigación Científica, Desarrollo Tecnológico e Innovación en Salud ante la Contingencia por COVID-19. (c) Lorena R. Rosas-Solórzano would like to thank CONACYT for the support number 960719

REFERENCES

[1] G. Rivera, C. Gómez, L. Cruz, R. García, F. Balderas, E. Fernández, & F. López. "Solution to the social portfolio problem by evolutionary algorithms." *International Journal of Combinatorial Optimization Problems and Informatics* 3, 2 (2012): 21-30.

[2] C. Gomez, L. Cruz-Reyes, G. Rivera, N. Rangel-Valdez, L. Morales-Rodriguez, & M. Perez-Villafuerte. "Interdependent Projects selection

ISSN 2395-8618

with preference incorporation." *New Perspectives on Applied Industrial Tools and Techniques*. Springer, Cham, 2018. 253-271. https://doi.org/10.1007/978-3-319-56871-3_13

[3] S. Bastiani, L. Cruz-Reyes, E. Fernandez., C. Gómez, & G. Rivera. "An ant colony algorithm for solving the selection portfolio problem, using a quality-assessment model for portfolios of projects expressed by a priority ranking." *Design of Intelligent Systems Based on Fuzzy Logic, Neural Networks and Nature-Inspired Optimization*. Springer, Cham, 2015. 357-373, https://doi.org/10.1007/978-3-319-17747-2_28

[4] A. Carazo, T. Gómez, J. Molina, A. Hernandez, F. Guerrero, R. Caballero, "Solving a comprehensive model for multiobjective project portfolio selection", *Computers & Operations Research*, pp. 630-639, 2010, https://doi.org/10.1016/j.cor.2009.06.012

[5] G. Fox, N. Baker, "Project Selection Decision Making Linked to a Dynamic Environment", *Management Science*, 1272-1285, 1985, https://doi.org/10.1287/mnsc.31.10.1272

[6] A. Salo, J. Keisler, A. Morton, "Portfolio Decision Analysis. Improved methods for resource allocation", *International Series in Operations Research & Management Science*, In A. Salo et al. (Eds), vol 162, Springer, New York, 3-27, 2011.

[7] F. Perez and T. Gomez, "Multiobjective project portfolio selection with fuzzy constraints", *Annals of Operations Research* 245(1–2) 7–29, 2016 https://doi.org/10.1007/s10479-014-1556-z.

[8] C. Carlsson, R. Fuller, M. Heikkila and P. Majlender, "A fuzzy approach to R&D portfolio selection", *International Journal of Approximate Reasoning* 44(2) 93–105, 2007 https://doi.org/10.1016/j.ijar.2006.07.003.

[9] H. Ke and B. Liu, "Project scheduling problem with mixed uncertainty of randomness and fuzziness", *European Journal of Operational Research* 183(1) 135–147, 2007, http://doi.org/10.1016/j.ejor.2006.09.055.

[10] T. Hasuike, H. Katagiri and H. Ishii, "Portfolio selection problems with random fuzzy variable returns", *Fuzzy Sets and Systems* 160(18), 2579–2596, 2009, https://doi.org/10.1016/j.fss.2008.11.010.

[11] K. K. Damghani, S. Sadi-Nezhad and M. B. Aryanezhad, "A modular decision support system for optimum investment selection in presence of uncertainty: Combination of fuzzy mathematical programming and fuzzy rule based system", *Expert Systems with Applications* 38(1) 824–834, 2011, http://doi.org/10.1016/j.eswa.2010.07.040.

[12] X. Huang, "Optimal project selection with random fuzzy parameters", *International Journal of Production Economics* 106(2) 513–529, 2007, https://doi.org/10.1016/j.ijpe.2006.06.011.

[13] D. Kuchta, "Use of fuzzy numbers in project risk (criticality) assessment", *International Journal of Project Management* 19(5) 305–310, 2001, http://doi.org/10.1016/S0263-7863(00)00022-3.

[14] T. Fliedner and J. Liesio, "Adjustable robustness for multi-attribute project portfolio selection", *European Journal of Operational Research* 252 931–946, 2016, http://doi.org/10.1016/j.ejor.2016.01.058.

[15] J. Liesio, P. Mild and A. Salo, "Preference programming for robust portfolio modeling and project selection", *European Journal of Operational Research*, 181(3), 1488–1505, 2007, http://doi.org/10.1016/j.ejor.2005.12.041.

[16] J. Liesio, P. Mild and A. Salo, "Robust portfolio modeling with incomplete cost information and project interdependencies", *European Journal of Operational Research* 190(3), 679–695, 2008, https://doi.org/10.1016/j.ejor.2007.06.049.

[17] F. Balderas, E. Fernandez, C. Gomez and L. Cruz-Reyes, "Metaheuristic robust optimization of project portfolios using an interval-based model of imprecisions", *International Journal of Combinatorial Optimization Problems and Informatics* 7(3) 101–118, 2016, Available at https://www.ijcopi.org/index.php/ojs/article/view/32.

[18] F. Balderas, E. Fernandez, C. Gómez, L. Cruz, Rangel, N., Morales, M.: "A grey mathematics approach for evolutionary multi-objective metaheuristic of project portfolio selection", *Fuzzy logic augmentation of neural and optimization algorithms: theoretical aspects and real applications*, Springer Cham, 379-388, 2018.

[19] A. Toppila and A. Salo, "Binary decision diagrams for generating and storing non-dominated project portfolios with interval-valued project scores", *European Journal of Operational Research* 260(1) 244–254, 2017, http://doi.org/10.1016/j.ejor.2016.12.019.

[20] M. Rabbani, M. A. Bajestani and G. B. Khoshkhou, "A multi-objective particle swarm optimization for project selection problem", *Expert Systems with Applications* 37(1) (2010) 315–321, https://doi.org/10.1016/j.eswa.2009.05.056.

[21] A. L. Medaglia, S. B. Graves and J. L. Ringuest, "A multiobjective evolutionary approach for linearly constrained project selection under uncertainty", *European Journal of Operational Research* 179, 869–894, 2007, https://doi.org/10.1016/j.ejor.2005.03.068

[22] F. Balderas, E. Fernandez, C. Gómez., G. Rivera, & L. Cruz-Reyes. "Uncertainty modelling for project portfolio problem using interval analysis." *International Journal of Combinatorial Optimization Problems and Informatics* 7, 3 (2016): 20-27.

[23] Q. Zhang, L. Hui, "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition", *IEEE Transactions on Evolutionary Computation*,712-731, 2007, https://doi.org/10.1109/TEVC.2007.892759

[24] B. Miller and D. Goldberg, "Genetic Algorithms, Tournament Selection, and the Effects of Noise". *Complex Systems* 9: 193-212, 1995.

[25] J. Holland. Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. U Michigan Press, 1975.

[26] R. Díaz. Optimización de estructuras reticuladas planas de madera mediante algoritmos genéticos. Valdivia, Chile, 2014.

[27] F. Wilcoxon. "Comparaciones individuales por métodos de clasificación". *Boletín biométrico*,80-83, 1945.

ISSN 2395-8618

Towards a Computational Intelligence Framework to Smartify the Last-Mile Delivery

Jhonny Pincay, Edy Portmann, and Luis Terán

Abstract—**Last-mile is the component of the supply chains that has the most potential to be optimized and give advantage to retailers and delivery companies. At the same time, it is the hardest to deal with. Factors such as traffic, weather, unexpected events, or the simple fact that a customer is not at home affect directly the efficiency of the overall shipping process. This work-in-progress article proposes a framework for the improvement of the first-try delivery by studying traffic on the streets and past delivery success as a way of approximating customers' presence at home. In contrast to existing solutions, it is proposed to work only with data that does not compromise the customers' privacy and to get insights about traffic features in cities without the need of deploying expensive equipment to obtain data. The main goal is to provide a route plan to the delivery team and route planners, which allows finishing the distribution of the parcels in the least amount of time, while being able to effectively deliver the highest amount of them. This will be translated into less resource consumption and increased customer satisfaction. The research work is conducted following the principles of design science research for information systems. The implementation will use methods of computational intelligence to address the lack of precise information, following a transdisciplinary approach as industrial partners support the development of this study.**

Index Terms—**Smart logistics, last-mile delivery, swarm intelligence, fuzzy logic.**

I. Background and Aim of the Research Work

ELECTRONIC commerce has grown tremendously in the last decades and shows no signs of declining any time soon. Day to day millions of transactions are performed and the same amounts of articles are being sent using postal services.

Currently, standing out from the competition and responding to costumers' expectations of delivery services are the main concerns for online retailers and postal service companies [1]. Besides quality products and good deals, most customers expect the shipment process of their parcels to be fast and efficient while being inexpensive [2], [3]. These aspects mean added complexity to supply chains and increasing costs of inventory management, packing and picking up, transport, customer service, among others.

Manuscript received on June 24, 2020, accepted for publication on July 31, 2020, published on December 30, 2020.

Jhonny Pincay and Luis Terán are with the Human-IST Institute, University of Fribourg, Bouleverad de Pérolles 90, Switzerland (e-mail: {jhonny.pincaynieves, edy.portmann}@unifr.ch).

Luis Terán us with the Human-IST Institute, University of Fribourg, Bouleverad de Pérolles 90, Switzerland, and the Universidad de las Fuerzas Armadas ESPE, Av. General Rumiñahui, s/n, Ecuador (e-mail: luis.teran@unifr.ch, lfteran1@espe.edu.ec).

While most delivery companies have optimized the majority of their distribution channels, there is still a lack of effective models for the optimization of the so-called "last-mile delivery." According to several authors (e.g., [4], [5], [6]), the last mile—the transportation of parcels from a hub to the final destination—is the component of the supply chains, which has the more potential to be optimized and give advantage to retailers and delivery companies. Nevertheless, the last mile is also the most difficult level to deal with, given the dynamic nature of the conditions where it occurs. Aspects such as traffic jams, weather conditions, unexpected events, or even that the customer is not at home when a postal worker rings the bell, affect directly the efficiency and resource consumption of the overall shipping process and the customers' satisfaction.

In the literature, solutions aiming to improve the last-mile delivery can be grouped into four groups [7]:

1) *Change in location:* Leaving a parcel in an alternative location.
2) *Change in time:* Delivering parcels at the times indicated by the customer.
3) *Change in route:* Closely related to changes in time, meaning that the route is adapted according to the time availability of the customers.
4) *Change in behavior:* Triggering customers to choose certain characteristics or delivery time, by offering cheaper fares for example.

While most of the existing solutions claim to have improved the efficiency of the last-mile delivery, most of them are context dependant are very closely related to the characteristics of an area [7], [5]. Moreover, they face challenges when it comes to obtaining the necessary data to develop prediction models since it comes incomplete, and becomes complex as a consequence of business operations (e.g., road and speed restrictions, multiple delivery stops, and waiting on customers). Another issue is the low sample rate as logistic companies might have unique vehicles covering certain routes. Further issues are related to the fact that is hard to accurately predict where a customer is going to be at a certain time [6], [8], given that privacy is a major concern and most of the time, customers are not willing to share their location with companies and some laws need to be accomplished (e.g., General Data Protection Regulation—GDPR) which increase complexity. Moreover, despite the alternatives provided to customers (e.g., pick-up places, ad-hoc carriers, trunk delivery, and parcel lockers), home delivery is the first choice of the

ISSN 2395-8618

vast majority of people. Therefore, efforts should be directed toward addressing the aforementioned issues.

This research work proposes a framework for the improvement of first-try delivery by studying the behavior of traffic on the streets and customers' presence at home. In contrast to existing solutions, it is proposed to work only with data that does not compromise the customers' privacy (i.e., avoiding the use of tracking data) and getting insights about traffic characteristics without the need of deploying a large number of vehicles or expensive sensors to obtain data. The main goal is to provide delivery route plan to the delivery team and route planners, which allows finishing the distribution of the parcels in the least amount of time, while being able to effectively deliver the highest amount of them, which will be translated into fewer resources consumption and increased customer satisfaction.

The aforementioned could be achieved through the use of fuzzy logic and computational intelligence, as a way of dealing with uncertainty and not accurate data [9], and following a transdisciplinary approach [10] to developing a solution that truly adjusts to the needs of the delivery companies and customers. Furthermore, this work will be conducted following the principles of the design science research for information systems methodology [11] and tested in a real-life scenario.

Considering the aspects presented in the previous section, the questions that need to be answered towards reaching the goals of this study are:

– *Which methods of computational intelligence are suitable towards finding ways of implementing a re-routing algorithm?* Given the nature of the problem, computational intelligence methods could be used to implement an improved routing of delivery. However, all these methods and theories should be studied in detail to find a viable one. Through pre-defined criteria, studying the needs of stakeholders, and a review of the state of the art, it will be possible to find an answer to these unknowns.
– *How can we reduce the first-try delivery failure without compromising customer privacy and while reducing resource consumption?* Considering aspects as the customer's presence at home and traffic characteristics, it will be possible to conceptualize and to build a framework, which will provide the means to process information in a human-like way to create a way that will allow to handling imprecision derived from the nature of the problem and the data studied. The framework will be evaluated utilizing a prototype implementation, experiments, and interviews with customers and experts.

II. Methodology

Design science research methodology guidelines will be used to conduct the present work. This research approach was chosen because it allows implementing artifacts systematically to extend existing knowledge while providing solutions to practical issues or organizational problems [11].

The design science research stages are to be executed in the following manner:

1) **Identify problems and motivation:** Through state-of-the-art review and interviews with people working on the logistics sector, it will be possible to define the current problems that afflict the last-mile delivery.
2) **Define Objectives of a solution**: Given the results of stage 1, the targets and requirements that the developed solutions must achieve can be clearly defined.
3) **Design and development:** After stage 2, it is possible to develop artifacts (i.e., a software prototype) that will allow us to conduct experimental tests to define if suitable solutions to the problems are implemented.
4) **Demonstration:** After the implementation of the artifacts, case studies will be executed towards demonstrating that they fulfill the requirements and meets their purpose.
5) **Evaluation:** By means of comparison with other existing solutions, the performance, and quality of the results provided by the artifacts can be evaluated. Moreover, expert interviews and satisfaction surveys can be conducted to evaluate the results from a qualitative perspective. Once this stage is completed, the researcher will decide if it is necessary to move back to stage 3. to improve the results or to move to the last process step which is:
6) **Communication** of the results, which translates into publishing the findings.

III. Theoretical Framework

This section provides details about the components that will conform a framework aimed at improving the first-try ratio delivery. Details about possible evaluation methods are also presented.

A. Building Blocks and Interplay

A theoretical framework aimed at improving the first-try delivery ratio is depicted in Figure 1. It is composed of four (4) main layers: *data*, *knowledge*, *intelligence*, and *visualization* layer.

The *Data layer* contains the different data sources used to the end of this work. Possible data sources that are the history of successful deliveries, the vehicles' log of events during service hours, and current route planning. The databases needed to conduct this research work are already available, given an existing partnership with an industrial partner.

The *Knowledge Layer* takes data from the data layer and processes it trough fuzzy logic methods; the output of this layer is the categorization of customers according to past deliveries success and critical traffic sectors in a city.

The output from the knowledge layer serves as input for the *intelligence layer* and it constitutes the basis to compute and find the most convenient route to complete deliveries in the least amount of time while saving resources consumption

ISSN 2395-8618

(i.e., less time spent on the streets translates into less fuel). A Swarm intelligence algorithm will be used to find the best candidate routes, considering the customer characteristics and if the routes have to pass by critical traffic areas. Options of swarm intelligence that are suitable for the present research work include Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO) [12].

Finally, the *visualization layer* allows displaying of the delivery route which requires the least amount of time to be covered. It will allow logistic planners and the delivery team to interact with the recommended route and to make adjustments and changes; customers will be able to get updates about the status of their delivery. Moreover, it will be possible for the users to provide feedback about how useful the recommended route was and this feedback can be used to improve future recommendations.

Further details about a possible implementation of the knowledge and intelligence layer are presented below.

B. Knowledge Layer

Before attempting to find better ways to increase the first-try ratio of deliveries, it is important to understand the customer and traffic behavior, as those are factors directly affecting whether delivery is successful or not. Moreover, it is necessary to identify routes that have a less successful delivery ratio to be considered when re-routing the vehicles and to have a baseline for comparison and evaluation.

Clustering methods are a powerful way to analyze data and have become essential when developing solutions, interpreting information, and data dimensionality reduction. Unlike hard clustering, fuzzy clustering assigns elements to clusters with gradual memberships, meaning that objects can belong to a certain degree to more than one cluster. This provides more detail when building models and can reveal how ambiguous or certain an element belongs to a cluster [13].

In this work, it is pretended to deal with data that describe non-precise information such as the success of past deliveries (e.g., *"always successful* and *"mostly successful"*) and traffic congestion (e.g., *"congested street"* and *"normal traffic"*). Thus, fuzzy clustering constitutes a better fit towards representing in a better way the information and processing more homogeneous groups rather than a large number of single elements.

In the following, the algorithms and methods that can be used towards building the proposed artifact are explained.

1) Past Deliveries Success: The inclusion of customers' presence at home as a way of improving the first-try delivery ratio is a trend that has been exploited lately [14], [6], [8]. Considering that this work proposes an improvement to the first-try delivery without compromising the customers' privacy, tracking data about their location will not be used as unlike solutions found in the literature. Therefore, the authors propose using data from previous deliveries, which is owned by delivery companies, as a way of approximate the customers' home presence.

Through the Fuzzy C-Means algorithm [15], we propose to cluster addresses of customers according to the incidence of successful deliveries. It is possible to obtain clusters that categorize customers' addresses where the deliveries are for example *"successful most of the times"*, *"usually successful"*, *"rarely successful"*, and *"always unsuccessful"* in a similar fashion as proposed in the work of Mangiaracina et al. [6], but as we are dealing with non-precise concepts, the usage of fuzzy clustering will allow to have more expressive partitions and identify customers that might be on the border of a cluster, and therefore, having a more realistic interpretation of the reality.

2) Critical Areas Identification: With the goal of spending less time on the roads, it is necessary to identify zones inside the cities where traffic anomalies occur since they might provoke delays covering the delivering routes.

Most delivery trucks are fitted with global positioning system (GPS) devices which record their position and events happening when the vehicles are in operation. These tracking data can be used to derive information about what is happening in the streets without the need of deploying a large number of vehicles or expensive equipment [16], [17]. The authors of this article have already performed two case studies to identify anomalies on the streets [18] and predict travel time [19] through the geospatial indexing technique Geohash, and aggregation and mathematical operations.

C. Intelligence Layer

The intelligence layer is the component responsible for determining and recommending the best-performing routes to cover the package delivery. Even if there are various shortest path and optimal route algorithms (e.g., Dijkstra, A* search algorithm, and Bellman-Ford algorithm) [20], [21], they are rather static and do not respond to the need of routing problems in dynamic environments where factors such as traffic conditions and weather play crucial roles when finding optimal routes.

Ant colony algorithms have been used as an alternative to address the aforementioned challenges [21], [22]. They are based on the behavior of ants; these simple insects are able to accomplish complex tasks such as finding food, by working as a unit and lying some sort of pheromone which helps them to find the shortest path from their nests to the food sources [21]. Moreover, ant colony-based algorithms have been applied successfully in different domains of routing problems. One example is the work of Durand et al. [23] in which the authors attempted to optimize the solution of air traffic conflicts; Jiajia & Zaien [24] performed a study for traffic signal timing optimization using ant-based algorithms and obtaining promising results. Further examples include the AntNet algorithm described in [25] which addresses the routing problem in packet-switched networks and the work of Jagadeesh et al. [26] that introduced a hierarchical routing algorithm that computes a near-optimal route in a large city road network.

ISSN 2395-8618

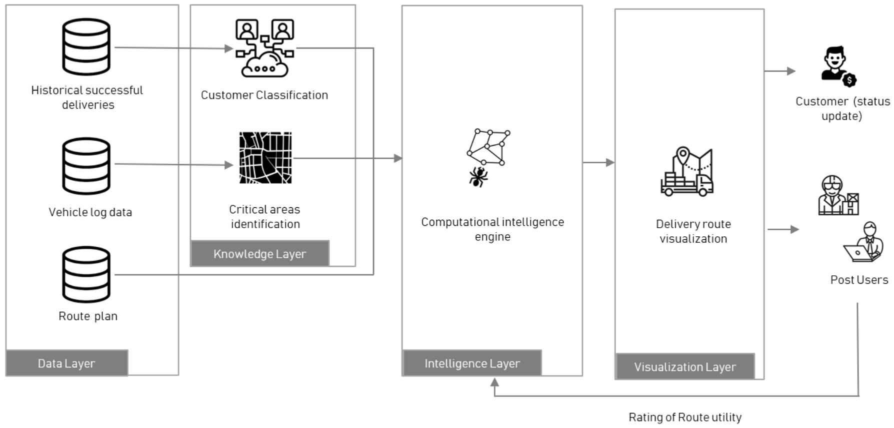

Fig. 1. Building blocks and interplay of the proposed framework.

1) Fuzzy Ant System: The ant colony optimization (ACO) is a metaheuristic that allows solving complex combinatorial optimization problems [22]. It was conceived by Dorigo and Di Caro [27], the researchers applied the ant system to the well-known Traveling Salesman Problem (TCP).

To solve the TSP problem using ACO, artificial ants are the entities in charge of finding a solution in the solution space. At the beginning of the search (time $t = 0$), the ants are located in different towns (nodes) that need to be visited or served by the salesman; then when an ant moves to an unvisited town, it leaves a pheromone (trail intensity) which helps other ants to decide the paths to choose in the future depending on the pheromone intensity [22]. Moreover, the next movement or so-called transition probability $p_{ij}^k(t)$ is defined by the following expression [28], [29]:

$$p_{ij}^k(t) = \begin{cases} \frac{\tau_{ij}(t)^{\alpha} {\eta_{ij}}^{\beta}}{\sum_{h \in \Omega_i^k(t)} \tau_{ih}(t)^{\alpha} {\eta_{ih}}^{\beta}}, & if\ j \in \Omega_i^k(t) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $\Omega_i^k(t)$ corresponds to all the possible nodes to be visited by an ant k; d_{ij} is the Euclidean distance between nodes i and j; $\eta_{ij} = 1/d_{ij}$ is the *visibility* and α, β are parameters that represent the importance given to the trail intensity and visibility.

One important concept from the previous explanation is one of visibility. It refers to the desirability of choosing a city j while currently being in town i [30], meaning that visibility is based on local information. In the case of trail intensity, the more important it is given, the more desirable the link becomes since many ants have already passed that way [22]. Moreover, every ant will complete a traveling salesman tour after n iterations; m iterations of the algorithm are called a cycle and after each cycle, the trail intensity needs to be updated, following the natural behavior of evaporation of pheromones. There are different methods to update the trail intensity (see [30], [27], [22]), and this is made with the goal of discovering a good solution through cooperation.

Teodorović and Lučić [28], [29] go one step further and propose the fuzzy ant system (FAS), which resulted from the combination of ACO with concepts of fuzzy logic. The principles are the same, differing on the way how the utility is calculated to visit the next node. The authors assumed that the ants can perceive the distance between nodes as *small*, *medium* and *big*, whereas the trail intensity can be anticipated as *weak*, *medium* or *strong*. Furthermore, an approximate reasoning algorithm to compute the utility when choosing the next link is composed of fuzzy rules similar to the following:

If d_{ij} is SMALL and τ_{ij} is STRONG
Then u_{ij} is VERY HIGH

where d_{ij} is the distance between node i and node j, τ_{ij} represents the pheromone along the link (i, j), and u_{ij} corresponds to the ant's utility when choosing the node j, considering that the ant is located in the node i.

With fuzzy logic acting as a separate module within an ACO implementation is possible to handle the uncertainty present in complex combinatorial problems [22]. Applying fuzzy rules reduces complexity and allows us to represent nature and reality with higher fidelity, leading to the generation of good implementations that can find solutions in a reasonable computational time [28].

D. Basis of the Proposed Solution

The fuzzy ant system proposes the usage of linguistic variables to approximate the distance between nodes when deciding which place to visit. Taking that approach further and landing it to the last-mile delivery, the authors of this research work propose an practical implementation on boundary of the principles of FAS, by adding further variables that come into play when performing the delivery of packages to households.

As illustrated in Figure 1, the intelligence layer of the proposed framework uses a customer classification according to the success of past deliveries and information related to critical traffic areas. Such concepts are uncertain and imprecise; thus, their representation as fuzzy variables is

ISSN 2395-8618

coherent. Traffic status can be represented in terms of linguistic variables as *"no traffic"*, *"normal traffic"*, *"heavy traffic"*, *"extremely heavy traffic"* for example; whereas the delivery success of parcels in the household could be defined as *"always unsuccessful"*, *"rarely successful"*, *"usually successful"*, *"most of the times successful"*, and *"always successful"*. Possible membership functions of these fuzzy sets are shown in Figures 2 and 3.

When an ant has to decide about the utility of the next link, meaning in this context the next address to visit to deliver a package, it can act in accordance with fuzzy rules composed in the following manner for example:

If d_{ij} is SMALL and τ_{ij} is STRONG and
ϕ_j is ALWAYS SUCCESSFUL and θ_{ij} is NO TRAFFIC
Then u_{ij} is VERY HIGH

where d_{ij} is the distance between node i and node j, τ_{ij} represents the pheromone along the link (i, j), in the same manner as the FAS proposed by Teodorović and Lučić [28]. ϕ_j represents the delivery success of the customer located in node j, and θ_{ij} is the traffic status when transiting from node i to j; u_{ij} corresponds to the ant's utility when choosing the node j, considering that the ant is located in the node i.

The usage of fuzzy rules of this type will allow to address the issues that partially known input brings along.

Furthermore, adapting the FAS [28], [29] algorithm to the conditions of the problem addressed in this work, the delivery routes can be created in the following way:

- **Step 1**: Describe the past delivery success of the households to be visited, in linguistic terms (see 2). Set the counter of the cycles to zero ($c = 0$)
- **Step 2**: Define the number the numbers of cycles C that the algorithm is going to be executed. If the number of cycles is reached, proceed to Step 4, otherwise proceed to Step 3.
- **Step 3:** Set the counter of ants to one ($k = 1$). All m ants are to be located at the starting point. Generate m sets of routes by m ants. Each ant generates a route. When all nodes are visited, ant k will finish with the route design. Increase the ant counter by one after creating one set of the routes. If the ant counter is equal to $m + 1$, increase the cycle counter by one and go to step 2. Otherwise, the next ant creates the set of routes within the considered cycle.
- **Step 4:** Take the routes that perform better in terms of time and that provide allow a higher delivery hit-ratio.
- **Step 5:** Recommend the top-performing routes to the delivery team.

For Step 5, it is possible to rank the top-performing routes found by the artificial ants in terms of the utility perceived by the delivery team. Thus, it is coherent to ask them if the computed routes improve the delivery ratio when delivering parcels. People from the delivery team could make use of natural language expressions to express their satisfaction with the suggested route (i.e., *"not useful"*, *"somewhat useful"*, and *"useful"*) in this way, routes that might not represent any improvements can be identified and adjusts can be made.

E. Evaluation

Controlled experiments, simulations and case studies can be performed to evaluate the quality of the results obtained from the implementation of the artifact. Moreover, given that it is proposed to execute this research work following a transdisciplinary approach, the practice partners are also to be involved in the evaluation of the solution, and thus, qualitative analysis in conjunction with interviews and expert opinions and are also to be performed to measure the performance of the framework.

IV. Outlook and Conclusions

In this article a framework that enables the improvement of the first-try delivery by studying the behavior of traffic on the streets and customers' presence at home. In contrast to existing solutions, it is proposed working only with data that does not compromise the customers' privacy as (i.e., avoiding the use of tracking data) and getting insights about traffic characteristics without the need of deploying a large number of vehicles or expensive sensors. The main goal is to provide a delivery route plan to delivery teams and route planners, which allows finishing the distribution of the parcels in the least amount of time while being able to effectively deliver the highest amount of them. This will be translated into less resource consumption and possible increased customer satisfaction.

The work to be conducted to implement the framework is practice-oriented and aimed at identifying and solving problems that affect societies, industries, and cities, while also expanding knowledge from it. Having the support of practice partners (transdisciplinarity) eases the fact of testing the developed solutions in real-life scenarios, and with that, it is possible to refine them towards finding an optimal solution to the problems addressed.

Moreover, with the execution of this research work, it is expected to contribute to developments in the field of smart logistics, in the assessment of how approximative methods (i.e., fuzzy logic, and computational intelligence) can be leveraged towards building working solutions that do not need large amounts of precise information (i.e., machine learning solutions), and additionally, the results to be obtained can be used as a basis to develop further solutions in the field of green logistics and urban planning. In the practical terms, this project will provide to our practice partners a tool to perform better delivery route plans which translate into less failed delivery, less resource consumption, and therefore, a possible nation-wide more sustainable delivery service.

Acknowledgment

The authors would like to thank the members of the Human-IST Institute at the University of Fribourg for contributing with valuable thoughts and comments. We especially thank

ISSN 2395-8618

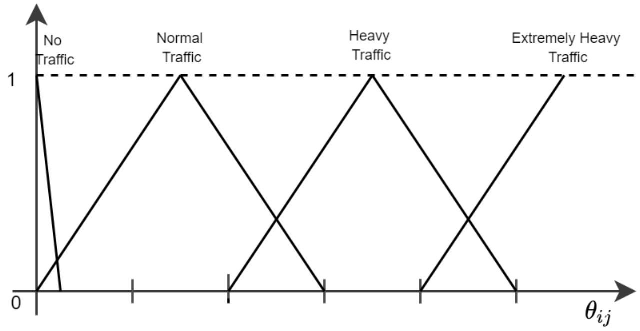

Fig. 2. Membership functions of the fuzzy sets depicting traffic condition between nodes.

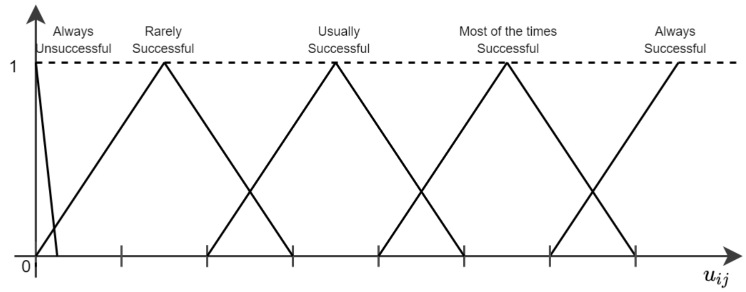

Fig. 3. Membership functions of the fuzzy sets depicting delivery success of parcels in the household.

the Secretariat of High Education, Science, Technology, and Innovation (SENESCYT) of Ecuador and the *Swiss Post* for their support to conduct this research.

References

[1] N. Agatz, A. Campbell, M. Fleischmann, and M. Savelsbergh, "Time slot management in attended home delivery," *Transportation Science*, vol. 45, no. 3, pp. 435–449, 2011.

[2] E. Morganti, S. Seidel, C. Blanquart, L. Dablanc, and B. Lenz, "The impact of e-commerce on final deliveries: alternative parcel delivery services in france and germany," *Transportation Research Procedia*, vol. 4, no. 0, pp. 178–190, 2014.

[3] Y. Qu, F. Wu, and W. Zong, "A vehicle routing problem with time windows for attended home distribution," in *International Conference on Data Science*. Springer, 2015, pp. 127–134.

[4] S. F. W. Lim, X. Jin, and J. S. Srai, "Consumer-driven e-commerce," *International Journal of Physical Distribution & Logistics Management*, 2018.

[5] J. Mańdziuk, "New shades of the vehicle routing problem: emerging problem formulations and computational intelligence solution methods," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 3, pp. 230–244, 2018.

[6] R. Mangiaracina, A. Perego, A. Seghezzi, and A. Tumino, "Innovative solutions to increase last-mile delivery efficiency in B2C e-commerce: A literature review," *International Journal of Physical Distribution & Logistics Management*, 2019.

[7] J. Van Duin, W. De Goffau, B. Wiegmans, L. Tavasszy, and M. Saes, "Improving home delivery efficiency by using principles of address intelligence for B2C deliveries," *Transportation Research Procedia*, vol. 12, no. 2, pp. 14–25, 2016.

[8] S. Pan, V. Giannikas, Y. Han, E. Grover-Silva, and B. Qiao, "Using customer-related data to enhance e-grocery home delivery," *Industrial Management & Data Systems*, 2017.

[9] L. A. Zadeh, "Fuzzy logic, neural networks, and soft computing," in *Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh*. World Scientific, 1996, pp. 775–782.

[10] G. H. Hadorn, H. Hoffmann-Riem, S. Biber-Klemm, W. Grossenbacher-Mansuy, D. Joye, C. Pohl, U. Wiesmann, and E. Zemp, *Handbook of transdisciplinary research*. Springer, 2008, vol. 10.

[11] A. Hevner and S. Chatterjee, "Design science research in information systems," in *Design research in information systems*. Springer, 2010, pp. 9–22.

[12] A. Abbas, L. Zhang, and S. U. Khan, "A survey on context-aware recommender systems based on computational intelligence techniques," *Computing*, vol. 97, no. 7, pp. 667–690, 2015.

[13] E. Portmann, "Fundamentals of fuzzy clustering methods," in *The FORA Framework*. Springer, 2013, pp. 37–61.

[14] A. M. Florio, D. Feillet, and R. F. Hartl, "The delivery problem: Optimizing hit rates in e-commerce deliveries," *Transportation Research Part B: Methodological*, vol. 117, pp. 455–472, 2018.

[15] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[16] W. Dong and A. Pentland, "A network analysis of road traffic with vehicle tracking data," in *AAAI Spring Symposium: Human Behavior Modeling*, 2009, pp. 7–12.

[17] J. Yoon, B. Noble, and M. Liu, "Surface street traffic estimation," in *Proceedings of the 5th international conference on Mobile systems, applications and services*. ACM, 2007, pp. 220–232.

[18] J. Pincay, A. O. Mensah, E. Portmann, and L. Terán, "Partitioning space to identify en-route movement patterns," in *2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG)*. IEEE, 2020, pp. 43–49.

[19] ——, "Forecasting travel times with space partitioning methods," in *GISTAM*, 2020, pp. 151–159.

[20] Y.-z. Chen, S.-f. Shen, T. Chen, and R. Yang, "Path optimization study for vehicles evacuation based on dijkstra algorithm," *Procedia Engineering*, vol. 71, pp. 159–165, 2014.

ISSN 2395-8618

[21] B. Tatomir and L. Rothkrantz, "Hierarchical routing in traffic using swarm-intelligence," in *2006 IEEE Intelligent Transportation Systems Conference*. IEEE, 2006, pp. 230–235.

[22] D. Teodorović, "Swarm intelligence systems for transportation engineering: Principles and applications," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 6, pp. 651–667, 2008.

[23] N. Durand and J.-M. Alliot, "Ant colony optimization for air traffic conflict resolution," in *ATM Seminar 2009, 8th USA/Europe Air Traffic Management Research and Developpment Seminar*, 2009.

[24] J. He and Z. Hou, "Ant colony algorithm for traffic signal timing optimization," *Advances in Engineering Software*, vol. 43, no. 1, pp. 14–18, 2012.

[25] G. Di Caro and M. Dorigo, "AntNet: Distributed stigmergetic control for communications networks," *Journal of Artificial Intelligence Research*, vol. 9, pp. 317–365, 1998.

[26] G. R. Jagadeesh, T. Srikanthan, and K. Quek, "Heuristic techniques for accelerating hierarchical routing on road networks," *IEEE Transactions on intelligent transportation systems*, vol. 3, no. 4, pp. 301–309, 2002.

[27] M. Dorigo and G. Di Caro, "Ant colony optimization: a new meta-heuristic," in *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)*, vol. 2. IEEE, 1999, pp. 1470–1477.

[28] D. Teodorović and P. Lučić, "Schedule synchronization in public transit using the fuzzy ant system," *Transportation Planning and Technology*, vol. 28, no. 1, pp. 47–76, 2005.

[29] D. TEODOROVIĆ and P. LUČIĆ, "The fuzzy ant system for the vehicle routing problem when demand at nodes is uncertain," *International Journal on Artificial Intelligence Tools*, vol. 16, no. 05, pp. 751–770, 2007.

[30] E. Bonabeau, M. Dorigo, D. d. R. D. F. Marco, G. Theraulaz, G. Théraulaz *et al.*, *Swarm intelligence: from natural to artificial systems*. Oxford university press, 1999, no. 1.

ISSN 2395-8618

Journal Information and Instructions for Authors

I. Journal Information

Polibits is a half-yearly open-access research journal published since 1989 by the *Centro de Innovación y Desarrollo Tecnológico en Cómputo* (CIDETEC: Center of Innovation and Technological Development in Computing) of the *Instituto Politécnico Nacional* (IPN: National Polytechnic Institute), Mexico City, Mexico.

The journal has double-blind review procedure. It publishes papers in English and Spanish (with abstract in English). Publication has no cost for the authors.

A. Main Topics of Interest

The journal publishes research papers in all areas of computer science and computer engineering, with emphasis on applied research. The main topics of interest include, but are not limited to, the following:

- Artificial Intelligence
- Natural Language Processing
- Fuzzy Logic
- Computer Vision
- Multiagent Systems
- Bioinformatics
- Neural Networks
- Evolutionary Algorithms
- Knowledge Representation
- Expert Systems
- Intelligent Interfaces
- Multimedia and Virtual Reality
- Machine Learning
- Pattern Recognition
- Intelligent Tutoring Systems
- Semantic Web
- Robotics
- Geo-processing
- Database Systems
- Data Mining
- Software Engineering
- Web Design
- Compilers
- Formal Languages
- Operating Systems
- Distributed Systems
- Parallelism
- Real Time Systems
- Algorithm Theory
- Scientific Computing
- High-Performance Computing
- Networks and Connectivity
- Cryptography
- Informatics Security
- Digital Systems Design
- Digital Signal Processing
- Control Systems
- Virtual Instrumentation
- Computer Architectures

B. Indexing

The journal is listed in the list of excellence of the CONACYT (Mexican Ministry of Science) and indexed in the following international indices: Web of Science (via SciELO citation index), LatIndex, SciELO, Redalyc, Periódica, e-revistas, and Cabell's Directories.

There are currently only two Mexican computer science journals recognized by the CONACYT in its list of excellence, *Polibits* being one of them.

II. Instructions for Authors

A. Submission

Papers ready for peer review are received through the Web submission system on www.easychair.org/conferences/?conf=polibits1; see also updated information on the web page of the journal, www.cidetec.ipn.mx/polibits.

The papers can be written in English or Spanish. In case of Spanish, author names, abstract, and keywords must be provided in both Spanish and English; in recent issues of the journal you can find examples of how they are formatted.

The papers should be structures in a way traditional for scientific paper. Only full papers are reviewed; abstracts are not considered as submissions. The review procedure is double-blind. Therefore, papers should be submitted without names and affiliations of the authors and without any other data that reveal the authors' identity.

For review, a PDF file is to be submitted. In case of acceptance, the authors will need to upload the source code of the paper, either Microsoft Word or LaTeX with all supplementary files necessary for compilation. Upon acceptance notification, the authors receive further instructions on uploading the camera-ready source files.

Papers can be submitted at any moment; if accepted, the paper will be scheduled for inclusion in one of forthcoming issues, according to availability and the size of backlog.

See more detailed information at the website of the journal.

B. Format

The papers should be submitted in the format of the IEEE Transactions 8x11 2-column format, see http://www.ieee.org/publications_standards/publications/authors/author_templates.html. (while the journal uses this format for submissions, it is in no way affiliated with, or endorsed by, IEEE). The actual publication format differs from the one mentioned above; the papers will be adjusted by the editorial team.

There is no specific page limit: we welcome both short and long papers, provided that the quality and novelty of the paper adequately justifies its length. Usually the papers are between 10 and 20 pages; much shorter papers often do not offer sufficient detail to justify publication.

The editors keep the right to copyedit or modify the format and style of the final version of the paper if necessary.

See more detailed information at the website of the journal.