Volume 4, Issue 2
November 2017
ISSN: 1390-9266

LAJC

LATIN-AMERICAN JOURNAL OF COMPUTING

FACULTAD DE INGENIERÍA DE SISTEMAS

QUITO - ECUADOR

Editor in Chief:

PhD. Jenny Torres, Escuela Politécnica Nacional, Ecuador

http://lajc.epn.edu.ec/

ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS

Estimation of the Contaminant Risk Level of Petroleum Residues Applying FDA Techniques

Miguel Flores, Ana Escobar, Luis Horna, and Lucía Carrión

***Abstract*—In the process of oil extraction, specifically in the refinement and industrialization of hydrocarbons, as is known, multiple wastes are highly polluting for the soil, water and air.**

In this work, the risk level of these wastes in affected areas is estimated thanks to the application of statistical models in the field of functional data analysis. These models have been implemented in a statistical software called $RStudio$ that allows an early measurement and evaluation of the level of risk by using semiquantitative and quantitative methods. This measurement is carried out by the staff of PETROECUADOR close to the affected place. It was used the laser-induced fluorescence technique (LIF). The data obtained using this technique was used to adjust the following models: Generalized Functional Linear Model (MLFG), which makes it possible to classify the spectrum generated in two pollution levels: Low and High. Functional linear regression model with scalar response and functional explanatory variable with the aim of directly estimating the percentage of contamination level. With these results it is verified that the shape of the laser fluorescence spectrum is highly related to the gasoline content in the sample.

***Index Terms*—Quality Control, Generalized Linear Functional Model, Linear Regression, Classification**

I. Introduction

THe filtration of oil (or its derivatives), transport and diffusion-dispersion are processes whose study is of vital importance due to the great impact they have on human activity and the environment.

The filtration of petroleum in soils causes a level of pollution that is a very complex problem to evaluate, this depends mainly on the following elements: soil type, porosity, hydraulic conductivity, and petroleum properties such as density and viscosity.[1]

For this reason, an important task is to determine the state of the system at all times, and as a priority at the initial moment, since it would allow the application of corrective measures in the ecosystem in a more efficient way.[7]

Oil is made up of a variety of compounds, some of which produce fluorescence when illuminated with ultraviolet light. The fluorescence of the petroleum depends to a great extent on its chemical composition (Celander, Freddricsson, Galle, and Svanberg, 1988). For this reason there are analytical techniques for the characterization of crude oil, in the parts that the intensity and life time of the fluorescence are related to the chemical composition and density (API) of the oil.[7]

If we combine a source of ultraviolet laser light, a spectrometer and oil, we will have a system to detect the presence of petroleum, such as contaminated lands (O'Neill, RA, Buja-Bijunos, L., Rayner, DM, 1980). Laser light produces fluorescence when there is oil in the earth, which is detected using the spectrometer. As each variety of oil has a characteristic fluorescence spectrum, fluorescence techniques are often used for identification.[1] The data obtained by this technique are used in the first instance to solve a problem of supervised classification and then to carry out a forecast of the level of contamination.

Among the statistical techniques that are used to solve a classification problem are: discriminant analysis, logistic regression and cluster analysis, depending on the objective. For example, in Anderson, Farrar, Thoms, (2009) determines the contamination of anthropogenic metals in the soil using the technique of discriminant analysis with clustered chemical concentrations. In the case of the prediction process, Lopez (2014) applies a linear regression in order to predict the air pollution of carbon dioxide produced by Hawaii's Mauna Loa volcano, in this case time is the independent variable and pollution is the dependent variable.

The statistical techniques mentioned are traditional and multivariate techniques, however, in recent times technological changes has been able to measure data in a faster and more precise way, and thanks to this evolution, it is possible to work with the functional form of the data.[2]

In this work, statistical techniques of functional data analysis (Functional data Analyzes - FDA -) are used. Specifically, a Generalized Functional Linear Model is applied to solve the classification problem and a Linear Regression Model with scalar response and functional explanatory variable to estimate the level of contamination. One of the advantages of using FDA is reducing the influence of noise or observation errors. (Ramsay, & Silverman, 2005).

Classification of functional data is one of the major branches of the FDA (Functional Data Analysis), there are two types of classification that are: supervised and unsupervised. In the case of unsupervised classification, its objective will be to make groups as homogeneous as possible, and at the same time the most distinct among them (Noguerales, 2010); the most common technique is clustering and the method for performing such technique is k-means.

For the supervised classification there are different classifiers that will help to classify the base between them: Linear Discriminant, k-NN (nearest neighbor method), Kernel,

Article history:
Received 09 September 2017
Accepted 28 November 2017

Miguel Flores is a professor, Escuela Politécnica Nacional
Ana Escobar is a student, Escuela Politécnica Nacional
Luis Horna is a professor, Escuela Politécnica Nacional
Lucía Carrión is a student, University of Technology Sydney

Table I
SAMPLES MADE FOR MFLG ADJUSTMENT AND VALIDATION

Sample	Level	Total	Sample 2	Level 2	Total 2
GE1	0.3	2	GE8	9.1	2
GE2	0.4	2	GA15	16.67	2
GE3	0.5	2	GE9	16.7	2
GA1	1.48	2	GE10	37.5	2
GA3	1.48	2	GA19	37.5	2
GE4	1.5	2	GE11	50	2
GE5	2.4	2	GA23	50	2
GA5	2.44	2	GE12	75	2
GE6	3.8	2	GE13	83.3	2
GA8	3.85	2	GA26	83.33	2
GE7	6.1	2	GE14	100	2
GA13	6.1	2	GA30	100	2
GA14	9.09	2			

PLS (generalized linear models), Generalized linear models (Noguerales, 2010).

A recent application of a Generalized Linear Functional Model can be found in Flores, Saltos and Castillo-Paz (2016), where the types of cancer are classified by DNA information.

In this study, the generalized functional linear model was used to classify the contamination level of the hydrocarbon residues, with a variable binary response. This model assumes that the content of the solid element remains constant and that it is indeformable.

It is important to note that the model has already been integrated in a software that interacts with the laser spectrometer, built by the team of engineers of the project developed by the National Polytechnic School of Ecuador.

On the other hand, the functional linear regression model with scalar response and functional explanatory variable, the model is used as predictor. This class of models has been applied to regional analysis associations as an alternative to standard multiple regression models (Luo, Zhu, Xiong (2012)).

For the development of the models, 25 tests (with two replicates) were carried out, which are differentiated by the percentage (level) of gasoline in the sample (see Table 1). For this work, if a spectrum has a gas percentage less than or equal to 10% it is classified in the low pollution group. Otherwise it is classified in the high pollution group. The model allows to consider any other level of gasoline to discriminate between spectra corresponding to samples with a low or high contamination.

II. MATERIALS AND METHODS

The methodology used in this study with functional data is that described in Bande and Fuente (2012), which consists of the following stages:

1) Explore and describe the functional data set highlighting its most important characteristics.
2) Explain and model:
   a) Find the relationship between a dependent variable and an independent variable using regression models
   b) Solve the problem of Supervised or Non-Supervised Classification of a set of data regarding some characteristic.
3) Contrast, validation and prediction.

But before describing the stages of the methodology is given a definition of a functional random variable.

Let X random variable it is functional if it takes values in the space which can be normalized and semi normalized. (Bande and Fuente, 2012).

One set of functional data $\{x_1, ..., x_n\}$ is the observation of n functional variables $\{X_1, ..., X_n\}$ Identically distributed (Bande and Fuente, 2012).

The most aware to work with functional data is to determine the space where it works to use right statistical techniques.

Let $T = [a, b] \subset \mathbb{R}$. It is generally assumed that there are elements of :

$$\mathcal{L}^2 = \{X : T \to \mathbb{R},\ such\ that\ \int_T |X|^2 dx < \infty\} \quad (1)$$

A. FDA exploratory analysis

The first step is the exploratory analysis of the data to make their characteristics known and know exactly how to manipulate them.

The methods of representation are: decrementation and the choice of a reduced basis of functions. One way to represent the functional data is in a nonparametric way. And in most cases this is the best representation.

In this work, we used the B-spline base to represent the functional data. And in most cases this is the best representation. It is given by a smoothing matrix S:

$$S_{ij} = \frac{1}{h} K(t_i - t_j)h \quad (2)$$

Where h is the bandwidth, which is calculated with cross-validation.

There are different kernel types $K()$, however the most used is **Gaussian**

$$K(u) = \frac{1}{\sqrt{2}} exp(-\frac{u^2}{2}) \quad (3)$$

This is to approximate and have the best representation of $\mathcal{X}$:

$$\widehat{x} = \sum_{i-1}^{n} s_i(x) Y_i \quad (4)$$

There are different methods for calculating $s_i()$: the nearest neighbor, Nadaraya-Watson, local linear regression.

X is defined as the functional variable of interest, spectrum generated through the LIF technique, which takes values in a normalized (or semi-normalized) space F, and is considered as functional data to the results of the 25 tests represented as the set $x_1, x_2, ..., x_n$ that come from n functional variables $X_1, X_2, ..., X_n$ identically distributed as X.

Definition 2. A base is a set of known functions $(\Phi_k)_{k \in N}$ such that any function can be approximated as well as desired by a linear combination of K of them with sufficiently large K.

In this way functional observation can be approximated as

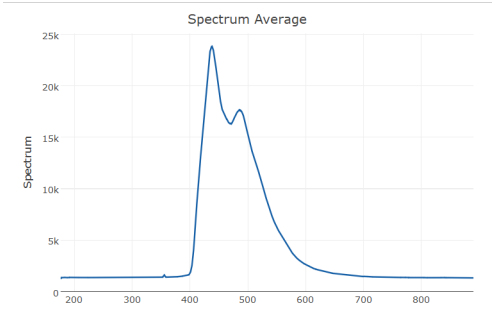


Figure 1. Functional average:

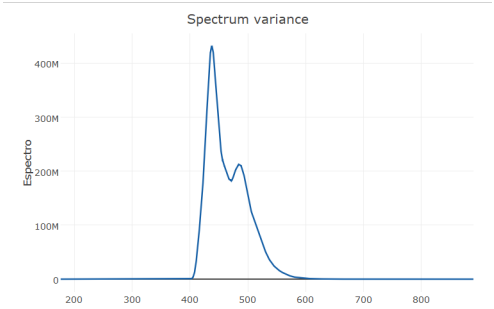


Figure 2. Functional variance:

$$X(t) = \sum_{k=1}^{n} c_k F_k(t) \tag{5}$$

The type of base will depend on the nature of the data, it is very common to use B-Spline bases, second step is to describe the functional data, functional exploratory analysis is used, in which several estimators such as the mean and the functional variance.

Then, depending on the purpose of the study, techniques such as functional regression, classification models, and others are used.

The present document aims to provide an overview of these techniques, in order to present the usefulness of their application to the environmental context.

The results are obtained through the statistical software R and its packets, such as the packet, $'fda.usc'$.

The mean and functional variance are defined below: Let $x_i(t)$, $i = 1, 2, ..., N$ be a sample of functional data curves, the mean and variance are given by (Plazola, 2013):

The FDA concepts and techniques used in this paper can be found in the books of Ramsay and Silverman, 2002 and Ramsay and Silverman, 2006. In both cases, all the included techniques are restricted to the space of $\mathcal{L}^2$ functions (the Hilbert space of all square integrable functions over a certain interval). The book by Ferraty and Vieu is another important reference that incorporates non-parametric approaches, as well as the use of other theoretical tools like Semi-norms that allow us to deal with norms or metric spaces.

X is defined as the functional variable of interest, spectrum generated through the LIF technique, which takes values in a

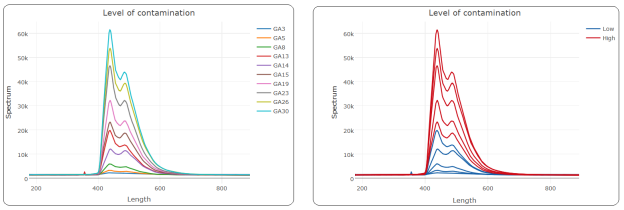


Figure 3. Spectra by level of contamination

normalized (or semi-normalized) space F, and is considered as functional data to the results of the 25 tests represented as the set $x_1, x_2, ..., x_n$ that come from n functional variables $X_1, X_2, ..., X_n$ identically distributed as X.

The functional data are discretized in a total of $3'648$ points that are in the range $[176.39, 890.62]$. These are represanted by the set of points t_j. In Figure 1 we can see in blue color the spectra of low level and in red color the spectros of high level.

It can be seen from the figure that the spectra tend to have the same shape however the spectra of high level have a greater amplitude than the low ones and it is increasing in relation to the percentage of gasoline that is in the sample. This varies from 16.67% to 100% (greater than 10%); on the other hand, the amplitude of the spectra of the low level samples are decreasing according to the percentage of gasoline in the sample. A spectrum is considered with a low level of pollution if the percentage of gasoline is less than 10%, that is from 3% to 9.1%.

For the representation of the functional data, a B-spline base was used using the $fda.usc$ package of the statistical software R (Bande and Fuente, 2012).

B. Generalized Functional Linear Model (MFLG)

Once the spectra are represented to functional data, a Generalized Linear Functional Model (MFLG) is fitted to estimate the probability that it belongs to one of the two groups. For the adjustment and implementation of the model, the $'fregre.glm'$ function of the $fda.usc$ package of the statistical software R was used.

The MFLG is also known in the literature as Functional Logistic Regression (FLR).

The model explains the relationship between Y (binary response) and a functional covariate $X(t)$ with representation based on $X(t)$ and $\beta(t)$. Π_i is the probability of occurrence of the event $Y_i = 1$, which in this case corresponds to a high contamination, conditioned to the covariate $X_i(t)$, which is expressed as follows:

$$Y_i = \pi_i + \epsilon_i, \qquad where\ i = 1, ..., n \tag{6}$$

$$\pi_i = P\left[Y = \frac{1}{x_i(t)} : t \in T\right] \tag{7}$$

$$= \frac{\exp(\int_T X_i(t)\beta(t)dt)}{1 + \exp(\int_T X_i(t)\beta(t)dt)} \qquad i = 1, \ldots, n \tag{8}$$

Where ϵ_i are independent errors with zero mean. It is defined as a functional covariate the spectrum denoted by: $X = X(t)$, and as a scalar (binary) response variable the

	High	Low
High	6	0
Low	0	5

type of pollution denoted by Y (0 = Low pollution, 1 = High pollution).

In this case, since the MFLG works with a binary response variable, this model provides a classification rule for the type of contamination (Bayes rule).

C. Linear regression with scalar response variable and functional explanatory variable

For this model the objective will be to understand how a response variable Y being this scalar is related to a vector of variables $X \in \mathbb{R}^p$

Therefore, the regression model is defined as follows (Ramírez, 2014),

$$Y = \langle X, \beta \rangle + \epsilon \tag{9}$$

$$Y = \langle X, \beta \rangle + \epsilon \tag{10}$$

$$= \frac{1}{\sqrt{T}} \int_T x(t)\beta(t)dt + \epsilon \tag{11}$$

In our case study we will analyze the following model

Where, $\langle \cdot, \cdot \rangle$we denote the usual internal product defined in $\mathcal{L}^2$ and ϵ is the random error with mean zero and variance σ^2. Specifically, we will perform a non-parametric functional regression model. An alternative to model (5) is

$$y_t = r(X_t(t)) + \epsilon_t \tag{12}$$

Where, the unknown real soft function is estimated using the Kernel estimate.

For the regression model it is considered as a scalar response variable the percentage of water presented by gasoline and our functional explanatory variable is the spectrum. To obtain the corresponding estimates, the $'fregre.np'$ function of the $fda.usc$ package has been used.

D. Validation of the models

For the validation of the statistical models, two samples are used.

A sample of training and validation. For each test taken, two replicates of the spectrum were obtained.

These are used as follows: one of the replicas is used for the training of the model and the other replica for the validation of the model.

In this work, to carry out the validation of the model a confusion matrix is used, with the following structure:

High Low / High True False negatives /Low False positives True negatives).

Where, true positives and true negatives correspond to correctly classified spectra in high and low contamination respectively, while false negatives and false positives are misclassified spectra by our model; Using the data of the spectra we obtain:

As seen in the previous matrix, the model has an efficiency of 100%.

On the other hand, for the validity of the functional linear regression model with scalar response and functional explanatory variable, the coefficient of determination R^2 was used, where a value of 99% was obtained, this implies that the model correctly explains the variability of the data in that percentage. We also calculate the mean absolute percentage error(MAPE) with nonparametric regressions, 7% was obtained.

III. RESULTS AND DISCUSSION

Prior to the modeling, an exploratory analysis of the data in the functional field has been carried out, i.e. the mean and functional variance for all data, as well as for the High and Low groups, have been estimated. The way to estimate these descriptive measures is in Gonzalez-Manteiga and Vieu, 2007.

From Figure 4, it can be seen that there is a clear distinction between the defined groups. This result facilitates and confirms the use of a functional supervised classification model.

In Figure 5, a graph of probabilities resulting from applying the GLFM model is presented.

Where it is appreciated that the spectra with a percentage less than 10% are classified as low level while the rest of the spectra are classified as high level. Therefore, the model correctly classifies 100% of the spectra in each group (Low and High). Figure 6 shows the level of contamination that the oil sample will present. It is important to mention that several tests were done with different methods trying to find the one that fit the best. It is also observed the two curves, both the estimation with the model and the actual pollution, coincide in almost every point.

IV. CONCLUSIONS

As mentioned in the introduction each sample has two replicates, one of which is used for model estimation and the other for its validation. In the case of the sample for the estimation we have that the percentage of correctly classified spectra in each group (Low and High) is 100%, while for the validation sample the percentage of correctly classified spectra is 99% with a MAPE of 7%.

It has been verified that the shape of the laser fluorescence spectrum is highly related to the gasoline content of the sample.

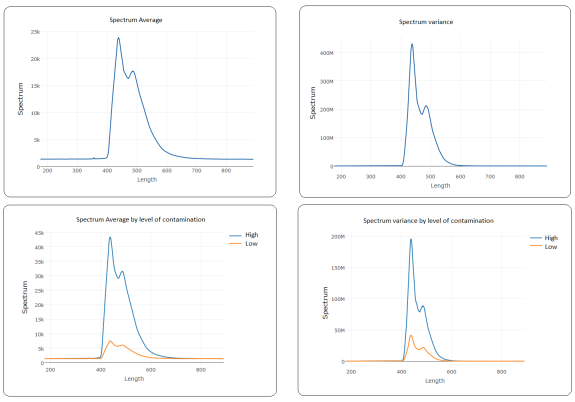


Figure 4. Functional descriptive measures by level of contamination.

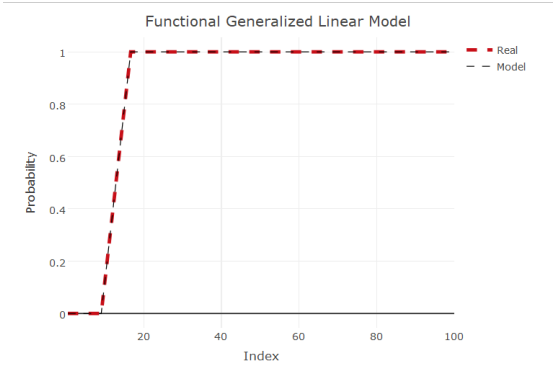


Figure 5. Estimated probabilities.

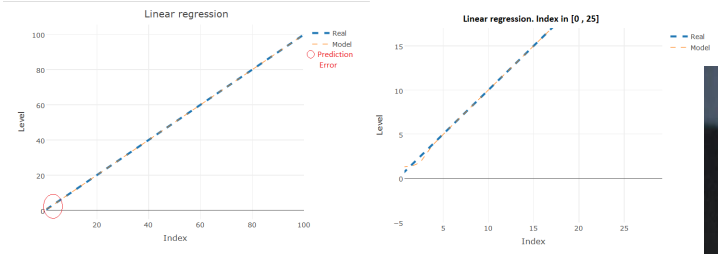


Figure 6. Actual pollution level vs. estimated by the functional regression model.

Therefore, due to its functional nature, the application of supervised FDA classification techniques provides a reliable solution for the identification of a high or low risk of contamination in potentially affected areas.

When applying the functional regression model, we have managed to explain the 99% R^2 of the variability, in addition to reach this result has been tested with several models. The corresponding validation tests of the model were also performed, which were statistically significant.

REFERENCES

[1] Celander, K., Freddricsson, B. Galle, S. y Svanberg. (1988). Investigation of Laser-Induced Fluorescence with application to remote sensing of environmental parameters, Goteborg Institute of Physics Reports GIPR-149.

[2] González-Manteiga, W. y Vieu, P. (2007). Statistics for functional data. Computational Statistics and Data Analysis, 51, 4788-4792.

[3] Li, J., Cuesta-Albertos, J. A., & Liu, R. Y. (2012). DD-classifier: Nonparametric classification procedure based on DD-plot. Journal of the American Statistical Association. Vol. 107, 737-753.

[4] López Miranda Claudio y Cesar Augusto Romero Ramos, (2014). Propuesta de proyecto de estadística: un modelo de regresión lineal simple para pronosticar la concentración de co2 del volcán Mauna Loa. EPISTEMUS, 17:63-69.

[5] [Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software, 51(4):1-28.

[6] Miguel Flores, Guido Saltos and Sergio Castillo-Paéz, (2016), Setting a generalized functional liner model (GFLM) for classification of different types of cancer, Latin American Journal Computing, 3 (2):41-48.

[7] O'Neill, R.A., Buja-Bijunos, L., Rayner, D.M. (1980). Field Performance of laser flourosensor for detection of oil spills. Appl. Opt. 19,863.

[8] Ramsay, J. O. and Silverman, B. W.2005. "Functional Data Analysis", 2nd ed., Springer-Verlag, New York, Pp. 147-325.

[9] Ramírez John. Matemática. (2014), Regresión funcional mediante bases obtenidas por descomposición espectral del operador covarianza, Matemáticas, 12 (2):15-27.

[10] R.H. Anderson, D.B. Farrar, S.R. Thoms, (2009). Application of discriminant analysis with clustered data to determine anthropogenic metals contamination. Elsevier, 408:50-56.

[11] Muñoz Dania, Silva Francisco, Hernández Noslen, Bustamante Talavera. (2014). Functional Data Analysis as an Alternative for the Automatic Biometric Image Recognition: Iris Application. Computación y Sistemas, 18 (1):111-121

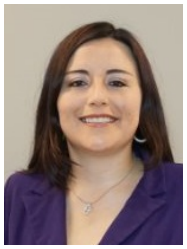
V. ACKNOWLEDGMENTS

The authors are grateful for the funding provided by the National Polytechnic School of Ecuador for the implementation of the project PII-DM-002-2016: 'Analysis of functional data in statistical quality control'.

Miguel Flores is a professor at the National Polytechnic School and a researcher at the Center for Modeling Mathematics at the National Polytechnic School in Quito, Ecuador. He is a BSc. in Statistical Computing Engineer from the Polytechnic School of the Coast. In 2006 he received an in MSc. in Operations Research from the National Polytechnic School, and in 2013 received a MSc. in Technical Statistics from the University of A Coruña. He is currently a doctoral student at the University of A Coruña in the area of Statistics and Operations Research. He has over 15 years professional experience in various areas of Statistics, Computing and Optimization, multivariate data analysis, econometric, Market Research, Quality Control, definition and construction of systems indicators, development of applications and optimization modeling.

Ana Julia Escobar Research assistant of the Mathematical Engineering career Master in Mathematics and interactions at the Paris-Saclay University. Data scientist Jr. Specialist in Operations Research.

Luis Horna Huaraca Mathematician of the Escuela Politénica Nacional, 1979. PhD. in Physical-Mathematical Sciences by the Lomonosov State University of Moscow, Rusia, 1985. Principal Professor attached to the Department of Mathematics, Faculty of Sciences, Escuela Politénica Nacional.



Lucia Carrion Gordon She received the B.S. and M.S. degrees in Systems Engineering and IT Management from the Pontifical Catholic University of Ecuador, Quito, in 2004 and 2010. She is currently PhD researcher at UTS University of Technology Sydney. Self-motivated Engineering graduate who enjoys collaborating with team members in order to achieve successful project outcomes. Adaptive, dedicated and fast-learning person. From 2011 to 2016, she was a Principal Lecturer in Ecuador, Austria and Australia in several universities. She is the author of seven book chapters, twelve conference papers and four proceedings. Her research interests include WSN Wireless Sensor Networks, IoT Internet of Things and DHP Data Heritage Preservation. Mrs. Carrion Gordon was a recipient of two awards in Australia in 2016. He is an author and co-author of several innovative investigations. She collaborated with research groups in University of Applied Sciences Upper Austria and University of Vienna. Her recent work in Heritage Preservation focus in the redefinition of terms and process with recognition in the research community

A novel approach based on multiobjective variable mesh optimization to Phylogenetics

Cristian Zambrano-Vega, Byron Oviedo Bayas, Stalin Carreño, Amilkar Puris, and Oscar Moncayo

***Abstract*—One of the most relevant problems in Bioinformatics and Computational Biology is the search and reconstruction of the most accurate phylogenetic tree that explains, as exactly as possible, the evolutionary relationships among species from a given dataset. Different criteria have been employed to evaluate the accuracy of evolutionary hypothesis in order to guide a search algorithm towards the best tree. However, these criteria may lead to distinct phylogenies, which are often conflicting among them. Therefore, a multi-objective approach can be useful. In this work, we present a phylogenetic adaptation of a multiobjective variable mesh optimization algorithm for inferring phylogenies, to tackle the phylogenetic inference problem according to two optimality criteria: maximum parsimony and maximum likelihood. The aim of this approach is to propose a complementary view of phylogenetics in order to generate a set of trade-off phylogenetic topologies that represent a consensus between both criteria. Experiments on four real nucleotide datasets show that our proposal can achieve promising results, under both multiobjective and biological approaches, with regard to other classical and recent multiobjective metaheuristics from the state-of-the-art.**

***Index Terms*—Multiobjective Optimization, Phylogenetic Inference, Evolutionary Computation, Bioinformatics.**

I. Introduction

The evolutionary history of mankind and all other living and extinct species on earth is a question which has been preoccupying mankind for centuries. Therefore, the construction of a "tree of life" comprising all living and extinct organisms on earth has been a fascinating and challenging idea since the emergence of evolutionary theory [1].

Typically, evolutionary relationships among organisms are represented by an evolutionary tree. Phylogenetic inference consists in finding the best tree that explains the genealogical relationships or evolutionary history of species from molecular sequences (DNA or protein data). The data used in this analysis usually come from aligned nucleotide or aminoacid sequences called Multiple Sequence Aligned [2], [3].

Various scientific fields can benefit thanks to the contributions of phylogenetic, such as evolutionary biology, physiology, ecology, paleontology, biomedicine, chemistry and others [4]. For all this, many scientists agree that phylogenetic inference is one of the most important research topics in Bioinformatics.

Article history:
Received 13 September 2017
Accepted 28 November 2017

The authors are professors at Universidad Técnica Estatal de Quevedo, Quevedo, Los Ríos, Ecuador.
Email: czambrano, boviedo, sdcarreno, apuris, omoncayo}@uteq.edu.ec

Handl et al. [5] discussed the applications of multiobjective optimization in several bioinformatics and computational biology problems, in this survey phylogenetic inference is one of the central problems in this area. Unfortunately, many interesting problems and algorithms in Bioinformatics, such as inference of perfect phylogenies or optimal multiple sequence alignment are NP-complete and computationally extremely intensive.

Recently several multiobjective proposed applied to phylogenetic inference have been published oriented to optimize trees under reconstruction criteria Maximum Parsimony and Maximum Likelihood: two bio-inspired techniques based in swarm intelligence algorithms: *MOABC* [4] an adaptation of the Artificial Bee Colony (ABC) and *Mo-FA* [6] a multiobjective adaptation of the novel Firefly Algorithm; and two other techniques based on the popular multi-objective metaheuristic the fast non-dominated sorting genetic algorithm (NSGAII): *PhyloMOEA* [7] and *MO-Phyl* [8] a hybrid OpenMP/MPI parallel technique.

In this work we present a phylogenetic adaptation of the multiobjective variable mesh optimization algorithm [9] called *PhyloMOVMO*, to infer phylogenetic trees optimizing two optimality criteria, simultaneously: the Maximum Parsimony and Maximum Likelihood, with the aim of allowing biologists to infer in a single run a set of trade-off phylogenetic topologies that represent a consensus between different points of both optimality criteria. In order to assess the performance of our proposal, we have carried out experiments on four nucleotide data sets extracted from the state-of-the-art, comparing the multiobjective and biological results with other popular and recient multiobjective metaheuritics applying multiobjective quality metrics. PhyloMOVMO has been implemented using funcionalities of the framework MO-Phylogenetics [10], a phylogenetic inference software tool with multi-objective evolutionary metaheuristics. The rest of the algorithms, the benchmark, the configurations and parameters files were taken from this framework.

The remainder of this paper is organized in the following way. In the Section II, we introduce concepts about the basis of phylogenetics, the complexity of the problem and the parsimony and likelihood methods. Section III explains the details about the PhyloMOVMO algorithm and the adaptation to phylogenetic inference. The followed experimental methodology to assess the performance of our proposal is described in the Section IV. The multiobjective and biological results are shown in Section V. And finally, Section VI summarizes some conclusions and future works about this topic.

II. Phylogenetic Inference fundamentals

Phylogenetic inference seeks to find the most accurate hypotheses about the evolution of species by combining statistical techniques and algorithmic procedures. In a phylogenetic analysis, we consider as input an alignment composed by n sequences of N characters (sites) that represent molecular characteristics of the organisms under review. Site values in the sequences belong to an alphabet Σ defined in accordance with the nature of the data, where for DNA sequences, Σ consists of four characters of the nucleotides {A, T, G, C} and for protein sequences, Σ consists of 20 characters of the amino acids {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}. The output of the inference process is a tree-shaped structure $\tau = (V, E)$, where V represents the set of nodes in the tree τ and E the branches that connect related nodes V in the tree τ.

A. Complexity of the problem

The main computational problem of phylogenetic inference is the large number of possible topologies in the search space, which grows exponentially with the number of species to be analyzed.

Given n organisms, the number of possible binary unrooted trees is defined by equation 1 [11]:

$$|SS| = \prod_{i=1}^{n}(2_i - 5) = \frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{1}$$

Due to the large number of possible combinations, the exhaustive methods become totally complex from a computational approach, when trying to infer phylogenies with more of ten species. Because of this "combinatorial explosion", the phylogenetic inference is considered as NP-Hard problem, formally demonstrated both under an approach Maximum Parsimony [12] and Maximum Likelihood [13].

In the following subsections we will introduce the basis of two of the most used criteria-based methods for phylogenetic reconstruction: maximum parsimony and maximum likelihood analysis.

B. Maximum Parsimony Approach

Among the different hypotheses that explain the nature of a system, Occam's reasoning suggests that the simplest hypothesis relative to a phenomenon must always be preferred. This statement is widely applied in a wide range of scientific domains, including Bioinformatics. The principle of parsimony is an analysis inspired by this reasoning.

The maximum parsimony method aims to find a tree that minimizes the number of character state changes (or evolutionary steps) that are needed to explain the data. It is preferred the tree whose topology implies a smaller amount of transformations at molecular level [14]. The problem of maximum parsimony is described as follows: Let D an input dataset containing n number of aligned sequences of species. Each aligned sequence has N sites (columns of characters), where d_{ij} is the state character of the sequence i at the site j. Given the τ tree with the set of nodes $V(\tau)$ and the set of branches $E(\tau)$, the parsimony value of the tree τ is defined as equation 2 [15].

$$PS(\tau) = \sum_{j=1}^{N} \sum_{(v,u)\in E(\tau)} w_j C(v_j, u_j) \tag{2}$$

where w_j refers to the weight of the site j, v_j and u_j are the character states of the nodes v and u in the site j for each branch (u, v) in τ, respectively, and C is the cost matrix, such that $C(v_j, u_j)$ is the cost to change the state v_j to state u_j.

In this work, we will use the algorithm proposed by Fitch [16] to compute the parsimony score of a phylogenetic tree.

Having defined the algorithm that minimizes $PS(\tau)$ for a tree τ, we have to find the tree τ^* such that $PS(\tau^*)$ is the score with the lowest value of parsimony in the whole space of trees.

C. Maximum Likelihood Approach

Likelihood is a statistical function that, applied to phylogenetics, indicates the probability that the evolutionary hypothesis involving a phylogenetic tree topology and a molecular evolution model Φ would give rise to the set of organisms observed in the input data D (set of aligned sequences) [15]. The maximum likelihood approach aims to find that tree representing the more likely evolutionary history of the organisms of the input data. It can be defined as follows: The likelihood of a phylogenetic tree, denoted by $L = P(D|\tau, \Phi)$, is the conditional probability of the data D given a tree τ and an evolutionary model Φ [14].

Given τ, $L = (\tau)$ can be defined as equation 3:

$$L(\tau) = \prod_{j=1}^{N} L_j(\tau) \tag{3}$$

where $L_j(\tau) = P(D_j|\tau, \Phi)$ is the likelihood in the site j, which is denoted as equation 4:

$$L_j(\tau) = \sum_{r_j} C_j(r_j, r).\pi_{r_j} \tag{4}$$

where r is the root node of τ, r_j refers to any possible state of r in the site j, π_{r_j} is the frequency of the state r_j and $C_j(r_j, r)$ is the conditional likelihood of the sub-tree rooted by r. Specifically, $C_j(r_j, r)$ is the probability of everything that is observed from the root node r to the leaves of the tree τ, in the site j and given r, has state r_j. Let u and v the descendant nodes next to r, $C_j(r_j, r)$ can be formulated as equation 5:

$$C_j(r_j, r) = \left[\sum_{u_j} C_j(u_j, u).P(r_j, u_j, t_{ru})\right]\left[\sum_{v_j} C_j(v_j, v).P(r_j, v_j, t_{rv})\right] \tag{5}$$

where u_j y v_j refers to any state of the nodes u y v, respectively. t_{ru} and t_{rv} are the branch lengths that join the node r with the nodes v and u, respectively. $P(r_j, u_j, t_{ru})$ is the probability of change from the state r_j to the state u_j while the evolutionary time t_{ru}. Similarly, $P(r_j, v_j, t_{rv})$ is the

probability of change from the state r_j to the state v_j in the time t_{rv}. Both probabilities are provided by the evolutionary model Φ.

In this work, to calculate L we will use the method proposed by Felsenstein [14], where L is obtained by a post-order traversal in τ. Usually, it is convenient to use logarithmic values of L, so that the equation (3) can be redefined as equation 6:

$$\ln L(\tau) = \prod_{j=1}^{N} \ln L_j(\tau) \tag{6}$$

III. A MULTIOBJECTIVE VARIABLE MESH OPTIMIZATION APPROACH FOR PHYLOGENETIC INFERENCE

In this section we describe the main features of our proposal, a phylogenetic adaptation of the Multiobjective variable mesh optimization algorithm (MOVMO) proposed by [9]. Algorithm 1 shows the PhyloMOVMO's general workflow. The parameters: Mesh size P, Neighborhood size k, Number maximun of evaluations C, Maximun archive size S are the same of the MOVMO algorithm. We have included the input dataset to infer phylogenies: the multiple sequence alignments with the set of aligned sequences, the initial phylogenetic trees, and the evolutionary model parameters for each dataset, which can be computed by using jModelTest [17]. The representation of the individuals is based on the standard tree template codification. The crossover operator is the Prune-Delete-Graft (PDG) recombination method [18]. The output of the algorithm will be a set of non-dominated solutions L (Pareto set approximation) that describes trade-off phylogenetic topologies.

The algorithm starts by generating the initial mesh Pop_0 and initializing the leaders archive L (using Algorithm 2) with all the non-dominated solutions in Pop_0 (Lines 1 and 2). These initial solutions are assigned randomly from a repository composed by phylogenies generated by a bootstrap analysis [14]. For each node n_i of the current mesh Pop, the following steps are carried out:

1) The best node n_i^* among the n_i's k nearest neighbors in the decision variable space is selected. The distance between the nodes (phylogenetic trees) is calculated according to the Robinson-Foulds metric and the best node is selected according to the multiobjective dominance criterion (Line 5).
2) If the local optimum dominates n_i, a new node n_l is generated by applying TreeCrossover operator using n_i^* and n_i (Line 7); otherwise n_i is the local optimum itself (Line 8).
3) A global leader n_g from the archive L is selected through Binary Tournament selection operator (Line 10). Two non-dominated solutions from L are randomly picked and the one with largest crowding distance in L is selected.
4) A TreeCrossover operator is applied over the global leader n_g with the local optimum n_l (Line 11) to generate a new solution n_x, which contains subtrees of both topologies (mesh nodes).

Algorithm 1: Phylogenetic Multiobjective Variable Mesh Optimization (PhyloMOVMO)

Input: mesh size P, neighborhood size k, number max. of evaluations C, maximun archive size S
Data: multiple sequence alignment, initial trees and evolutionary model
Result: An approximation L of the true Pareto set L^*

1 $Pop \leftarrow$ Initialize_Evaluate_Population(P);
2 L=Initialize archive with each mesh node n_i by Algorithm 2;
3 $c \leftarrow 1$;
4 **while** *node* n_i *in the current mesh* Pop **do**
5 $n_i^* \leftarrow$ the best among the k neighbors of n_i
6 **if** $n_i^* \prec n_i$ **then**
7 $n_l \leftarrow$ PDGTreeCrossover(n_i^*,n_i);
8 **else**
9 $n_l \leftarrow n_i$
10 $n_g \leftarrow$ Select a global leader from L (BinaryTournament_Selection);
11 $n_x \leftarrow$ PDGTreeCrossover(n_l,n_g);
12 n_x:PhylogeneticOptimization(PPN&PLL);
13 evaluateFitness(n_x);
14 add n_x to the Pareto set approximation L (see Algorithm 2);
15 **if** $n_x \preceq n_i$ **then**
16 Replace n_i with n_x in the current population Pop
17 $c \leftarrow c + 1$;
18 return L

5) A phylogenetic optimization method is applied on n_x (Line 12), a Local Search provided by MO-Phylogenetics [10] based on two highly optimized techniques to explore the tree space, pllRearrangeSearch [19] and PPN [20], to optimize the likelihood and parsimony objectives, respectively.
6) The new n_x node is evaluated and, if is a new non-dominated solution, is added to the Pareto set approximation L. All dominated solutions by n_x are deleted in L (Line 13 and 14).
7) Finally, if n_i is dominated by n_x, it is replaced with n_x in the current mesh Pop (Line 16).

PhyloMOVMO returns the leaders archive L as the approximation of the Pareto optimal set found.

Algorithm 2 describes the addition of a new mesh node n_x to the bounded leader archive L. First, all nodes in L that are dominated by the incoming solution are deleted from the archive prior to n_x's addition. If the archive reached its maximum size, we drop the node with the lowest crowding distance. This ensures that a well-spread set of non-dominated solutions is maintained in L.

Evolutionary Crossover Operator

A wide range of recombination operators can be found in the literature [21], [22]. We have used in our proposal the Prune-Delete-Graft (PDG) recombination operator [18] available in

Algorithm 2: Add the n_x solution to the leader archive L)

Input: Solution n_x, archive L
Result: Archive L

```
foreach n_j of L do
    if n_x ≺ n_j then
        L ← L − n_j ;      /* remove n_j from the
         archive */
    else if n_x = n_j ∥ n_j ⪯ n_x then
        exit ;                    /* discard n_x */
L ← L ∪ n_x ;    /* add n_x to the archive */
if L : size() ≻ L : maxSize() then
    recompute crowding distances in L;
    L ← L− {L:worstByCrowdingDistance} ;
     /* remove most crowded solution */
```

MO-Phylogenetics. This operator takes a random subtree from one of the tree and inserts it in the other tree at a randomly selected insertion point, deleting duplicated species. Fig. 1 ilustrates the operator.

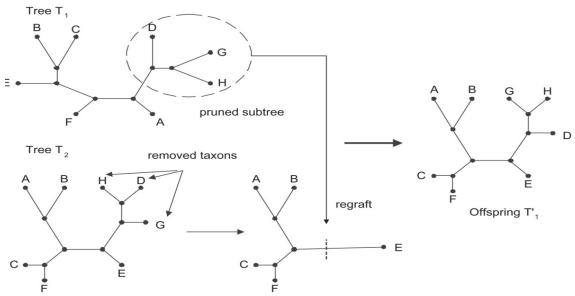


Fig. 1. Example of the Prune-Delete-Graft crossover operator.

IV. Experimental Methodology

In this section, we summarize the experimental methodology used to assess the performance achieved by our proposal PhyloMOVMO.

To comparate the results of our proposal, we have selected three representative multiobjective algorithms of the state-of-the-art, the classical reference NSGA-II and two other moderm techniques MOEA/D and SMS-EMOA, which are representative techniques of decomposition and indicator-based algorithms, respectively.

- NSGA-II [23] is a generational genetic algorithm based on generating new individuals from the original population by applying the typical genetic operators (selection, crossover and mutation). A ranking procedure is applied to promote convergence, while a density estimator (the crowding distance) is used to enhance the diversity of the set of found solutions.
- MOEA/D [24] is based on decomposing a multi-objective optimization problem into a number of scalar optimization subproblems, which are optimized simultaneously, only using information from their neighboring subproblems. This algorithm also applies a mutation operator to the solutions.
- SMS-EMOA [25] is a steady-state evolutionary algorithm that uses a selection operator based on the hyper-volume measure combined with the concept of non-dominated sorting.

We have used four multiobjective quality indicators: the Hypervolume (I_{HV}) and the Inverted Generational Distance Plus or IGD$^+$ (I_{IGD^+}) to take into account both the convergence and diversity of the Pareto front approximations, the Unary Additive Epsilon ($I_{\epsilon+}$) and the Spread or Δ (I_Δ) indicators, that are used as a complement to measure the degree of convergence and diversity, respectively. As we are dealing with real-world optimization problems, the Pareto fronts to calculate these two metrics are not known, so we have generated a reference Pareto front for each nucleotide dataset by combining all the non-dominated solutions computed in all the executions of all the algorithms. This strategy allows to make a relative performance assessment of the metaheuristics, because if the behavior of all the compared techniques is poor we know which of them yields the best fronts, but we do not know if they are near or far from the true Pareto front.

The experiments were carried out on four nucleotide data sets from the literature [26]: *rbcL_55* 55 sequences with 1314 nucleotides per sequence of the rbcL gene, *mtDNA_186* 186 sequences with 16608 nucleotides per sequence of human Mt DNA, *RDPII_218* 218 sequences with 4182 nucleotides per sequence of prokaryotic RNA and *ZILLA_500* 500 sequences with 759 nucleotides per sequence of rbcL plastid gene, under the reliable General Time Reversible (GTR+Γ) evolutionary model [27]. For each combination of algorithm and nucleotide dataset problem we have carried out 20 independent runs, and we report the median, $\tilde{x}$, and the interquartile range, IQR, as measures of location (or central tendency) and statistical dispersion, respectively, for every considered indicators. When presenting the obtained values in tables, we emphasize with a dark gray background the best result for each problem, and a clear grey background is used to indicate the second best result; this way, we can see at a glance the most salient algorithms. To check if differences in the results are statistically significant, we have applied the unpaired Wilcoxon rank-sum test. A confidence level of 95% (i.e., significance level of 5% or p-value under 0.05) has been used in all cases. The results of these tests have been summarized in tables where each cell contains the results of this test for a pair of algorithms. Three different symbols are used: "–" indicates that there is no statistical significance between these algorithms, "▲" means that the algorithm in the row has yielded better results than the algorithm in the column with statistical confidence, and "▽" is used when the algorithm in the column is statistically better than the algorithm in the row.

All the algorithms use the same parameters, the crossover and mutation probabilities are 0.8 and 0.2. The population size is 100. The initial population is generated by using a set of user phylogenetic trees performed by bootstrap analysis [14]. The parameters of the evolutionary model are computed by jModelTest [17].

TABLE I
MEDIAN AND INTERQUARTILE RANGE IQR OF THE VALUES OF THE $I_{\epsilon+}$ INDICATOR.

	PhyloMOVMO	NSGAII	MOEAD	SMSEMOA
rbcL_55	$2.10e-01_{1.4e-01}$	$1.01e+00_{5.0e-01}$	$1.75e-01_{4.1e-02}$	$2.50e-01_{8.5e-02}$
mtDNA_186	$3.33e-01_{1.5e-01}$	$3.38e-01_{1.1e-01}$	$3.89e-01_{1.7e-01}$	$4.44e-01_{2.1e-01}$
RDPII_218	$1.18e-01_{3.6e-02}$	$1.36e-01_{2.7e-02}$	$1.59e-01_{5.6e-02}$	$1.51e-01_{4.7e-02}$
ZILLA_500	$7.68e-01_{3.8e-01}$	$9.33e-01_{2.0e-01}$	$8.13e-01_{3.7e-01}$	$9.38e-01_{3.0e-01}$

TABLE II
MEDIAN AND INTERQUARTILE RANGE IQR OF THE VALUES OF THE I_{Δ} INDICATOR.

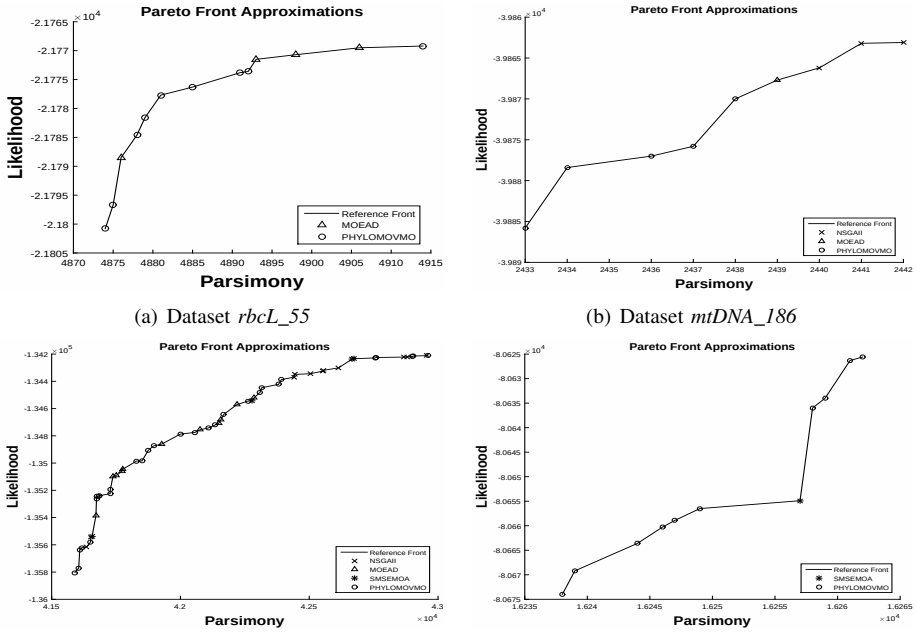
	PhyloMOVMO	NSGAII	MOEAD	SMSEMOA
rbcL_55	$9.91e-01_{2.5e-01}$	$1.01e+00_{3.4e-01}$	$1.14e+00_{2.8e-01}$	$8.76e-01_{3.8e-01}$
mtDNA_186	$1.31e+00_{3.9e-01}$	$7.97e-01_{5.0e-01}$	$1.30e+00_{1.5e-01}$	$1.13e+00_{7.7e-01}$
RDPII_218	$8.89e-01_{1.8e-01}$	$7.99e-01_{6.8e-02}$	$1.13e+00_{9.4e-02}$	$9.11e-01_{1.8e-01}$
ZILLA_500	$1.09e+00_{1.6e-01}$	$8.44e-01_{1.1e-01}$	$1.16e+00_{8.4e-02}$	$9.74e-01_{2.4e-01}$

TABLE III
MEDIAN AND INTERQUARTILE RANGE IQR OF THE VALUES OF THE I_{HV} INDICATOR.

	PhyloMOVMO	NSGAII	MOEAD	SMSEMOA
rbcL_55	$6.34e-01_{1.7e-01}$	$0.00e+00_{0.0e+00}$	$6.83e-01_{5.5e-02}$	$5.81e-01_{1.3e-01}$
mtDNA_186	$3.07e-01_{1.3e-01}$	$2.56e-01_{1.1e-01}$	$2.75e-01_{1.6e-01}$	$2.40e-01_{1.6e-01}$
RDPII_218	$6.18e-01_{4.2e-02}$	$6.08e-01_{5.5e-02}$	$5.87e-01_{8.9e-02}$	$5.99e-01_{4.1e-02}$
ZILLA_500	$1.57e-02_{1.0e-01}$	$0.00e+00_{1.1e-02}$	$2.88e-03_{8.7e-02}$	$0.00e+00_{3.1e-02}$

TABLE IV
MEDIAN AND INTERQUARTILE RANGE IQR OF THE VALUES OF THE I_{IGD^+} INDICATOR.

	PhyloMOVMO	NSGAII	MOEAD	SMSEMOA
rbcL_55	$1.10e-01_{1.2e-01}$	$9.35e-01_{5.2e-01}$	$8.71e-02_{4.0e-02}$	$1.48e-01_{7.7e-02}$
mtDNA_186	$1.90e-01_{1.2e-01}$	$2.33e-01_{8.9e-02}$	$2.26e-01_{2.0e-01}$	$2.37e-01_{2.2e-01}$
RDPII_218	$6.87e-02_{2.5e-02}$	$7.55e-02_{3.5e-02}$	$8.57e-02_{5.0e-02}$	$7.77e-02_{3.2e-02}$
ZILLA_500	$5.56e-01_{6.3e-01}$	$8.25e-01_{3.2e-01}$	$6.97e-01_{4.7e-01}$	$5.99e-01_{2.0e-01}$



(a) Dataset *rbcL_55* (b) Dataset *mtDNA_186*

(c) Dataset *RDPII_218* (d) Dataset *ZILLA_500*

Fig. 2. Reference Pareto fronts and best Pareto front approximations obtained by all the algorithms (PhyloMOVMO, NSGAII, MOEA/D and SMSEMOA) over 20 independent runs solving the nucleotide datasets a) *rbcL_55*, b) *mtDNA_186*, c) *RDPII_218* and d) *ZILLA_500*.

V. Empirical Results and Statistical Analysis

In this section we analyze the PhyloMOVMO's multiobjective and biological performance compared to NSGA-II, MOEA/D and SMSEMOA solving four nucleotide datasets from benchmark based on the experimental methodology described in Section IV.

Multiobjective results

The median values, $\tilde{x}$, and the interquartile range, IQR of the quality indicators $I_{\epsilon+}$, I_{Δ}, I_{HV} and I_{IGD^+} are reported in the Tables I, II, III and IV, respectively. We have to consider the highest values of I_{HV} and I_{IGD^+} and the lowest of $I_{\epsilon+}$ and I_{Δ}

The results of $I_{\epsilon+}$, I_{HV} and I_{IGD^+} show that PhyloMOVMO obtains the best median values for all the datasets, except for the *rcbL_55* instance, where MOEA/D shows a better performance. And for the results of I_{Δ} occurs the same, NSGAII obtains the best median values for all the datasets, except for the *rcbL_55* instance, where SMSEMOA shows a better performance.

Pareto Front approximations

To ilustrate graphically the multiobjective quality indicators results, we ilustrate in the Figure 2 the reference Pareto front (described in Section IV), and the best Pareto front approximations obtained by all the algorithms (PhyloMOVMO, NSGAII, MOEA/D and SMSEMOA) over 20 independent runs solving the nucleotide datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

We can observe that all reference Pareto fronts of the Figure 2, are mostly conformed by the non-dominated solutions (phylogenies) of the Pareto front approximations of PhyloMOVMO, considering only a few solutions of MOEA/D and SMSEMOA for the *rbcL_55* and *ZILLA_500* datasets,respectively. Furthemore, in the Figure 2c we can observe a high competitive performance that exists between all the algorithms solving the *RDPII_218* nucleotide dataset.

Furthermore, the Figure 3 shows the reference Front and the Pareto front approximations of each algorithm of each nucleotide dataset, from the best values of I_{HV} and I_{IGD^+} indicators, respectively, considering that both take into account the convergence and diversity of the Pareto front approximations. All these Pareto fronts approximations confirm the multiobjective quality indicators results of the Tables I, II, III and IV.

Statistical Analysis

The Tables V, VI, VII and VIII show the the Wilcoxon rank-sum test results. These results confirm that PhyloMOVMO yielded better performance at 95% significance level on the $I_{\epsilon+}$, I_{HV} and I_{IGD^+} values for the datasets *mtDNA_186*, *RDPII_218* and *ZILLA_500*, except for the *rbcL_55* instance where MOEA/D reports a better performance overall the algorithms. Furthermore, these results confirm the best performance of NSGAII for the I_{Δ} values, and although the SMSEMOA reports the best performance over the *rcbL_55* instance in this indicator, the Wilcoxon test indicates that they are not statistically significant with the rest of the algorithms except for MOEA/D.

TABLE V
Results of the Wilcoxon rank-sum test for the $I_{\epsilon+}$ values for the datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

	NSGAII				MOEAD				SMSEMOA			
PhyloMOVMO	▲	–	–	▲	▽	▲	▲	–	▲	▲	▲	▲
NSGAII					▽	▲	▲	▽	▽	▲	▲	–
MOEAD									▲	–	–	▲

TABLE VI
Results of the Wilcoxon rank-sum test for the I_{Δ} values for the datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

	NSGAII				MOEAD				SMSEMOA			
PhyloMOVMO	–	▽	▽	▽	–	–	▲	▲	–	–	–	–
NSGAII					–	▲	▲	▲	–	–	–	▲
MOEAD									▽	–	▽	▽

TABLE VII
Results of the Wilcoxon rank-sum test for the I_{HV} values for the datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

	NSGAII				MOEAD				SMSEMOA			
PhyloMOVMO	▲	▲	–	▲	▽	–	▲	▲	▲	▲	▲	▲
NSGAII					▽	–	▲	▽	▽	–	▲	–
MOEAD									▲	–	–	▲

TABLE VIII
Results of the Wilcoxon rank-sum test for the I_{IGD^+} values for the datasets *rbcL_55*, *mtDNA_186*, *RDPII_218* and *ZILLA_500*.

	NSGAII				MOEAD				SMSEMOA			
PhyloMOVMO	▲	▲	–	▲	–	▲	▲	▲	▲	▲	▲	–
NSGAII					▽	▽	▲	–	▽	–	▲	–
MOEAD									▲	▲	–	–

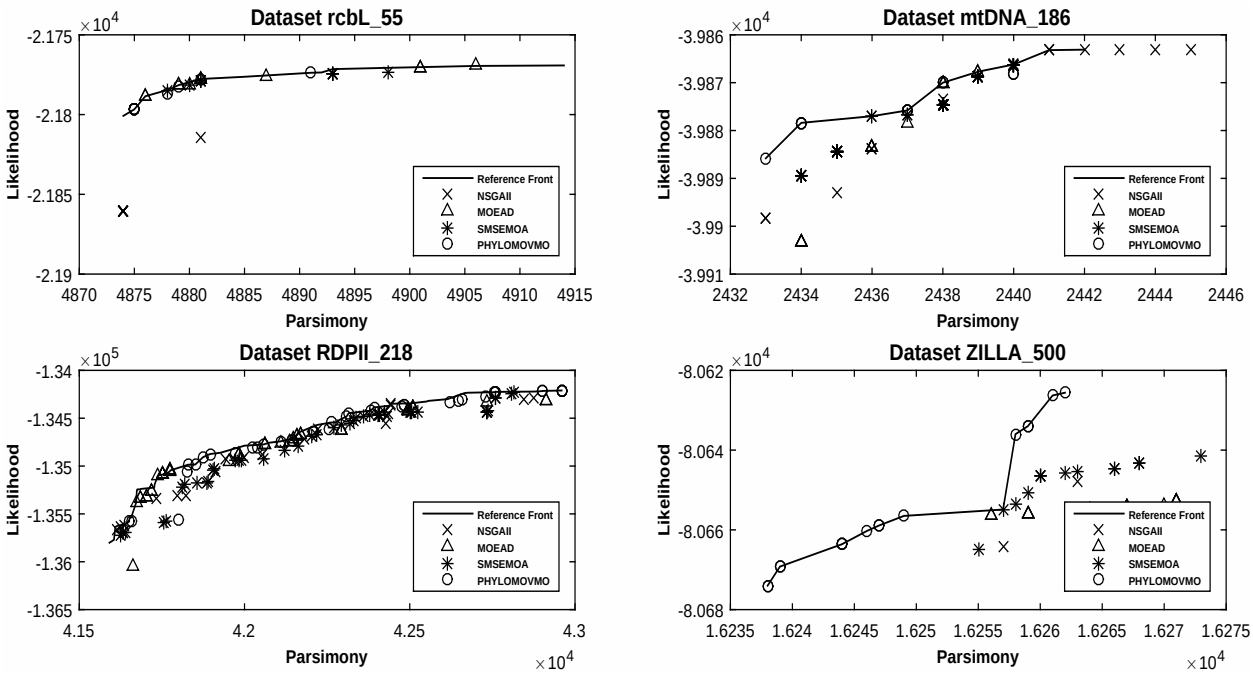
Biological results

The Table IX shows the best maximum parsimony and maximum likelihood scores obtanied by all the algorithms solving the four nucleotide datasets.

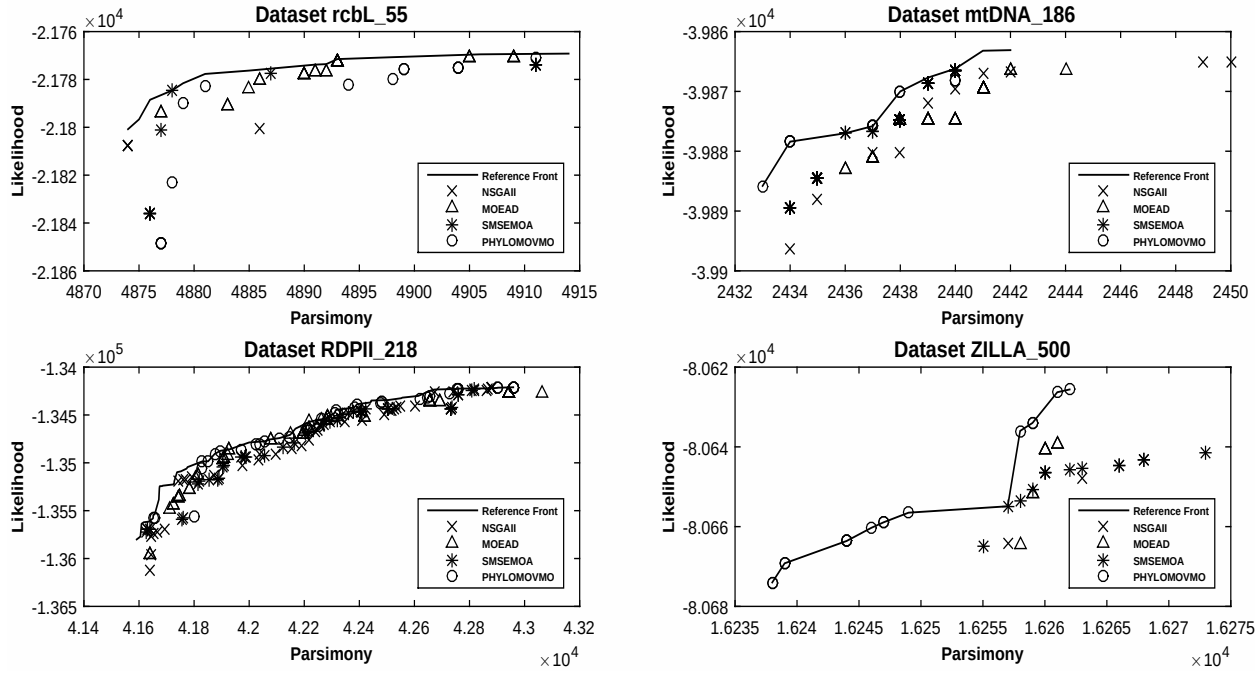
TABLE IX
Phylogenetic results. Comparing Parsimony and Likelihood scores of PhyloMOVMO with other multiobjective metaheuristics.

Dataset	PhyloMOVMO		NSGAII		MOEAD		SMSEMOA	
	Par.	Lik.	Par.	Lik.	Par.	Lik.	Par.	Lik.
rbcL_55	**4874**	**-21769.22**	4874	-21800.81	4874	-21769.53	4874	-21773.63
mtDNA_186	**2133**	-39865.52	2433	**-39863.11**	2434	-39866.60	2434	-39864.73
RDPII_218	**41589**	**-134210.40**	41613	-134211.03	41634	-134238.30	41618	-134224.12
ZILLA_500	**16238**	**-80625.60**	16251	-80630.91	16251	-80639.42	16250	-80628.03

We can observe that our proposal achieves a significant improvement with regard to the parsimony and likelihood scores reported by the other algorithms, except for the *rbcL_55* dataset where all the algorithms generate the same parsimony scores and for the *mtDNA_186* dataset where NSGAII performs a better likelihood score overall the algorithms.



(a) I_{HV} Pareto front approximations



(b) I_{IGD+} Pareto front approximations

Fig. 3. Pareto front approximations from the best I_{HV} and I_{IGD+} values obtained by all the algorithms (PhyloMOVMO, NSGAII, MOEA/D and SMSEMOA) over 20 independent runs resolving each nucleotide dataset.

Run-time analysis

Table X shows the computational processing times (in seconds) required to perform a complete execution of each algorithm (PhyloMOVMO, NSGAII, MOAED and SMSEMOA) using a single thread, making a phylogenetic analysis on each considered nucleotide dataset (rbcL_55, mtDNA_186, RDPII_218 and ZILLA_500).

TABLE X
SEQUENTIAL PROCESSING TIMES (SECS).

Dataset	PhyloMOVMO	NSGAII	MOAED	SMSEMOA
rbcL_55	5230.02	6376.34	22039.08	7500.19
mtDNA_186	39987.29	41870.29	47730.54	32362.21
RDPII_218	82901.13	79871.19	96085.18	84399.27
ZILLA_500	84090.31	82981.27	99098.10	86780.32

We can observe that the run-time of the algorithms is very expensive, specially for the large-scale datasets, requiring hours for mtDNA_186 dataset and almost a whole day for the RDPII_218 and ZILLA_500 datasets. The single-thread version of PhyloMOVMO and NSGAII perform a faster execution than the other algorithms.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, we have presented the PhyloMOVMO algorithm, a novel approach based on a multiobjective variable mesh optimization technique for inferring phylogenies, optimizing both parsimony and likelihood criteria simultaneously. Unlike to other multiobjective proposals applied to phylogenetic inference, the solutions selection based on the Robinson-Foulds distance metric, adds a new perspective to the exploration of the tree-space.

With the aim of evaluating its multiobjective and biological performance, we have carried out experiments on four nucleotide datasets and applying multiobjective quality indicators with other classical and recent multiobjective metaheuristics. With the purpose of making a fair comparison, all the algorithms were configured using the same parameters.

The obtained results reveal that in the context of the adopted parameter settings, the experimentation methodology, and the solved datasets, PhyloMOVMO shows a very competitive performance, under both multiobjective and biological approaches. The reference Pareto fronts of each dataset, are almost totally composed by the non-dominated solutions generated by PhyloMOVMO. Furthermore, the values of the multiobjetive quality indicators shows a promising perfomance of our proposal and to confirm these results, a Wilcoxon rank-sum analysis indicates the significant statistically differences of our proposal. Finally, under biological approach, PhyloMOVMO obtanied the best parsimony and likelihood scores overall data sets, except for the mtDNA_186 dataset where NSGAII provided a better likelihood score.

In summary, preliminary results have shown that PhyloMOVMO can make relevant contributions to phylogenetic inference. Moreover, there are several aspects that can be investigated to improve the current approach, such as: a parameter sensitivity study (including the use of different phylogenetic optimization methods), improve the funcionality of the recombination operator, add new evolutionary models to support protein data sets and, PhyloMOVMO requires several hours to find acceptable Pareto-solutions if initial trees are poorly estimated, so performance can be improved using the benefits of shared-memory and distributed-memory programming paradigms to efficiently inferring phylognies of large-size sequences data sets with a lot of number of species.

REFERENCES

[1] A. Stamatakis, "Distributed and Parallel Algorithms and Systems for Inference of Huge Phylogenetic Trees based on the Distributed and Parallel Algorithms and Systems for Inference of Huge Phylogenetic Trees based on the Maximum Likelihood Method," Ph.D. dissertation, Technische Universität München, Germany, 10 2004.

[2] C. Zambrano-Vega, A. J. Nebro, J. J. Durillo, J. García-Nieto, and J. Aldana-Montes, "Multiple sequence alignment with multiobjective metaheuristics. a comparative study," *International Journal of Intelligent Systems*, vol. 32, no. 8, pp. 843–861, 2017. [Online]. Available: http://dx.doi.org/10.1002/int.21892

[3] C. Zambrano-Vega, A. J. Nebro, J. García-Nieto, and J. Aldana-Montes, "Comparing multi-objective metaheuristics for solving a three-objective formulation of multiple sequence alignment," *Progress in Artificial Intelligence*, pp. 1–16, 2017. [Online]. Available: http://dx.doi.org/10.1007/s13748-017-0116-6

[4] S. Santander-Jiménez and M. A. Vega-Rodríguez, "Applying a multiobjective metaheuristic inspired by honey bees to phylogenetic inference," *Biosystems*, vol. 114, no. 1, pp. 39–55, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0303264713001615

[5] J. Handl, D. B. Kell, and J. Knowles, "Multiobjective optimization in bioinformatics and computational biology." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 4, no. 2, pp. 279–92, 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17473320

[6] S. Santander-Jiménez and M. A. Vega-Rodríguez, "A multiobjective proposal based on the firefly algorithm for inferring phylogenies," in *European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Springer, 2013, pp. 141–152.

[7] W. Cancino and A. C. Delbem, "A Multi-objective Evolutionary Approach for Phylogenetic Inference," in *Evolutionary Multi-Criterion Optimization*, ser. Lecture Notes in Computer Science, S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, and T. Murata, Eds. Springer Berlin Heidelberg, 2007, vol. 4403, pp. 428–442. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-70928-2_34

[8] S. Santander-Jiménez and M. A. Vega-Rodríguez, "A hybrid approach to parallelize a fast non-dominated sorting genetic algorithm for phylogenetic inference," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 3, pp. 702–734, 2014. [Online]. Available: http://doi.wiley.com/10.1002/cpe.3269

[9] Y. Salgueiro, J. L. Toro, R. Bello, and R. Falcon, "Multiobjective variable mesh optimization," *Annals of Operations Research*, pp. 1–25, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10479-016-2221-5

[10] C. Zambrano-Vega, A. J. Nebro, and J. Aldana-Montes, "Mophylogenetics: a phylogenetic inference software tool with multi-objective evolutionary metaheuristics," *Methods in Ecology and Evolution*, vol. 7, no. 7, pp. 800–805, 2016. [Online]. Available: http://dx.doi.org/10.1111/2041-210X.12529

[11] A. Edwards, L. Cavalli-Sforza, V. Heywood *et al.*, "Phenetic and phylogenetic classification," *Systematic Association Publication No. 6*, pp. 67–76, 1964.

[12] W. H. Day, D. S. Johnson, and D. Sankoff, "The computational complexity of inferring rooted phylogenies by parsimony," *Mathematical biosciences*, vol. 81, no. 1, pp. 33–42, 1986.

[13] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees: hardness and approximation," *Bioinformatics*, vol. 21, no. suppl 1, pp. i97–i106, 2005.

[14] J. Felsenstein, *Inferring Phylogenies*. Palgrave Macmillan, 2004. [Online]. Available: http://books.google.fr/books?id=GI6PQgAACAAJ

[15] D. Swofford, G. Olsen, P. Waddell, and D. Hillis, "Phylogeny reconstruction," in *Molecular Systematics*, 3rd ed. Sinauer, 1996, ch. 11, pp. 407–514.

[16] W. M. Fitch, "Toward defining the course of evolution: Minimum change for a specific tree topology," *Systematic Biology*, vol. 20, no. 4, pp. 406–416, 1971. [Online]. Available: http://sysbio.oxfordjournals.org/content/20/4/406.abstract

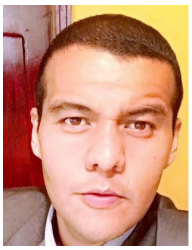
[17] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada, "jmodeltest 2: more models, new heuristics and parallel computing," *Nature methods*, vol. 9, no. 8, pp. 772–772, 2012.
[18] C. Cotta and P. Moscato, "Inferring phylogenetic trees using evolutionary algorithms," in *International Conference on Parallel Problem Solving from Nature*. Springer, 2002, pp. 720–729.
[19] T. Flouri, F. Izquierdo-Carrasco, D. Darriba, A. Aberer, L.-T. Nguyen, B. Minh, A. Von Haeseler, and A. Stamatakis, "The Phylogenetic Likelihood Library," *Systematic Biology*, vol. 64, no. 2, pp. 356–362, 2015.
[20] A. Goëffon, J.-M. Richer, and J.-K. Hao, "Progressive tree neighborhood applied to the maximum parsimony problem." *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 5, no. 1, pp. 136–45, 2008. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18245882
[21] H. Matsuda, "Construction of Phylogenetic Trees from Amino Acid Sequences using a Genetic Algorithm," *Sciences, Computer*, p. 560, 1995.
[22] C. B. Congdon, "Gaphyl: An Evolutionary Algorithms Approach For The Study Of Natural Evolution," in *GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference*, 2002, pp. 1057–1064.
[23] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.
[24] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.
[25] N. Beume, B. Naujoks, and M. Emmerich, "Sms-emoa: Multiobjective selection based on dominated hypervolume," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1653–1669, 2007.
[26] W. Cancino and A. Delbem, "A multi-criterion evolutionary approach applied to phylogenetic reconstruction," in *New Achievements in Evolutionary Computation*, P. Korosec, Ed. Rijeka: InTech, 2010, ch. 06. [Online]. Available: http://dx.doi.org/10.5772/8051
[27] C. Lanave, G. Preparata, C. Sacone, and G. Serio, "A new method for calculating evolutionary substitution rates," *Journal of molecular evolution*, vol. 20, no. 1, pp. 86–93, 1984.

Cristian Zambrano-Vega Docente investigador de la Carrera de Ingeniería en Sistemas de la Facultad de Ciencias de la Ingienería de la Universidad Técnica Estatal de Quevedo. Doctor en Ingeniería del Software e Inteligencia Artificial de la Universidad de Málaga. Su línea de investigación abarca las técnicas de optimización multiobjetivo aplicadas a la Inferencia Filogenética y al Alineamiento Múltiple de Secuencias. Email czambrano@uteq.edu.ec

Byron Oviedo Bayas Director del Departamento de Investigación y Docente titular principal de la Universidad Técnica Estatal de Quevedo. Doctor en el programa oficial de Tecnologías de la Información y Comunicación de la Universidad de Granada - España. Email: boviedo@uteq.edu.ec.

Stalin Carreño Sandoya Ingeniero en Sistemas y Master en Conectividad y Redes de Ordenadores. Docente de la Unidad de Estudios a Distancia. Líder de la Unidad de Tecnologías de la Información y Comunicación de la Universidad Técnica Estatal de Quevedo. Email: sdcarreno@uteq.edu.ec

Amilkar Puris Cáceres Ph.D. en Ciencias Técnicas por la Universidad Marta Abreu de las Villas, Cuba. Sus principales investigaciones han sido en el área de las Metaheurísticas Poblacionales para la solución de problemas complejos. Actualmente se desempeña como docente e investigador en la Universidad Técnica Estatal de Quevedo. Email: apuris@uteq.edu.ec.



Oscar Moncayo Carreño Docente titular de la Carrera de Ingeniería en Gestión Empresarial de la Facultad de Ciencias Empresariales de la Universidad Técnica Estatal de Quevedo. Magister en Dirección de Empresas con Énfasis en Gerencia Estratégica en la Universidad Regional Autónoma de los Andes. Email omoncayo@uteq.edu.ec

Children learning of programming: Learn-Play-Do approach

Julián-Andrés Galindo, and Monserrate Intriago-Pazmiño

***Abstract*—Writing computer programs is a skill that can be introduced to children and adolescents since early ages. Although children can gain skills in coding, there is a lack of motivation and easiness at the time to write logic structures. It raises the question, how can children be encouraged to code in a successful environment of learning and fun?. To address this question, this paper shows an experimental approach called "Learn-Play-Do" for introducing children in the programming. It shows that (1) it is feasible for children to learn about programming by following the proposed approach with (2) encouraging levels of learning, usefulness content and self-learning programming in (3) a developing country context. The results of an empirical experimentation with forty-one children are reported. This work was implemented as a social project linking the university with the community.**

***Index Terms*—Computers & programming, children learning, Scratch.**

I. Introduction

Writing computer programs is a skill who can be introduced since early ages [1], [2]. Papert argued the main learning benefit is called the "Piagetian learning," or commonly called "learning without being taught" [3]. It has been reported as an effective device for a cognitive process instruction focus on teaching *how* rather than *what*. Through that, children can model abstract concepts to help them to develop skills such as classification, meta-cognition, left and right orientation, verbal memory and creative thinking.

Many issues have been reported about children learning of programming, documented by the UK's Computing Research Committee [4]. These issues involve a lack of motivation, easiness, and formal teachers training to guide young learners with enthusiasm and pro-activity. As a result, programming is seen as a boring, difficult and frustrating activity. Therefore, it seems like children and teachers need an approach more interactive to overlap engagement, fun, and learning.

Furthermore, some interactive environments has been released such as LEGO WeDo [5], Raptor [6], Scratch [7], Tinker [8] and Turtle Math [9]. These tools allow children to access to a complete set of features to build programs in online and offline settings. These features mainly include audio and video, events, logic sequences, conditions, loops and images control. However, technology itself is still not enough. First, children have been reported with cognitive issues to learn programming such as divergent thinking, awareness of comprehension failure, reflectivity and impulsivity, operational competence and receptive vocabulary [1]. Second, science learning emerges other children issues which include (1) children conception of the world and their influence at science learning, (2) language disabilities, (3) the role of the science teacher, (4) analysis of a teaching model and (5) the implications in the curriculum and teacher education [10]. Hence, technological solutions promote the art of programming. It requires a global and transversal approach.

This complex vision may be addressed by the implementation of an experimental project [11]. We introduce its core component Learn-Play-Do and We shall show the results of the first round with University students (fulfilling the instructor role) and children attending primary school. The key findings of using Learn-Play-Do reveals that (1) children learn by following its two stages (play and do) with (2) a valuable degree in learning, content usefulness and self-learning in programming with (3) children in an Ecuadorian context.

The rest of this article is organized as follow. The second section presents the related work about programming interactive environments for children. The third section describes the Learn-Play-Do approach. In the fourth section, we describe the design of our experimental study. In the fifth section, we report our experiment's results that compile the first experiences with this approach. The sixth section contains some discussions and limitations of this study. Finally, some conclusions and future works are presented.

II. Related work

There are many approaches to teach children about programming. To begin, the book "Teach your kids to code" shows a traditional manner to learn where children are exposed to a console to write commands in Python and then check its output in a Graphical User Interface (GUI) (see Fig. 1). Although, this project restates the need of exploring coding in a fun environment with valuable principles such as "do it together", "Coding = Solving problems" and

Article history:
Received 15 September 2017
Accepted 28 November 2017

J.A. Galindo is a professor at the Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito, Ecuador. He is also a PhD candidate at the Laboratoire d'informatique de Grenoble, Grenoble Alps University (UGA), Grenoble, France (e-mail: julian.galindo@epn.edu.ec)

M. Intriago-Pazmiño is a professor at the Departamento de Informática y Ciencias de la Computación, Escuela Politécnica Nacional, Quito, Ecuador (e-mail: monserrate.intriago@epn.edu.ec)

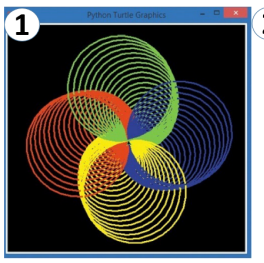


Fig. 1. Python project [2]. 1) program output, 2) source code.

Fig. 2. Turtle math project. a) turtle turner tool and b) the label lines tool.

”Explore!”, the method remains in text commands which is useless for children in level 0 (2-7 years) and level 1 (5-10 years) as reported in the project ”Should Your 8-year-old Learn Coding?” [2]. Thus, this traditional approach should be taught to children at level 3 (12 years and up) which rises a learning curve wall for others.

In the Turtle math project (see Fig. 2) building text and graphical relationships are exposed. It can be considered as a version of Logo for learning mainly mathematics [9]. This approach is based on six research principles which expose children (in upper elementary grades) and teachers in an interactive environment. It includes visual elements and text commands to control paths, shapes, scaling, coordinates, motions, drawing and number activities. The main principle related to programming is ”maintain close ties between representations”. This argues that explicit relationships between programming codes and drawings are essential. Children often lose these connections so that they need to write, save commands and see them run immediately. Although this approach underlines programming, fun seems like a fuzzy element in the interaction. The User Interface (UI) does not encourage children to play as coding because it has a lack of aesthetics and usability expressed mainly in the activity windows. For instance, children need to code by hand commands which may cause compiling errors (low usability) as well as a growth of negative user learning. A feature which can be faced by the introduction of interactive UI elements such as drag and drop widgets.

Another interesting approach is writing computer programs by using digital storytelling. It allows children to draw stories with a computer program where their imagination and composing skills are mixed with digital elements. In the 1990s, it became more popular including visual images and written text that expanded the student comprehension [12].

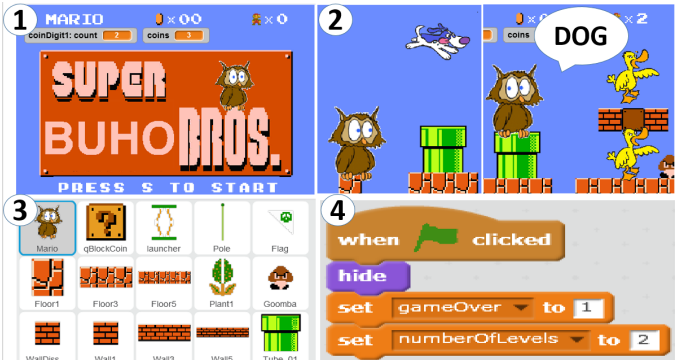


Fig. 3. Scratch project interface. 1) starting UI, 2) interactive UI, 3) visual elements and 4) script UI.

Then, Mitch Resnick in his TED conference emphasized the production of digital content by the expression “Learn to code and code to learn” where children develop writing and community learning skills by coding [13], [14].

Consequently, mixing writing skills and technology to combine audio, text, and video emerges a growing strategy to engage youth into computer programming. It was shown by Kelleher at designing a Storytelling program called Alice [15]. She argues that many girls begin to turn away from math and science-related disciplines (computer science) at the middle school. This programming environment provides to middle school girls a positive initial experience with computer programming as a means to the end of storytelling rather than an end in itself. A motivating activity for middle school girls at Pittsburgh.

Following this last approach, another robust research project was found named Scratch [7]. Scratch as a method to collect and test kids’ imagination allows them to create stories by a drag-and-drop block process. Kids stick the blocks together, forming code scripts as similar as developers create code lines in a web language such as python, C#, Java, and others. Then, when the code script is finished, kids can run it to bring to life the scratch characters of the screen. Using this tool, kids can create robust digital stories because every scratch block can represent text, audio or a video element to create an interactive story. For instance, Fig. 3 shows a Scratch project called “Super Buho Bros”. It is part of the ”Red Juega y Aprende” social project [16]. “Super Buho Bros” project aims to learn English vocabulary of animals as playing in a super Mario and fun environment. When a kid run the project, he will interact with the animations, read the English words and listen to the audios. Thus, Scratch projects represent a child ability to coordinate a different set of blocks to create projects as complex as they want to.

In spite of this, digital storytelling demands further examination in (1) the efficiency and clarity of the scripts produced by children, (2) the potential relation of programming and content, (3) analysis of imaginative and aesthetics features and (4) more broad studies to validate its

impact in children learning [17].

Overall, all approaches clarify the challenge to balance UI interaction, learnability, and playfulness. First, the traditional method (text-based) may be found difficult for children at early ages. Second, although there are advances in GUI, there is still a lack of aesthetics, usability in conjunction with enjoyability. Third, the storytelling approach by using visual elements helps children to learn about coding by mixing elements such as narration, creativity, and communication. Despite this, further examinations are mainly needed to validate the relation between digital content production and coding in children learning by these highlights.

III. Learn-Play-Do approach

From the related work, it is highlighted that children may learn to code by harmonizing a rich UI in a fun way. Since this lesson learned, our experimental approach relies on two simple stages: Play and Do. The first stage exposes children to play interactive games (made for tutors) which aims to introduce children to the environment by playing instead of coding. Then, with this gained knowledge, children have the opportunity to create their own programs with a tutor assistance at the Do stage which attempts to promote a cognitive [18], social [19] and emotional [20] child development. It is expected that as children gain knowledge - in a Play and Do spiral - the tutoring will be less required.

Now, to ensure the children learning process during these stages, the approach is also underlined by the Suzuki methodology [21] [22] [23]. This learning method can be summarized by: *results* = *desire* + *repeat* [24]. Suzuki argues that one learns only by continual practice of basic or main concepts. For instance, when children are taught mathematics in an exiting(fun) and interesting manner they develop a desire to repeat the learned activities [24]. Consequently, in our context, *results(Learn)* = *desire(Play)* + *repeat(Do)*. *Desire* will be attached to our Play stage which should encourage children to do or perform activities again (*repeat*). Therefore, desire (Play) + repeat (Do) should evoke results to keep children in a continuous learning growth.

To be consistent, the approach should cover also the following factors or principles of Suzuki's methodology:

- Listening
- Memory
- Motivation
- Vocabulary
- Repetition
- Parental Involvement
- Step by Step Mastery
- Love

Listening: By listening and watching to video recordings of the interactive lessons given by tutors, children should learn the coding language and UI interaction just as they absorb the sounds of their mother tongue to interact with others.

Memory: Through repetition and listening to the recordings, the child will code from memory which is a skill they can use with other educational aspects (reading and maths). Coding by text commands is postponed until the child is able to code by drag and drop elements, just as we only teach children to read after they can speak fluently. In this way, the tutors can concentrate of the child's coding development of main factors such as start and stop programs (events and sensing), adding pictures, videos, widgets (look and sound), repetitive and conditional actions (control) and widget actions (motion and operators).

Motivation: Daily practice of games is encouraged to build the child's abilities and confidence. As the child masters a particular game (program) the motivation and sense of achievement will move them on to the next game in a desire to learn more. All students follow the same games sequence so that an standard repertoire provides strong motivation as younger children want to code games they hear older students code. Parent involvement will motivate children and give them a sense of achievement and makes playing an enjoyable experience.

Vocabulary: At the beginning, children should regularly repeat all previously learned games and code exercises to expand his instructions knowledge(vocabulary) and o reuse it in future programs. Just as in learning to speak, the entire vocabulary and grammar are used, not just the most recently learned words. In this way, the child gradually expands his cognitive abilities to solve problems by reusing code.

Repetition: Through constant repetition of games, children strengthen old skills and gain new ones. The technique is developed through the study and repetition of these games. Students can interact with their games to see the progress they have made.

Parental Involvement: It is required a three-way partnership between the child, the tutor, and the parent by working together. Parents need to go to tutor lessons to serve as home teachers. With this, a more enjoyable environment is made where children can consolidate the teaching given by the tutor.

Step by Step Mastery: Every child learn by building small coding steps so that they need to start with easy games or programs to master coding gradually.

Love: Tutors and parents should have a strong level of empathy, patient, tolerance, and creativity to guide and reinforce children learning.

IV. Experimental study

The experimental study aims at exploring how feasible is children learning by using Learn-Play-Do approach. We will show that children learn programming following two stages (Play and Do) with an interactive tool "Scratch". The following section will underline the experimental protocol.

A. Goal and hypothesis

The objective of this experiment is to examine the degree of learning and likeness during the exposition of children with Learn-Play-Do as a first attempt to understand how to teach children about programming. So the experiment's hypothesis are:

- H1: children learn by their exposition with Learn-Play-Do approach.
- H2: children like the games' content.
- H3: children are able to code by themselves basic programs.

B. Experimental method

A quantitative research method was used [25], where children could interact with all games by following their preferences in a face-to-face tutor support. First, children interacted with the games only by playing. In Play stage, every two-children had a tutor who introduced them how to play by almost 20 minutes. Second, one tutor taught all the group how to create a single game by interacting with Scratch in a 20 minutes session. During Do stage, children are supported by their tutor to understand clearly the basic coding instructions by following the drag-and-drop interface from Scratch. Then, every child was challenged to make their own game by reusing the code of the previous stage. In this last part of the Do-stage, children had the opportunity to use their creativity and learned skills to develop a basic but fun game in a 20-minutes-period. Finally, a survey was provided by tutors to children to gather all their impressions about the programming experience.

C. Participants

For all sessions, 41 children participated in the experiment from the Dominicas de la Inmaculada Concepción primary school with 20 tutors (University students). This children group is distributed in 19 (under 6 years old), 1 (7 to 9), 18(10 to 11) and 3 from 12 to 14 years old. Furthermore, and 2 professors at Computer Science Faculty who guided whole experiment sessions. They supported and reinforced the learning process.

D. Materials

Five games were made for University students by using Scratch to expose children to interact with. Those ones include the body, the instruments, the numbers, the animals and the transport game that we will explain below.

1) Body Parts: The game has two options for body parts (see Fig. 4). It has its own audio and image. Once both are present, a question concerning launches into a body part which must be correctly selected. They must complete a total of five hits to win, otherwise, they lose. It reinforces the body parts in English.

Fig. 4. Body Part Game

Fig. 5. Instruments Game

2) Instruments: The game features two choices of musical instruments (see Fig. 5), which have their own audio and image, once both presented a question regarding a musical instrument which shall be selected appropriately launches. They must complete a total of five hits to win, otherwise, they lose. It aims to help children to recognize musical instruments in English.

3) Numbers: It allows children to do a review of the numbers 1 to 9 in English while showing us the correspondent quantity of different elements per number (see Fig. 6). For instance, at showing the number "3", three soccer balls are

Fig. 6. Numbers Review

Fig. 7. Animals Game

Fig. 8. Transport Game

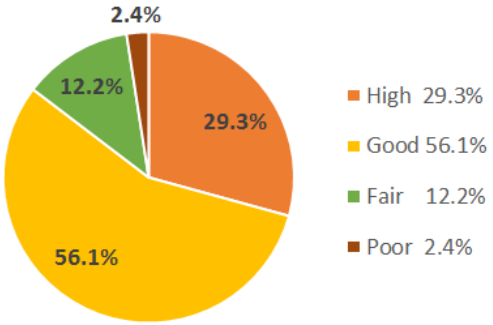


Fig. 9. About learning degree. Answers to the question: What grade of learning did you achieve through the game used?

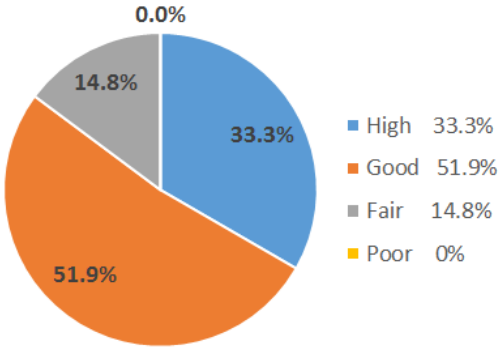


Fig. 10. About content usefulness. Answers to the question: How useful is the exposed content?

shown and the English word 'three' is playing. Hence, children can learn about the quantity and also the relation with sound and visual interaction.

4) The animals game: The game is designed to learn to differentiate vertebrates invertebrates, supporting English (see Fig. 7). We also have three ways to play. Vertebrate or invertebrate, here children need to locate the correct animal in the respective box classification, vertebrate or invertebrate. For the Roulette questions, it is necessary to spin the wheel with the space-bar, then we generated a question which has three options A, B and C, where one of these is the correct answer.To catch animals, an augmented reality game is shown for which it is necessary to use a webcam.

5) Transport: The game features two transportation options, each has their own audio and image (see Fig. 8). The main goal is a reinforcement of the means of transport in English. During the game, once both are present audio and image a question appears concerning a means of transport which shall be selected appropriately. Children must complete a total of five hits to win, otherwise, they lose.

6) Survey: A paper-based questionnaire was used to gather children impressions. It involved six questions regarding with the age, gender, learning degree, content usefulness, training likeness and coding independence.

V. Results

Two resulting products are reported. Fully developed games for children and their responses to the Learn-Play-Do approach. First, children games were published in [26]. Almost all games had a basic interaction as expected; however, a few children were able to modify the game "Super Mario Bros" by including new characters. Second, the following section will explore the key points regarding the gathered responses.

1) Learning degree: The results confirm that children learned coding with almost 56% and 29% for an acceptable and sufficient level (see Fig. 9). Reaching approximately 15% of low gained knowledge.

2) Content usefulness: Children liked the content of games (materials section) revealing almost 85% of a positive content reception and 15% as a negative one (see Fig. 10).

3) Coding independence: Mostly all children remained with confidence at programming with approximately 78% (see Fig. 11). In consequence, an group of 8 children did not learn enough to develop their own program.

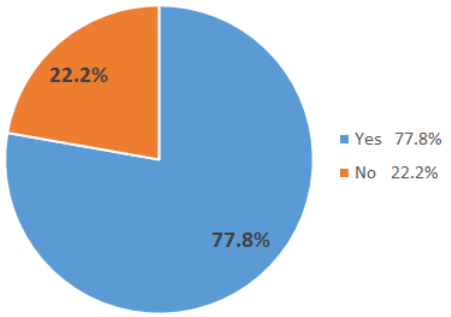


Fig. 11. About writing new code. Answers to the question: After this training, could you create your own code?

VI. Discussion

Programming can be introduced to children through many interactive platforms since plaint-text tools until rich user interfaces. However, this activity becomes harder at facing discouragement and good support at coding logic structures. Hence, it is valuable enough to explore an approach to encourage children at coding as keeping a learning and fun environment.

In this study, we have identified three main findings. First, most children reflected a positive learning degree with almost 85% after their exposition with "Learn-Play-Do" (H1). This fact is related to previous studies where children's performance reached a mean score of 64% programming by using Scratch [2]. Here, there was also a valuable factor to encourage learning beyond technology identified as affect. In fact, it was seen not only as an accompaniment but also as a source of motivation. As stated by Duncan "Learning will hardly progress without motivation, and that is stirred and maintained by positive affect". It supports the Learn-Play-Do approach where learning should be driven by a combination of fun (play) and repetitive activities (do). Thus, it suggests that learning may gain important levels in children when affect elements are shown such as motivation and fun aligned with technology.

Second, a low level of 15 percent was revealed in content likeness by children at interacting with the pack of 11 developed Scratch games (H2). Although it could be seen also a matter of novelty in children, there is positive evidence of longer children exposition with contents developed in Scratch shown in [2]. These contents included interaction with geometric shapes, sprites, scratch cards and audio files trough tasks such as order, selection, sequence, movement, coordination, and synchronization. Furthermore, it was argued that Scratch was beneficial and fun in an 8-weeks-period for children overpassing mere novelty. Therefore, a cyclical and longer exposition may be needed to confirm or validate the positive attraction of the content games.

Third, the majority of children showed a high level of coding independence with 78% (H3). Similarly, Burke and Kafai found that 9 out of 10 children knew more programming after their exposition with Scratch into a Storytelling process [17]. These technical skills included programming concepts such as object-oriented and sequential, sprite to sprite conditionals, looping, boolean variables, sampling scripts and if-then statements. Hence, children are able to increase their development skills trough Scratch; however, our study does not measure the performance of individual logic structures(e.g. if-then or do-while). Hence, it suggests a more deep analysis in the manner of children code which may confirm the gained levels in coding by logic structure toward a bottom-up approach.

1) Experiment limitations: Although these initial experimental results are encouraging, some restrictions should be noted. First, children should have more exposition to the approach which can clarify the impact of learning. Second, Learn-Play-Do approach used Scratch as an interactive tool for programming. However, there are other tools which could release other insights such as Lego Boost and Lego Mindstorms. Lastly, even when the initial games were quite easy to understand and explore their programming; it should be advisable to increase games difficulty according to children age and also the definition of a formal method of games assessment.

Overall, this experimental results reveals mainly that it is possible to use the Learn-Play-Do approach to achieve initial levels of (1) learning, (2) content usefulness, and (3) coding independence which need a deeper and longer exposition to validate and/or extend its degree of learning by children.

VII. Conclusions and Perspectives

This paper presents an approach denominated Learn-Play-Do for learning about code programming for children and an experiment to show its initial results. It validates that it is feasible for children to introduce them in learning programming skills by following the stages Play (fun) and Do (repetitive) with the interactive tool "Scratch" in the first round of the social project. Children revealed a valuable level of learning, content usefulness and coding independence. More experiments will be necessary to prove the knowledge reception in order to validate and/or extend the approach(stages and principles) with more children groups and experiment conditions. From the created games for children, more analysis in the script(code) could reveal the level of learning gained per child with more deep detail in logic sequences, variables and UI actions definition.

Acknowledgment

The authors wish to thank God, reviewers and our priceless family as well as University students for their support, highlights and quality time.

References

[1] D. H. Clements and D. F. Gullo, "Effects of computer programming on young children's cognition," *Journal of educational psychology*, vol. 76, no. 6, p. 1051, 1984.

[2] C. Duncan, T. Bell, and S. Tanimoto, "Should Your 8-year-old Learn Coding?" in *Proceedings of the 9th Workshop in Primary and Secondary Computing Education*, ser. WiPSCE'14. New York, NY, USA: ACM, 2014, pp. 60–69.
[3] J. Lochhead and J. Clement, *Cognitive Process Instruction. Research on Teaching Thinking Skills.* ERIC, 1979. [Online]. Available: https://eric.ed.gov/?id=ED234997
[4] UKCRC, "UK Computing Research Committee," 2010. [Online]. Available: http://www.ukcrc.org.uk/
[5] K. Mayerov?, "Pilot activities: LEGO WeDo at primary school," in *Proceedings of 3rd International Workshop Teaching Robotics, Teaching with Robotics: Integrating Robotics in School Curriculum*, 2012, pp. 32–39.
[6] M. C. Carlisle, T. A. Wilson, J. W. Humphries, and S. M. Hadfield, "RAPTOR: A Visual Programming Environment for Teaching Algorithmic Problem Solving," in *Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education*, ser. SIGCSE '05. New York, NY, USA: ACM, 2005, pp. 176–180.
[7] M. Resnick, J. Maloney, A. Monroy-Hern?ndez, N. Rusk, E. Eastmond, K. Brennan, A. Millner, E. Rosenbaum, J. Silver, B. Silverman, and others, "Scratch: programming for all," *Communications of the ACM*, vol. 52, no. 11, pp. 60–67, 2009.
[8] H. Lieberman, "Tinker: A programming by demonstration system for beginning programmers," *Watch what I do: programming by demonstration*, vol. 1, pp. 49–64, 1993.
[9] D. H. Clements and J. S. Meredith, "Turtle math," *Montreal: Logo Computer Systems (LCSI)*, 1994.
[10] R. Osborne and P. Freyberg, *The Implications of Children's Science*. Heinemann Educational Books, Inc, Jan. 1985.
[11] G. Julian and I. Monserrate, "Proyecto EPN-FIS de Vinculación Social, Programación para niños Red Juega y Aprende." 2015.
[12] G. Kress, "Visual and verbal modes of representation in electronically mediated," *Page to screen: Taking literacy into the electronic era*, p. 53, 1998.
[13] M. Resnick, "Learn to Code, Code to Learn," 2013. [Online]. Available: https://scratch.mit.edu/
[14] ——, "Scratch day," *EdSurge, May*, 2013.
[15] C. Kelleher, J. Hodgins, and S. Kiesler, "Motivating Programming: using storytelling to make computer programming attractive to more middle school girls," 2006.
[16] Red juega y aprende, "Interactive game, Super Buho Bros at the Scratch MIT network," 2015. [Online]. Available: https://scratch.mit.edu/projects/45950952/#editor
[17] Q. Burke and Y. B. Kafai, "Programming & storytelling: opportunities for learning about coding & composition," in *Proceedings of the 9th international conference on interaction design and children*. ACM, 2010, pp. 348–351. [Online]. Available: http://dl.acm.org/citation.cfm?id=1810611
[18] J. Piaget, "Part I: Cognitive development in children: Piaget development and learning," *Journal of research in science teaching*, vol. 2, no. 3, pp. 176–186, 1964.
[19] A. Kozulin, *Vygotsky's educational theory in cultural context*. Cambridge University Press, 2003.
[20] N. Chomsky, "A review of BF Skinner's Verbal Behavior," *Language*, vol. 35, no. 1, pp. 26–58, 1959.
[21] S. Suzuki and W. Suzuki, *Nurtured by love: The classic approach to talent education*. Alfred Music, 1983.
[22] E. Hermann, *Shinichi Suzuki: The Man and His Philosophy (Revised)*. Alfred Music, 1999.
[23] S. Suzuki and M. L. Nagata, *Ability development from age zero*. Alfred Music, 2014.
[24] D. G. Hazlewood, S. Stouffer, and M. Warshauer, "Suzuki meets P?lya: teaching mathematics to young pupils," *The Arithmetic Teacher*, vol. 37, no. 3, p. 8, 1989.
[25] N. Mandran, "Méthode de conduite de la recherche en informatique centrée humain : processus et inclusion d'une démarche centrée utilisateur," Ph.D. dissertation, Nov. 2015, working paper or preprint.
[26] Red juega y aprende, "Scratch - Imagine, Program, Share," 2015. [Online]. Available: https://scratch.mit.edu/studios/932042/

Julian-A. Galindo was born in Quito, Ecuador, in 1982. He received a bachelor degree in informatics engineering from Central University, UCE, Quito, Ecuador, in 2007. The Master in Information Technology from James Cook University, JCU, Townsville, Australia (2012). In 2014, he joined the Faculty of Engineering in Systems, National Polytechnic School, EPN, as a Professor. Since October 2016, he has been working with the "Laboratoire d'informatique de Grenoble", LIG, Grenoble Alps University, UGA, Grenoble, France as a PHD student in the domain of Human Computer Interfaces. His research interests include children's learning, user interfaces, programming, adaptation, user modeling and music analysis.

Monserrate Intriago-Pazmiño was born in Chone, Ecuador, in 1984. She received the B.S. degree in Computer Science Engineering from National Polytechnic School, Quito, Ecuador, in 2007 and the M.S degree in Computer Science from the Technical University of Madrid, Spain, in 2011. She is currently a Ph.D. candidate in Computer Science at Technical University of Madrid. From 2008 to 2009, she was Assistant Professor with the Department of Informatics and Computer Science, National Polytechnic School. Since 2011, she has been a member of the Biomedical Informatics Group, Technical University of Madrid. Since 2014, she has been Professor at the Department of Informatics and Computer Science, National Polytechnic School. Her research interests include programming, software development, biomedical informatics, machine learning.

Strategic Scanning and innovative design: fuel the C/K method through Strategic Scanning information

Marie-Laurence Caron-Fasan, Justine Fasquelle, and Nicolas Lesca

Abstract— **The aim of this article is to study the role of Strategic Scanning in innovation processes. We seek to answer the following question: how can Strategic Scanning feed an innovative design activity? We mobilized the C-K theory which models the logic of creation in companies and the method of the same name. Based on a case study of 65 participants, we conducted a Strategic Scanning study to feed an innovative C-K design approach. The results show that Strategic Scanning helps to provide knowledge in a C-K process. It helps either to build the knowledge base of novice participants, or to validate the existing knowledge of expert participants. The Strategic Scanning activity also makes it possible to start the first disjunction mechanism C→K from the C0 concept.**

Index Terms— **C-K theory, Innovative design, knowledge space, Strategic Scanning.**

I. Introduction

ORGANIZATIONS now live in a world of "projects" where changes, sources of income, creation of competitive advantages ... arise from processes based on project-based organizations. Innovation projects are an illustration of this. Numerous companies, large and small, have established innovation projects that are more or less transversal with the aim of increasing their capacity to develop new products and/or competitive services. Others, in parallel, have implemented Strategic Scanning (SScan) projects to understand and anticipate changes in their external environments in order to reduce uncertainties related to decision making [1] and identify new opportunities in the market. Some have also linked their SScan activity to their innovation process with the hope of feeding innovation them and their related decisions with richer, more relevant and more anticipative information.

SScan is not always clearly defined in literature, nor is it really homogeneous inside a company. In their study, [2] admit that for all authors, SScan is an informative process whose the changes in its external environment and to support decisions [3]. Depending on the authors and the contexts of the studies, SScan process can take very different forms (ie. [4]-[6]). It can be individual, informal and unstructured, or organized and centralized. For example, it could take shape in the form of a cell, a service or an observatory [7]. The position of technology to support this process can also be very contrasting. Thus, SScan process can be completely computerized and use a dedicated platform for it, but it can also be based on a combination of numerous tools that are not very specific to SScan and that are weakly integrated and urbanized (for example, using Google search engines or curatorship tools such as Scoop it! to collect information coupled to emails or tweets for their diffusion).

However, there is not a single definition of an innovation process. Innovation is a particularly complex concept to address due to its multifaceted nature. It can respond to the desire to "do better, do things differently, do something else, do things faster, do fewer things or do things together" [8]. In a study dedicated to identifying the success factors of industrial innovation, [9] highlighted 5 generations of innovation processes. Since then, a sixth generation emerged at the beginning of the 2000s: The Open Innovation model, as shown in Table I.

The 6th generation process is the most implemented process in companies today. However, it should be noted that a significant number of companies are using 3rd generation innovation processes, such as the well-known Cooper Stage-Gate process. But, in fact, companies are also adopting a hybrid innovation process that mixes the 3rd and 6 th generations. Therefore, a company can adopt the Stage-Gate process (3rd generation) and consequently register its innovative activity in a context of open innovation of the 6th generation. Organized in more or less linear and/or parallel stages, each of these innovation processes involves making decisions with strong

TABLE I
THE 6 GENERATIONS OF INNOVATION PROCESS. ADAPTED FROM [9]

Generation	Date	Type of innovation process
1st generation	1950 - 1965	Technology-Push
2nd generation	1965 – 1970	Market-Pull
3rd generation	1970 - 1980	Interactive Model
4th generation	1980 - 1990	Integrated Model
5th generation	1990 - 2000	Model System integration and Networking
6th generation	2000- …	Open innovation

Marie-Laurence Caron-Fasan and Nicolas Lesca are both full professors at the University Grenoble Alpes in France. They are researchers of the CERAG-CNRS FRE 3748 (e-mail: marie-laurence.caron@univ-grenoble-alpes.fr, nicolas.lesca@univ-grenoble-alpes.fr).

Justine Fasquelle is a PhD Student at the Ecole Doctorale de Sciences de Gestion of the University Grenoble Alpes in France (e-mail: fasquelle.justine@gmail.com).

impacts on the progress of an innovation project.

The purpose of this article is to study the role of SScan in innovation processes. Its objective is to answer the following question: how can Strategic Scanning feed an innovative design activity? This research question forces us to direct our field of study to companies that carry out or have carried out SScan studies and that integrate these results into their innovative design activity. We use the C-K theory developed in the "École de Mines de Paris" by [10], which models the logic of expansion and creation in companies. We adopted a research approach of the type of case study by conducting an experiment in a school of creativity in the presence of professionals, academics and doctoral students. We used the CK method (declined of the theory of the same name), which is an innovative method for designing new products/services. The first results suggest that the use of previously directed and analyzed SScan information can help an innovative design process through the emergence of innovative concepts and new knowledge.

The first part presents the justification for the field of study. The second part explains the theoretical framework by presenting the C-K theory. In the third part, we establish the link between the innovative design by means of the C-K theory and the SScan process. The fourth part presents the research methodology of the case study, the results of which are presented and discussed in the fifth part.

II. The role of the Strategic Scanning in innovation projects

Research on the contributions of SScan activities to innovation processes is, to our knowledge, limited. Some authors, however, have demonstrated the role and importance of SScan in the development of innovation.

[11] explains that entrepreneurs who wish to develop their creativity should proactively and frequently look for information from sources, especially when their environment is very turbulent. [12] explain that creativity is nourished by the internal and external environment of the company. As a result, SScan. by means of contribution of knowledge of the external environment, is likely to fuel creativity.

Other research has shown that scanning the environment to collect information is a critical activity for identifying opportunities to develop and invigorate the innovation process [14]. The role of "Champions", defined by [13] as "heroes of innovation", stands out as well. They are people who carry out SScan to promote innovation to interest groups, overcome resistance to innovation and obtain the essential resources for the development of innovation. These people can, according to [14], correlate technical problems with external scientific knowledge and technical developments with market demand by identifying innovations with potential. [15] have shown that "champions" can accelerate the product innovation process by collecting and applying external information to development activities.

[16] showed that monitoring the technological environment would facilitate the design and introduction of market innovations. They specify that the companies that introduce the best new products in the market are those that have flexible SScan process that are adapted to the environment.

Finally, [17] explains that anticipatory SScan oriented towards usages would reduce the risk of divergence between market needs and innovation. This type of SScan leads to "identifying future uses of emerging technologies. Its general principle is based on the observation of the dynamics of technological innovation, as well as on the dynamics of social innovation in order to anticipate their potential convergences" [17].

Several authors have shown that companies that follow up (including technology) are more likely to innovate. Thus, SScan seems to play a role in the innovation capabilities of organizations. However, these studies do not show or have not tried to understand how SScan could fuel an innovation process to make it more efficient.

In conclusion, we can say that SScan, as an activity of gathering and processing external information, can play a role in an innovation project.

III. Innovation Project and Innovative Design: The C/K Theory

Like [18], we suggest that "innovation is an (non-systematic) output of innovative design". In fact, a design activity implements reasoning, one or more models and associated performance criteria that allow the development of projects and, therefore, innovation.

Developed initially by Armand Hatchuel and Benoit Weil, and later by Pascal Le Masson, the C-K theory is a theory about the design of innovative products or services [19]. Developed in the years 2002 and 2003 [10], [20], this theory aims to present a unified approach of design in order to provide a theoretical framework that integrates all kind of design activities (regulated and innovative).

This theory distinguishes two fundamental notions: the notion of "knowledge" and the notion of "concept". The authors discuss a knowledge space called space-K and a concept space called space-C.

The knowledge space is defined as a set of propositions that have a logical state. This space describes all the objects and truths, in other words, established knowledge. The space K is expandable as new facts and truths become available.

The space-C is defined as a proposition without a logical state. A concept has an unknown or undefined part. The concepts are the starting points for an innovative design process. Without a concept, the design is reduced to optimizing the existing ones and solving problems.

In the C-K theory, the innovative design process is based on a back and forth between space C and space K, a transformation of concepts into knowledge and vice versa. There is, therefore, a gradual expansion of the two spaces by mutual enrichment. This back and forth between spaces is modeled by external and internal operators.

There are two external operators: (1) the K→C disjunction that makes it possible to convert knowledge into the formulation of a concept by the addition of new properties or attributes, and (2) the C→K conjunction, which transforms a

concept into knowledge and therefore corresponds to a validation of the concept, and in a practical form, to a "finished design".

There are two internal operators: the C→C operator that defines the partition process of the set of concepts, and the K→K operator, which defines the expansion of knowledge.

The C-K theory implements several steps:

- Transform an initial proposition into a concept C0. This concept is therefore derived from a K→C disjunction. This disjunction must respond to two principles: (1) that all the terms of this proposition belong to propositions of K, and (2) that this proposition has no logical status, otherwise it would be an acquaintance of K. Because it belongs to the domain of concepts, this proposition has no logical status. It is neutral, neither false nor true.
- Add attributes and properties from the K space to the C0 concept. The goal is to expand the concept domain by proposing new concepts.

At this stage, two options are possible:

- The designers may consider that they know how to design one of the new concepts, in which case this concept acquires a logical, true state and becomes known. Consequently, the design reasoning can be stopped.
- Or, these new concepts are "neutral", in which case it is necessary to prolong the reasoning by making a new partition with the help of knowledge.

The C-K theory allows the formalization of the appearance of new concepts as well as the development of new knowledge. It offers a structured framework for the increasingly advanced development of an initial concept of C0 and a developed knowledge space at the beginning of the innovative design process.

IV. Innovative design and Strategic Scanning

We have seen that SScan can play a role in innovation, foster creative approaches, identify opportunities for development and revitalization of the innovation process, help introduce new products into markets or even reduce the risk of divergence between market needs and the new products/services.

However, we cannot find any research detailing the role of SScan in innovation and even less research that links innovative design and SScan. However, when we read professional journals and interview innovation leaders, the link between SScan, innovation and innovative design is clear, almost as obvious as it is shown in the Table II.

SScan has several purposes: it can gather information to build a state-of-the-art of current knowledge of the field/topic on which you want to work. You can pretend to anticipate future changes that do not yet exist by collecting weak signals [21] that can identify new threats and opportunities.

Whether for the construction of a state-of-the-art of current knowledge or for the identification of future changes, SScan is based on increasingly sophisticated and accurate computer tools used in the collection and analysis of information. Developed by service companies and often paid but sometimes free, these tools can collect a lot of information from a wide variety of sources, analyze it automatically and then present it in the form of panels and/or graphics. This is the case, for example, of data mining tools such as Thomson Innovation patent database or Space Net (a free tool) that allow, thanks to a keyword search, to analyze existing patents. Tools like Ixxo, E-Perion or Izi'Nov allow one to track the web, visible and invisible, and provide information of all kinds that will be filtered according to the needs of the user. Digimind offers monitoring software that monitors social networks and allows the company to be attentive.

TABLE II
ANSWERS TO THE QUESTION OF THE LINK BETWEEN SSCAN AND INNOVATIVE DESIGN DURING INTERVIEWS WITH DIRECTORS OF LARGE GROUPS

Innovation Director (March 2017)
"If you want to put yourself in a correct C-K dialogue, you should navigate in C and look at the known and unknown K and do the iterations recommended by the method, so we are in the construction of the K tree in SScan before embarking on the prototype to make sure we have the right knowledge and that we will do well in the project in question".
Director of Technology and Innovation by Uses (March 2017)
"Then, SScan is K. Knowledge, full of observations, full of learning. We are smarter, so being smarter, we may have a little more C, a little more concept. "

The use of data mining tools for innovative design has already been proven. [22] compared the results of creativity sessions of students in a school of engineering. Some groups used data mining tools, others did not. The conclusion is clear: the teams that used data mining developed more sophisticated offers than those that did not explore the data. They were able to identify solutions they had not considered before and offered a more advanced product to potential users.

These tools allow creative teams to think beyond what already exists [23]. As information is growing in size and will not stop growing, it is difficult (not to say impossible) for businesses to capture it all. Data mining techniques capture a lot of information, including weak signals, and thus offer the possibility to explore new trends or interesting functionality for a team wishing to go beyond their acquired knowledge.

Therefore, we consider that computer tools for SScan and, in particular, data mining tools can support innovative design, either through the collection of information that allows us to develop the latest advances in knowledge or by identifying weak signals useful to detect new trends. Based on these research results, we created an experiment with the objective of using a computer SScan tool to feed an innovative design approach through the C-K theory. We developed this case study in the following section.

V. Research methodology and data collection

A. The case study method

The research method used in this research is case study. The work of [24] has contributed to its development and legitimacy

by highlighting its scientific interest. We define a case as a set of empirical data that is related to a reality and that fits a situation which constitutes a unit of analysis. The concept of context is very important: all the results obtained must be analyzed with respect to the specific context of the case study.

The case study is based on the two principles of internal validity and external validity. [25] define the internal validity of qualitative research as the existence "on the one hand, of "only", "authentic" and "plausible" results in relation to the field(s) of study, and on the other hand, of results related to a previous or emerging theory".

External validity refers to the generalization of results. This validity is often presented as an important limitation of the case study. However, it is still possible under certain conditions of representativeness and transferability of the results [25].

However, the recognition of the case study is mainly based on internal validity. That is the only topic that interests us in this article: the measure of this internal validity. We do not intend to generalize the results obtained during the experiment described below.

Finally, it should be noted that two of the three authors of this article had observer status. One of the researchers participated in the design and preparation phase of the experiment; the other researcher participated in the experiment.

B. Context

The experiment took place in the Winter School of Creativity in Grenoble organized by Promising. Coordinated by the University of Grenoble Alpes, Promising is a Future Investment Program IDEFI financed by the French State for 5 million euros for 7 years. Its objectives are, among other things, to develop the collective intelligence of innovation and to explore the relationships between understanding and acting in situations of innovation. The ambition to contribute to the development of a collective intelligence of innovation involves the exploration of new forms of experimentation regarding projects that involve doctoral students, professors and the socio-economic world.

The Winter School of Creativity is a training program in the form of workshops. It is aimed at managers, employees of the private and public sector, teacher, researchers and doctoral students. This is an active learning path with theoretical contributions, tools and methods, case studies, and concrete techniques, all taught through practice. The Grenoble Winter School of Creativity is part of an international network of School of Creativity under the auspices of MOSAIC - School of Creativity HEC Montreal. The network includes the cities of Lille, Bangkok, Strasbourg and Grenoble.

In 2016, the Grenoble Winter Creativity School welcomed 65 participants, including 25 companies, 12 public organizations and 27 teacher-researchers and doctoral students. One of the 17 workshops was dedicated to experiencing the relationship between SScan activity and creativity in relation to the C-K theory.

C. Conducting the experiment

The objective of the experiment was to practice the C-K method (derived from the C-K theory) in half a day by explicitly linking it with a SScan activity. The innovative aspect that we wanted to test was related to make participants use the results of a SScan activity by mobilizing an innovative design method in order to bring innovative concepts to light.

1) Preparatory phase

This experiment required a preparatory phase of work in two stages.

Preparatory step 1: First, the identification of concept 0. For this, four experts (a consultant specialized in the C-K method, an innovation consultant with good knowledge of the field and sports-related issues, a researcher from the Ecole des Mines, from which the C-K theory came from, and a researcher specializing in the field of SScan) conducted a two-hour brainstorming session to identify and establish a concept 0 as clearly as possible.

Concept 0: Make stadium fans enjoy in an innovative way

Preparatory step 2: The goal was to build an initial K knowledge space. To this, we worked with a company specialized in SScan that also offered consulting and services in research and analysis of data and information. It proposed, in particular, SScan studies, benchmarking studies, thematic states-of-the-art, patent mappings, and networks of actors, partners, and clients.

During a one-hour meeting, we worked with an employee of the company and a specialist in the use of SScan tools to identify keywords related to the concept 0. Both the consultants and the innovation researcher tried to verbalize what they wanted to say by "making stadium goers enjoy in an innovative way". Based on this very informal discussion, the SScan expert identified keywords. Then he looked for intelligence information related to concept 0. His research approach is shown schematically in Fig. 1. However, the first results obtained, which were already satisfactory, were refined by a new research work based on new keywords.

Thus, following two iterations of information search via monitoring tools, we obtained the information detailed in Table III.

Then, an information analysis tool structured this large amount of information. In this way, the analysis of the information allowed extracting 44 documents to feed the K knowledge space. These 34 documents were structured as shown in Table IV.

These 34 documents, which can be either patents, excerpts from scientific publications or web sources, were gathered in a 39-page document for distribution to the participants of the Grenoble Winter School of Creativity. Fig. 2 is an excerpt from this document.

2) Mobilization phase of the C-K method

During the half day of the workshop, the 65 participants at the Grenoble Winter School of Creativity were divided into 11 groups. For 45 minutes they were informed about the work expected of them (presentation of the purpose of the session, the C-K method, the concept 0 and the information constituting the initial K space (the 34 documents selected)). Each participant received a document that contained the SScan information. Then, they worked independently (with the help of 2 specialists of the C-K method if necessary).

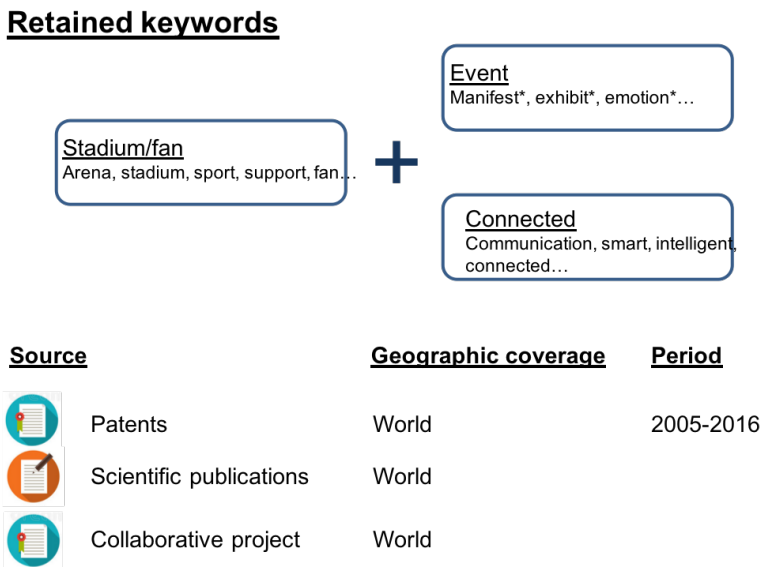


Fig. 1. Information search process.

Using the documents of concept 0 and the other 34 documents, they tried to propose a concept 1 and a concept 2. Each time it was necessary, and on the basis of concepts 1, 2, 3, etc., they could contribute to space K using the 34 documents but also their own knowledge, since some of the participants were sports professionals or members of sports companies, and others were passionate about sports. The iterations between the concepts C-space and K-knowledge space were completed when each group considered that the identified concept was sufficiently new and had a non-neutral logical state. Then, each group presented its concept to the other groups. It should be noted that all the groups were able to propose a new concept. Finally, one of the specialists of the C-K method presented his own reasoning on the basis of the concept 0 and the 34 documents of the space K (Fig. 3).

VI. Results and Conclusion

The purpose of this article is to begin a reflection on the link between SScan and innovative design. The experiment executed with 65 people shows that SScan can be integrated into an innovative design approach of type C-K. In half a day, each group, based on an initial K-space, succeeded in identifying one or more innovative concepts. Although these concepts are still underdeveloped and should be elaborated upon to produce real tradable innovations.

The research question was to understand how SScan activity can fuel an innovative design activity. The first results give some answers.

A. Expansion of space K and space C

The experiment showed that the participants were able to increase the knowledge space K (knowledge) as well as the space C (concepts).

1) The knowledge space K

Without any difficulty, the participants seized the document that contained the SScan information. They saw it as a starting point, as the initial knowledge space that they could mobilize. The monitoring document has thus replaced the K knowledge space. Therefore, SScan has come to fuel the knowledge space.

SScan served as a catalyst for the innovative design work of the C0 concept. All participants were curious to discover what this document might contain. All used it as an initial knowledge base, that is, initial K space. However, there are two types of behavior. The first was that of the beginners, who had little or no knowledge related to the C0 concept. These beginners used the reservation information as support for ideation and divergence reasoning. On the basis of certain information, they did not hesitate to diverge on ideas of original concepts, which were sometimes very original and disconnected from any realistic consideration. The latter, sports experts or sports enthusiasts, used the SScan information as validation for the knowledge they already had. They also managed to draw concepts. It is important to bear in mind that they generated many more concepts than the novices. However, some of these concepts were very originals and strongly anchored in themselves in pragmatism and feasible concepts. These experts also played the role of moderators against the concepts of beginners. They mobilized their own knowledge, their own K space to modify (and sometimes judge) novice concepts.

TABLE III
Information related to Concept 0 after search

2000 patents
40 collaborative projects
500 scientific publications
3000 other information from the web
2000 patents
40 collaborative projects

TABLE IV
Information related to Concept 0 after Analysis

11 documents related to connected stadiums	2000 pate
9 documents related to connected textiles	40 collabo
7 documents related to intra-stadiums	500 scient
7 documents related to visual displays in sports enclosures	40 collabo

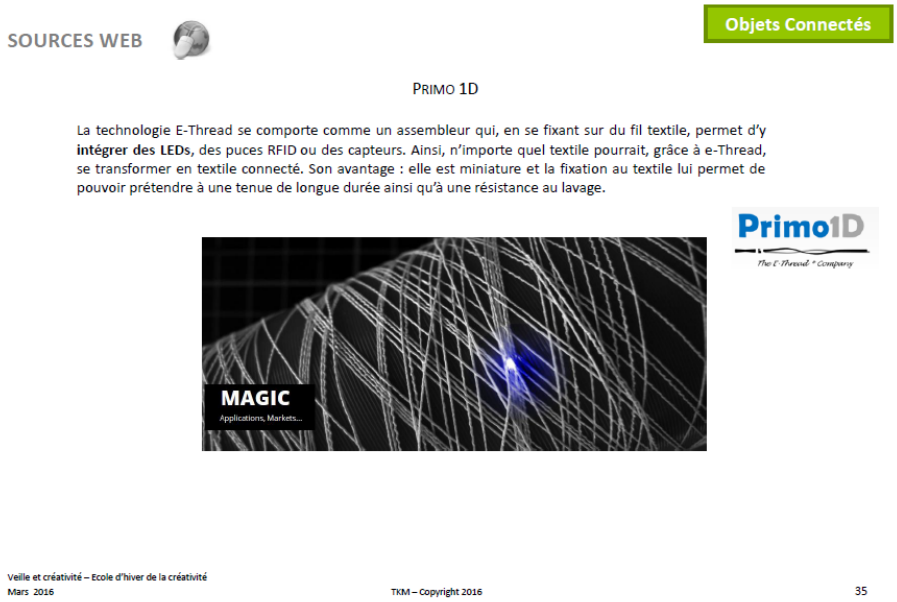


Fig. 2. Extract of web-based information about connected objects

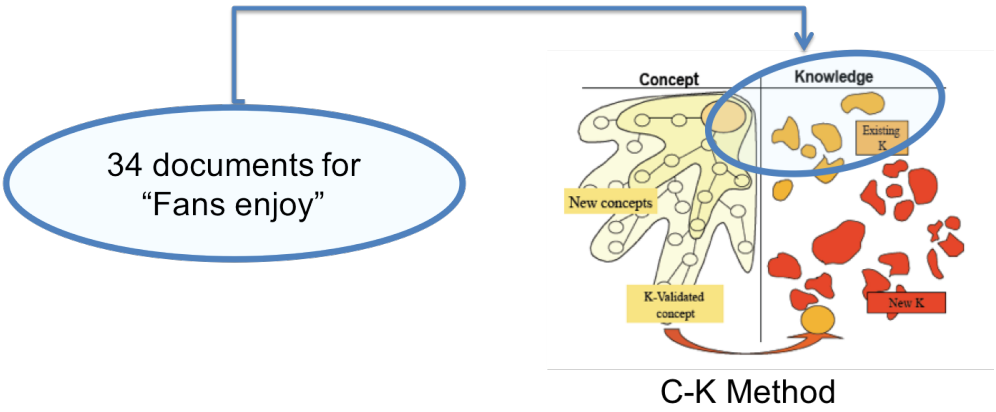


Fig. 3. Iteration between space C and space K according to the documents of the concept 0 and the 34 documents of the SScan work.

Therefore, we can say that the groups have used a disjunction mechanism K→C that allows movement from a set of knowledge to the formulation of one or more concepts by adding innovative properties or attributes. In addition, the initial knowledge space has been increased through informal exchanges of information between participants or through the creation of new knowledge. Finally, and especially among beginners, the internal operator K→K has allowed an expansion of knowledge.

2) Concept space C

The concept space has been enriched. Therefore, each group was able to present an innovative concept related to the C0 concept. Some groups were able to draw several concepts. Other groups had more difficulties and sometimes stopped at the appearance of a concept. However, and probably due to lack of time, the concepts presented were not successful concepts. Several participants regretted that they did not have more time to continue the experiment. Therefore, the conjunction mechanism C→K was not mobilized due to lack of time and support for the mobilization of the C-K method. The mobilization of operator C→C was neither observed.

B. A real help but limited for fueling C0 concepts

The definition of the concept C0 is the catalyst step of C-K. SScan carried out in the experimentation had a limited role in helping to define this initial concept.

1) Definition of concept 0

This is the primordial stage of the C-K theory. However, it is used very little in the C-K method and, sometimes, it is difficult to put into practice [26]. In the context of experimentation, the construction of the C0 concept was carried out with the help of specialists in innovation and the C-K method. It was based on their personal knowledge, whether in the field of sports and innovation, or the C-K method. SScan and the collection of information that could help identify the C0 concept, was not mobilized. However, we can question here the role that SScan could play. The collection of very general information about sports and its automatic analysis could have provided knowledge that would have helped in the development of the C0 concept. This knowledge could have played the role of K0 and could represent a state-of-the-art form in the treated field.

Therefore, it would be interesting to respond to the dissatisfactions of the C-K method on the elaboration of the C0 concept, to study the contribution of SScan, in particular, in very general information about the field studied. Therefore, a research track could study the role of SScan in the development of the C0 concept.

2) Fueling the C0 concept by identifying the information needs

The implementation of the document delivered to the participants as an initial knowledge space implies some conditions. First, the SScan expert's ability to listen and reformulate in order to identify keywords representative of the C0 concept is vital. Round trips are also needed to refine and validate the final document. Similarly, it also assumes the ability of the initiators of the C0 concept to verbalize their ideas in a clear and unambiguous manner. Knowing that a concept is defined as a proposition without a logical state, this task is not easy. The identification of the need for SScan information depends on the capacities of the different actors to verbalize, exchange, understand and reformulate the C0 concept. In a traditional SScan activity, the identification of needs is also a prerequisite for the collection of information because no organization has the resources to scan all of its environment [27]. It corresponds to the identification of strategic objectives and priorities in terms of information collection to optimize the allocation of resources necessary for the observation activity, obtaining useful results and avoiding the failure of the SScan project [28],[29]. Detailed and instrumented by [3] and [21], the identification of SScan information needs to identify the part of the environment that needs to be monitored as a priority. [21] implemented this step by constructing a method called Target®. It helps to identify the actors (competitors, customers, partners, etc.) and the issues (e.g., regulatory, technological, etc.) that will be prioritized under SScan. It would be interesting to study to what extent this scanning tool that targets the environment to be monitored could be used in the definition of the C0 concept. Is it relevant to use the notions of actors and themes to define the C0 concept? Could the actors and issues identified thus be used as keywords for the search for information?

3) Analysis and restitution of information

Fueling the C0 concept in information means being able to construct a readable and usable document: A document that presents the information in an easily understandable text and visual form. Therefore, it is about having a SScan tool with analysis and advanced graphic functions. Today, monitoring tools of data mining style have all these characteristics. Therefore, this step is not a problem as long as you have been able to gather potentially interesting information.

The experiment carried out during the Winter Creativity School shows the role of SScan in an innovative design activity. It suggests that SScan is an activity that contributes to the expansion of the knowledge space K and the concept space K. It also suggests that SScan can fuel specific information to the C-K method and more specifically to the C0 concept. However, this role is subject to a number of conditions, including the ability to fuel relevant information into the initial idea of the C0 concept. Its results make it possible to provide a beginning of answer to our research question. The internal validity criterion is thus satisfied. However, these results should be considered only in relation to the context of the experiment and not to claim any generalization.

Some lines of research have emerged. Experimentation shows that SScan cannot be limited to collecting information and producing a document. It would be interesting to use information identification methods to better support the construction phase of the C0 concept. In fact, after this first experiment, other experiments were carried out. They are being analyzed, but the first results show that professionals find great interest in coupling SScan and the C-K method during their innovative design activities.

REFERENCES

[1] B. Walkers, J. Jiang and G. Klein, "Strategic information and strategic decision making: the EIS/CEO interface in smaller manufacturing companies", *Inform. Manage.*, vol. 40, no. 6, pp. 487-495, 2003

[2] E. Loza-Aguirre, M.-L. Caron-Fasan, Lesca, N. and M. C. Chalut-Sauvannet, "Drivers and barriers to pre-adoption of strategic scanning information system in the context of sustainable supply chain", *Systèmes d'Inform. Manage.*, vol. 20, no. 3, pp. 9 – 46, Sept. 2015.

[3] C. W. Choo, "The Art of Scanning the Environment", *ASIS Bull.*, vol. 25, no. 3, pp. 13-19, 1999.

[4] C.-P. Wei and Y.-H. Lee, "Event detection from online news documents for supporting environmental scanning", *Decis. Support Syst.*, vol. 36, no. 4, pp. 385-401, 2004

[5] R. Y. K. Lau, S. S. Y. Liao, K. F. Wong and D. K. W. Chiu, "Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions", *MIS Quart.*, vol. 36, no. 4, pp. 1239-1268, 2002.

[6] J. H. Mayer, N. Steinecke, R. Quick and T. Weitzel, "More applicable environmental scanning systems leveraging « modern » information systems", *Inf. Syst. E-Bus. Manag.*, vol. 11, n°4, p. 507-540, 2012

[7] C. Belmondo, "Les phases de création des connaissances dans une cellule de veille: comparaison de deux processus", *Systèmes d'Inform. Manage.*, vol. 2, no. 8, pp. 41-68, 2003

[8] Giget M., "L'innovation dans l'entreprise", *Traité Génie industriel, Techniques de l'ingénieur*, Mai. 1994.

[9] R. Rothwell, "Successful Industrial Innovation: Critical Factors for the 1990s", *R&D Manag.*, vol. 22, no. 3, pp. 221-38, 1992

[10] A. Hatchuel and B. Weil. "La théorie C-K : Fondements et usages d'une théorie unifiée de la conception", Colloque Sciences de la conception, 2002.

[11] J. Tang, "Linking personal turbulence and creative behavior: the influence of scanning and search in the entrepreneurial process", *J. Bus. Res.*, Vol. 69, no. 3, pp.1167-1174, 2016.

[12] Y. Bertacchini and C. Strasser, C, "Intelligence économique et créativité au sein de la PME/PMI : une compétence offensive à organizer", *Revue internationale d'intelligence économique*, vol 3, no. 1, pp. 13-35. 2011.

[13] J. M. Howell and C. M. Shea, "Individual differences, environmental scanning, innovation framing, and champion behaviour: key predictors of project performance", *J. Prod. Innov. Manag.*, vol.18, no. 1, pp. 15-27, 2001.

[14] R. Burgelman and L. Sayles, *Inside Corporation Innovation*, New York: Free Press, 1986.

[15] R. T. Kessler and A. K. Chakrabrati, "Innovation speed: a conceptual model of context, antecedents, and outcomes", *Acad. Manage. Rev.*, vol. 21, no. 4, pp.1143-1191, 1996.

[16] N. Ahituv, J. Zif and I. Machlin, "Environmental scanning and information system in relation to success in introducing new products", *Inform. Manage.*, vol. 33, no. 4, pp. 201-211, 1998

[17] M-L. Caron-Fasan, "Accompagner l'innovation dans les entreprises » De la veille technologique à la veille usage anticipative", *La Revue des Sciences de Gestion*, vol.3, no. 231-232, pp. 19-26, 2008.

[18] G. Garel, Le processus d'innovation, conception innovante et croissance des entreprises, Paris: Hermès Lavoisier, 2006.

[19] P. Le Masson, B. Weil and A. Hatchuel, Les processus d'innovation, conception innovante et croissance des entreprises. Paris: Hermès, 2006.

[20] A. Hatchuel, P. Le Masson and B. Weil, "C-K Theory in Practice: Lessons from Industrial Applications", 8th International Design Conference, D. Marjanovic, (Ed.), Dubrovnik, 18–21 May 2004: 245–257.
[21] H. Lesca and N. Lesca, Strategic decisions and weak signals: anticipation for decision-making, London: ISTE Ltd, 2014
[22] M.L. Escandon-Quintanilla, M. Gardoni and P. Cohendet, "Improving concept development with data exploration in the context of an innovation and technological design course", *Int. J. Interact. Design Manuf.*, vol 12, no. 1, pp. 161-172. 2017.
[23] M. Agogué, A. Kazakçi, A. Hatchuel, A., P. Masson, B. Weil, N. Poirel, and M. Cassotti, "The impact of type of examples on originality: explaining fixation and stimulation effects", *J. Creative Behav.*, vol. 48, no. 1, pp. 1-12. 2014.
[24] R. K. Yin, Case Study Research: Design and Methods, Applied Social Research Methods Series, London: Sage Publications. 2003
[25] C. Ayerbe and S. Missonier, "Validité interne et validité externe de l'étude de cas: principes et mise en œuvre pour un renforcement mutuel", *Finance Contrôle Stratégie*, vol. 10, no. 2, pp 37- 62. Juin 2007
[26] O. Pialot, "L'approche PST comme outil de rationalisation de la démarche de conception innovante", Ph.D. dissertation, France, 2009
[27] M. Franco, H. Haase, A. Magrinho, and J. Silva, J. (2010), "Scanning practices and information sources: an empirical study of firm size", *J. Enterp. Infor. Manage.*, vol. 24 no. 3, pp. 268-287. 2010.
[28] N. Lesca and M.-L. Caron-Fasan, "Strategic Scanning Project Failure and Abandonment Factors: Lessons Learned", *Eur. J. Inform. Syst.*, vol. 17, no. 4, pp. 371-386, 2008
[29] X. Zhang, S. Majid and S. Foo, "Environmental scanning: an application of information literacy skills at the workplace", *J. Inform. Sci.,* vol. 36, no. 6, pp. 719-732. 2010.

Marie-Laurence Caron-Fasan is a Full Professor of Management of Information System at the Université Grenoble Alpes, and Head of the research department of Management of Information Systems and Flow at the laboratory CERAG CNRS FRE 3748. Her research lies in the study of anticipative strategic scanning system, the use of weak signals, and the links between innovation processes and strategic scanning systems. She has co-authored books on Information Systems and Strategic Scanning, and she has published in French and International journals such as the European Journal of Information Systems, Information & Management and the French Journal of Management Information Systems. Her research has been presented at leading scholarly conferences such as EURAM and IEEE Conference on Enterprise Systems.

Justine Fasquelle is a PhD student at the Université Grenoble Alpes and researcher at the laboratory CERAG CNRS FRE 3748. Managed by Pr Marie-Laurence Caron-Fasan, her dissertation examines the links between Strategic Scanning and innovation. More specifically, it concerns the role of information stemming from an activity of strategic scanning on decision-making throughout the innovation process. She has also been working for the last 3 years as expert in strategic scanning.

Nicolas Lesca is a Full Professor of Management of Information Systems at the Université Grenoble Alpes, researcher at the laboratory CERAG CNRS FRE 3748, and Vice-president in charge of Academic Affairs at Université Grenoble Alpes. His research interests lie in designing and experimenting with methods and systems to develop managerial skills and abilities to detect, disseminate, share, interpret and make sense of weak signals to support anticipation and decision-making. He is also interested in the adoption, the diffusion and the appropriation of such methods and systems in organizations, as well as in their success and failure factors. He has authored and co-authored books, book chapters, and articles in academic journals such as the European Journal of Information Systems, the European Journal of Information Systems, Information & Management, and the French Journal of Management Information Systems. His research has been also presented at leading scholarly conferences.

Published by:

Escuela Politécnica Nacional
Facultad de Ingeniería de Sistemas
Departamento de Informática y Ciencias de la Computación
Ecuador

http://lajc.epn.edu.ec/
lajc@epn.edu.ec

November 2017