CONGRESO INTERNACIONAL

Congreso Internacional de
Ciencias de la Computación y Sistemas de Información

Universidad Champagnat - Universidad Nacional de San Juan
Mendoza, Argentina - 19 al 20 de noviembre de 2020

CICCSI 2020
Libro de Actas

Universidad Champagnat, Universidad Nacional de San Juan, Universidad Católica de Pereira.

Anales del IV Congreso Internacional de Ciencias de la Computación y Sistemas de Información: CICCSI 2020 / Fernando Pinciroli ... [et al.] ; compilación de Fernando Pinciroli. - 1a ed. - Godoy Cruz : FUSMA Ediciones, 2021.
Libro digital, PDF

Archivo Digital: descarga
ISBN 978-987-45683-9-7

1. Computación. 2. Sistemas de Información. I. Pinciroli, Fernando, comp.
CDD 004.0711

Diseño de portada y logo: Pinciroli Braceli, María de los Ángeles.
Editor LaTeX: Méndez-Garabetti, Miguel.

Congreso Internacional de Ciencias de la Computación y Sistemas de Información CICCSI 2020

Mendoza, Argentina, 19 y 20 de noviembre de 2020.
Instituciones Organizadoras:

Universidad Nacional de San Juan, San Juan Argentina.
Universidad Champagnat, Mendoza, Argentina.
Universidad Católica de Pereira, Colombia.

Fecha de Catalogación: abril de 2021

Comité Científico

Juan Iglesia, Universidad Champagnat, Argentina
Miguel Méndez-Garabetti, Universidad Atlántida Argentina, Universidad Champagnat, LICPaD, UTN-FRM, Argentina
José Luis, Barros Justo, Universidade de Vigo, España
Laura Zeligueta, Universidad Champagnat, Argentina
María Inés Lund, Universidad Nacional de San Juan, Argentina
Nelson Rodríguez, Universidad Nacional de San Juan, Argentina
Laura, Nidia Aballay, Universidad Nacional de San Juan, Argentina
Manuel Ortega, Universidad Nacional de San Juan, Argentina
Mario Mirallas, Universidad Champagnat, Argentina
Carolina Canessa, Universidad Champagnat, Argentina
Alejandra Malberti, Universidad Nacional de San Juan, Argentina
Raúl Klenzi, Universidad Nacional de San Juan, Argentina
Juan Aranda, Universidad Nacional de San Juan, Argentina
María Murazzo, Universidad Nacional de San Juan, Argentina
Daniel Díaz, Universidad Nacional de San Juan, Argentina
Fabiana Píccoli, Universidad Nacional de San Luis, Argentina
Sandra Rodríguez, Universidad Nacional de La Rioja, Argentina
Rosana Giménez, Universidad del Aconcagua, Argentina
Patricia Ontiveros, Universidad Tecnológica Nacional-FRM, Argentina
Fabiane Benitti, Universidade Federal de Santa Catarina, Brasil
Ania Cravero, Universidad de La Frontera, Chile
Samuel Sepúlveda, Universidad de La Frontera, Chile
Alonso Toro Lazo, Universidad Católica de Pereira, Colombia
Juan Carlos Blandón Andrade, Universidad Católica de Pereira, Colombia
Álex Armando Torres Bermúdez. Corporación Universitaria Comfacauca, Colombia
José Luis Arciniegas Herrera, Universidad del Cauca, Colombia
Martín Monroy Ríos, Universidad de Cartagena, Colombia
Sonia Mora, Universidad Nacional de Costa Rica, Costa Rica
Santiago Matalonga, University of the West of Scotland, Escocia
Gilberto Borrego Soto, Universidad Autónoma de Baja California, México
Gisela Torres, Universidad Tecnológica de Panamá, Panamá
Gerardo Maturro, Universidad ORT, Uruguay
Lucía Cortés, Facultad de Ingeniería, Universidad Nacional de Cuyo, Argentina
Jorge Guerra Torres, Facultad de Ingeniería, Universidad Nacional de Cuyo, Argentina
Martin Marchetta, LABSIN, Facultad de Ingeniería, Universidad Nacional de Cuyo, Argentina
Pablo Godoy, ITIC, Facultad de Ingeniería, Universidad Nacional de Cuyo, Argentina
Facundo Bromberg, DHARMA, Facultad Regional Mendoza, Universidad Tecnológica Nacional, Argentina
Gabriel Caffaratti, LABSIN, Facultad de Ingeniería, Universidad Nacional de Cuyo, Argentina
Osvaldo Marianetti, Universidad de Morón, Argentina
Yisel Gari, ITIC, Facultad de Ingeniería, Universidad Nacional de Cuyo, Argentina
Sebastian Perez, DHARMA, Facultad Regional Mendoza, Universidad Tecnológica Nacional, Argentina
Carlos Diaz, DHARMA, Facultad Regional Mendoza, Universidad Tecnológica Nacional, Argentina
Elina Pacini, CONICET, Universidad Nacional de Cuyo, Argentina
Pablo Vidal, CONICET, Universidad Nacional de Cuyo, Argentina

Josefina Huespe, Facultad de Ingeniería, Universidad Nacional de Cuyo, Argentina
Ana Diedrichs, DHARMA, Facultad Regional Mendoza, Universidad Tecnológica Nacional, Argentina
Lucas Iacono, Facultad de Ingeniería, Universidad Nacional de Cuyo, Argentina
Marisa Haderne, Facultad de Ingeniería - UNCUYO, Argentina
Carlos Taffernaberry, GRIDTICS, Facultad Regional Mendoza, Universidad Tecnológica Nacional, Argentina
Ana Olivera, CONICET, Facultad de Ciencias Médicas, Universidad Nacional de Cuyo, Argentina
Rodrigo Gonzalez, Universidad Nacional de Cuyo, Argentina
Paula Martinez, Universidad Nacional de Cuyo, Argentina
Carlos Garcia Garino, ITIC, Universidad Nacional de Cuyo, Argentina

Comité Académico

General Chair:
Raymundo Forradellas, Universidad Nacional de San Juan, Argentina

Program Chairs:
José Luis Barros Justo, Universidade de Vigo, España
Javier Rosenstein, Universidad Champagnat, Argentina
Enzo Completa, Universidad Champagnat, Argentina
María Inés Lund, Universidad Nacional de San Juan, Argentina

Workshops Chairs:
Fernando Pinciroli, Universidad Champagnat, Argentina
Laura Nidia Aballay, Universidad Nacional de San Juan, Argentina

Research Meeting Chairs:
Nelson Rodríguez, Universidad Nacional de San Juan, Argentina
Carlos Catania, Universidad Nacional de Cuyo - Universidad Champagnat, Argentina

Professors Meeting Chairs:
Manuel Ortega, Universidad Nacional de San Juan, Argentina
Laura Zeligueta, Universidad Champagnat, Argentina

Comité Organizador

Fernando Pinciroli, Universidad Champagnat
Susana Dueñas, Universidad Champagnat
Javier Rosenstein, Universidad Champagnat
María Inés Lund, Universidad Nacional de San Juan
Carlos Catania, Universidad Champagnat
Verónica Miguez, Universidad Champagnat
Laura Nidia Aballay, Universidad Nacional de San Juan
Manuel Ortega, Universidad Nacional de San Juan
Miguel Méndez-Garabetti, Universidad Champagnat
Sandra Oviedo, Universidad Nacional de San Juan
Marcelo Palma, Universidad Champagnat
Line Yazmin Becerra Sánchez, Universidad Católica de Pereira
Lina María Suárez Vásquez, Universidad Católica de Pereira

Prólogo

El Congreso Internacional de Ciencias de la Computación y Sistemas de Información, CICCSI, es la actividad académica de referencia de la región y se ubica entre las más importantes de la Argentina. Su objetivo es reunir en un evento de periodicidad anual a investigadores, profesionales y alumnos de diferentes países, al promover un espacio para la difusión de los conocimientos generados en los ámbitos académico y de la industria. Con esta actividad se busca potenciar el intercambio entre investigadores, profesionales y estudiantes de informática, como así también de todos los participantes al evento.

En esta cuarta edición, el CICCSI 2020, que fue organizado por la Universidad Champagnat y coorganizado por la Universidad Nacional de San Juan y la Universidad Católica de Pereira, nuevamente marcamos un hito importantísimo en la historia de los congresos internacionales de informática del interior de la Argentina, al ser el primer evento que alcanza cuatro ediciones consecutivas.

Esto pone de manifiesto el compromiso de las instituciones organizadoras en seguir potenciando el nivel académico con el que están comprometidos y también la aceptación de todos los participantes, tanto de aquellos que nos acompañaron como protagonistas en las conferencias, presentaciones de artículos y reuniones, como así también de parte de quienes asistieron a las diferentes actividades, con cuyos comentarios y aporte de ideas nos permite seguir mejorando CICCSI año tras año.

A raíz de la pandemia de COVID que sigue sufriendo el mundo, este año, todas las actividades se realizaron completamente en línea, pero esto no impactó en el interés y la participación de los asistentes al evento. Entre sus actividades, CICCSI 2020 contó con quince conferencias de expositores internacionales de la academia y la industria, de Alemania, Escocia, Colombia, España y Argentina, una conferencia Magistral desde los Estados Unidos, un encuentro entre dos de los autores del Manifiesto Ágil, Arie van Bennekum y Alistair Cockburn, un encuentro de profesores de las áreas de Ciencias de la Computación y de Sistemas de Información y uno de investigadores, con representantes de diferentes países de América Latina, y con su actividad principal, que es la presentación de artículos científicos y posters de autores de diferentes instituciones académicas de diversos países. Todas estas actividades quedaron disponibles en línea.

Nuevamente, agradecemos la participación de todos los que hicieron sus aportes científicos, participaron de las diferentes actividades, nos acompañaron e hicieron posible CICCSI 2020. Mendoza, noviembre de 2020.

Fernando Pinciroli
Susana Dueñas
Javier Rosenstein
María Inés Lund
Carlos Catania
Verónica Miguez
Laura Nidia Aballay
Manuel Ortega
Miguel Méndez-Garabetti
Sandra Oviedo
Elisa Galdame
Marcelo Palma
Line Yazmin Becerra Sánchez
Lina María Suárez Vásquez

ÍNDICE

Artículos Científicos **1**

2 Hybrid Cuckoo Search for solving DNA Fragment Assembly Problem
Matias G. Rojas, Ana C. Olivera, Pablo J. Vidal

13 Pruebas de Aleatoriedad de Segundo Nivel para Secuencias Binarias Aleatorias
Andrés Francisco Farías, Pablo Marcelo García, Germán Antonio Montejano

25 Avoided traffic jams optimising the scheduling of signalised intersection of Mendoza Downtown and surroundings
Pablo Javier, Ana Carolina Olivera

35 Analítica de Negocio en la Gestión de Vuelos Gubernamentales: Una Revisión Sistemática de la Literatura
María Inés Lund, Silvina Migani, Alejandro Riveros, Susana B. Ruiz

47 Ontology of Cardiovascular and Respiratory Diseases
William da Rosa Fröhlich, Sandro José Rigo, Clarissa Almeida Rodrigue, María Eugenia Ruiz

59 Consideraciones Éticas de los Datos en Business Intelligence
María del Carmen Becerra, María Romagnano, Alicia Aballay, Adriana Sarmiento

71 Privacy Preserving Multi Party Computation for Data-Analytics in the IoT-Fog-Cloud Ecosystem
Julio Lopez-Fenner, Samuel Sepúlveda, Luiz Fernando Bittencourt, Fabio Moreira Costa, Nikolaos Georgantas

83 Evaluación heurística de usabilidad utilizando indicadores cualitativos para sistemas detectores de intrusión
Cristian Cappo, Cristian Aceval

95 APPADS. Propuesta de una Aplicación Open Source Interactiva, Reactiva y Dinámica para el Análisis Simbólico de los Datos
Scheffer María del Valle, Herrera Myriam Beatriz, Romagnano María

107 Development of Information Systems for Sexual and Reproductive Health: A systematic review
Leydi Erazo, Erika Gallego, Helder Castrillón Cobo

123 Sistema de información para la gestión de datos del laboratorio de Acuponía Lestoma
Andrés Camilo Tinoco Barreto, Cristian Camilo Medina Albarracín, Alexander Espinosa García, Jaime Eduardo Andrade Ramirez

131 Computational strategies for implementation of 2D elastic wave modeling in GPU
Anderson Paez Chanaga, Ivan Sanchez Galvis, Ana Ramirez Silva

143 Traducción de imagen con redes generativas antagónicas condicionales aplicada a la resolución automática de captchas
Tomás Ortiz, Pablo Sández, Rosana Giménez y Miguel Méndez-Garabetti

155 Reflexiones sobre aprendizaje presencial vs. remoto. un caso de estudio
Laura Nidia Aballay, Maria Inés Lund

169 Implementación de práctica DevOps en aplicación serverless
Federico Sanchez, Ana Laura Molina, Lorena Parra, Nelson Rodríguez, María Murazzo

181 Comparando un Sistema de Bases de Datos de Grafos con un Sistema de Bases de Datos Relacional
Silvina Migani, Cristina Vera

197 A maturity model for the Integrated Agile Transformation Model TM
Fernando Pinciroli

209 Beyond Random Split for Assessing Statistical Model Performance
Carlos Catania, Jorge Guerra, Juan Manuel Romero, Gabriel Caffaratti, Martin Marchetta

221 Estudio de eficiencia de redes de sensores para Inteligencia Ambiental
Adrián Puebla, Rosana Giménez, and Pablo Sández, Miguel Méndez-Garabetti

233 3D to 2D Projections: A Preliminary Study of Dental Segmentation using Watershed Algorithm
J.D. Tamayo-Quintero, J.B. Gómez-Mendoza, S.V. Guevara-Pérez

245 Interoperabilidad de funciones en el Modelo de Programación de Serverless Computing
Nelson Rodríguez, María Murazzo, Hernán Atencio, Martín Gomez, Diego Medel, Lucas Vargas, Cynthia Aguilera

257 Sistema de Detección y Reconocimiento Facial Mediante Algoritmos de Visión Computacional
Emiliano Gomez Blotta, Rosana Giménez, Miguel Méndez-Garabetti, Pablo Sández

263 Biometría como Comprobación de Identidades
Facundo Collado, Giménez Rosana, Sández Pablo Javier, Méndez-Garabetti Miguel

271 Análisis Comparativo de Herramientas de Data Warehouse orientadas a Negocios Inteligentes
Rocha Saturnino, Giménez Rosana, Sández Pablo Javier, Méndez-Garabetti Miguel

Pósters **283**

284 Implementación de Business Intelligence para la Gestión de Información en Organismos Públicos
Ezequiel Moyano, Martín Villarreal, Abraham Jurado

289 Una Aproximación al Uso de Rúbricas en la Evaluación de la Ingeniería de Software
Ezequiel Moyano, Matías Moncho

293 Diseño de Estrategias Didácticas para la Enseñanza de Ingeniaría de Requerimientos
Ezequiel Moyano, Martín Villarreal, Matías Moncho, Marcela Jerez, Lucas Romano

297 Técnicas de ciencias de datos para el análisis de la teoría moderna de optimización de carteras de inversiόm en el marco del mercado argentino de valores
Rodrigo Gonzalez, Carlos A. Catania, Santiago Eguren

301 Inteligencia artificial en diagnósticos médicos
Emmanuel Ricardo Orozco Gazzo

307 Integración de una BCI para rehabilitación de problemas cognitivos en pacientes neurológicos infantiles implementando VRPN como protocolo de comunicaciones
Javier J. Rosenstein, Rodrigo Gonzalez, Nicolás Ignacio Zárate Alvarez, Matías Campos

313 Redes bayesianas para el diagnóstico de COVID-19
Maschke Julián

317 Aplicación que te acerca al Reino Fungi junto al uso de la Realidad Aumentada
Giuliana Pessina

321 Sistema de Detección de Noticias Falsas
Augusto Lombino

325 Aplicación de redes neuronales para la predicción de series temporales de consumo de productos
Di Giorgio Bruno

Lista de autores **329**

ARTÍCULOS CIENTÍFICOS
TRABAJOS COMPELTOS

Hybrid Cuckoo Search for solving DNA Fragment Assembly Problem

Matias G. Rojas[1]0000-0003-3881-0888, Jessica A. Carballido[2,4]0000-0001-6284-1049, Ana C. Olivera[1,3,4]0000-0001-7825-1959, and Pablo J. Vidal[1,3,4]0000-0001-6502-8010

[1] Instituto para las Tecnologías de la Información y las Comunicaciones, Universidad Nacional de Cuyo, Mendoza, Argentina rojasmatias994@gmail.com
[2] Instituto de Ciencias e Ingeniería de la Computación, *Universidad Nacional del Sur*, Bahía Blanca, Argentina {jac}@cs.uns.edu.ar
[3] Universidad Nacional de Cuyo, Facultad de Ingeniería, Mendoza, Argentina
[4] Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
{pjvidal,acolivera}@conicet.gov.ar

Abstract. The Deoxyribonucleic Acid Fragment Assembly Problem represents an interesting research topic in the field of bioinformatics. This problem consists in reconstructing a DNA chain as close to the original sequence from a set of DNA fragments taken randomly. Several authors have proposed different approaches to solve Deoxyribonucleic Acid Fragment Assembly Problem. In particular, nature-inspired algorithms has been used for its resolution. Among these techniques, the Cuckoo Search is a recent method based on the breeding of the cuckoo birds. In this paper, a discrete Cuckoo Search algorithm is adapted to DNA-FAP hybridising with the behaviour of two different local searches to improve the search capabilities of the canonical algorithm. A numerical evaluation of both proposals and the canonical version is carried out using a well-known benchmark dataset. Results demonstrate that the hybridisation improves the search and transform the cuckoo search in a robust procedure with the potential to deal with longer sequences or sequences of unknown length as well.

Keywords: Metaheuristic, DNA Fragment Assembly Problem, Combinatorial Optimization, BioInformatics, Bio-inspired algorithm.

1 Introduction

The DNA-FAP (Deoxyribonucleic Acid Fragment Assembly Problem) is crucial for the successful of a genome project as the remaining phases strongly depend on the accuracy of the results at this stage. The DNA-FAP is a NP-Hard combinatorial optimisation problem [1].

Due to its difficulty, a lot of tools have been proposed to automate DNA sequencing. Among these tools, PHRAP [2], TIGR assembler [3], STROLL [4], CAP3 [5], Celera assembler [6] and EULER [1] may be cited. Each one automates

fragment assembly using a variety of algorithms, being the most widely used those based on greedy techniques.

Metaheuristic algorithms are known as powerful techniques to solve different complex problems from various fields. In particular, bio-inspired algorithms have proven to obtain promising results in the DNA strand assembly section [7–9]. In the same way, to increase the search capacity of these techniques, different hybridisations have been performed. Hybridisation allows us to combine two or more methods to improve the exploration and exploitation capacity of the original procedure.

The Cuckoo Search Algorithm (CSA) is a bio-inspired algorithm that comes from the breeding method of the cuckoo birds developed by Yang and Deb [10]. In recent times, it attracted significant attention due to its ability to solve out different kind of real-life problems such as biological modelling [11], electronic distribution problems [12], among others.

In this work, we propose a Hybrid Cuckoo Search Algorithm (HCSA) for solving the DNA fragment assembly problem. The achieved performance when Canonical CSA is hybridised with both Problem Aware Local Search (PALS) and Simulated Annealing (SA) is evaluated. The different approaches are tested using eight popular benchmarks [13].

The contributions of the paper can be summarised in the following:

- The definition and study of the structure of a novel Hybrid Cuckoo Search (HCSA).
- Analysing the behaviour of HCSA to solve the DNA assembly problem with different state-of-the-art instance sizes.

In the remainder of this paper, we describe the Fragment Assembly Problem in Section 2. After that, Section 3 introduces the canonical DFA and the details of our algorithmic proposal. Section 5 describes the experimental settings, including a brief explanation about DNA-FAP instances. Then, an analysis of the results is presented. Finally, Section 6 provides the conclusions and also highlights future research directions.

2 The DNA Fragment Assembly Problem

In a DNA fragment, there are four types of nucleotides which are represented by a symbol or character. They are Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). Structure of DNA contains double-helix of nucleotides that the number of Adenine (A) equals to the number of Thymine (T), and also the number of Guanine (G) equals to the number of Cytosines (C). Due to the number of naturally occurred repetitive nucleotides and partial data, assembling the DNA sequence is a challenging process. The assembling process starts with broken large DNA strands into a small number of fragments called *shotgun* sequencing [14]. Then, it merges the overlapped fragments to generate common DNA sequence, which is known as *contig*.

The steps are defined as Overlap, Layout and Consensus: (OLC) [15]

- **Overlap**: Finding potentially overlapping among fragments.
- **Layout**: Finding the order of the fragments based of similarity scores.
- **Consensus**: Deriving the DNA sequence from the layout.

The overlap problem consists in finding the best match between the suffix of one read and the prefix of another one. The layout is the hardest step [8]. The naturally repeated sequences are the most difficulty when deciding whether two fragments overlap (sequencing errors cause their differences), or they actually come from two different copies of a repeat.

In order to evaluate a solution for DNA-FAP, the follow function can be defined: for a possible order of l fragments (a permutation) $i = [0, \ldots, a, a+1, \ldots, l]$ the Equation 1 shows the value of the sequence i for DNA-FAP.

$$f(i) = \sum_{a=0}^{l-1} w_{a,a+1} \tag{1}$$

where $w_{a,a+1}$ is the pairwise overlap strength of fragments a and $a+1$ [16].

The consensus step of fragment assembly amounts to correcting errors in sequence reads. The coverage distribution is used to measure the quality of the consensus. Coverage at a base position is calculated as the number of elements at that position. It is a metric of the redundancy, and it denotes the average number of pieces where a given nucleotide in the target DNA is expected to appear. Coverage calculus consider the number of bases read from fragments over the target DNA's length [9].

$$Coverage = \frac{\sum_{i=1}^{l} \text{length of the fragment } i}{\text{target sequence length}}, \tag{2}$$

where l stands for the number of fragments, thus, the higher the coverage and the smaller the number of gaps better the results. Partial coverage is achieved provoked when cannot assemble a given set of fragments into a single contig. More precisely, a contig in biology is defined as a layout consisting of contiguous overlapping fragments. Overlapping between adjacent pieces must be greater than or equal to a predefined threshold (i.e., cut-off parameter).

The assembly of DNA fragments into a consensus sequence corresponding to the parent sequence constitutes the DNA-FAP, which is a NP-hard problem [1].

In recent years, many works related to DNA Fragment Assembly Problem have been performed. In [17], a Discrete Particle Swarm Optimisation (DPSO) was presented. It proposes a novel way to deal with DNA Fragment Assembly problems without using the Smaller Position Value (SPV) approach. The propose has reached a better performance compared with state-of-art algorithms.

There are works in that hybridisation with Problem Aware Local Search (PALS) was presented. PALS It is a technique that has achieved excellent results with many types of instances. In, a Restarting and Recentering Genetic Algorithm (RRGA) hybridised with PALS has been proposed. In [18], a Crow Search Algorithm is hybridised with PALS. At both cases, great improvements were achieved.

Finally, there are also some works in that Cuckoo Search are applied to DNA Fragment Assembly Problem. An example is a work presented in [19], where a canonical Cuckoo Search Algorithm (CSA) was used.

For our knowledge, there is not any work presenting a Hybrid CSA approach for solving DNA Fragment Assembly Problem.

3 Hybrid Cuckoo Search

In this section, the Hybrid Cuckoo Search Algorithm (HCSA) is presented. Firstly, its canonical version is introduced, and then the hybridisation with both the Problem-Aware Local Search (PALS) and the Simulated Annealing (SA) algorithm, are explained.

3.1 Canonical Cuckoo Search

The Cuckoo Search Algorithm (CSA) [10] is one of the recently emerged optimization algorithms. It is a population-based algorithm that comes from breeding behaviour of cuckoo birds.

In nature cuckoos eliminate eggs in nests from other birds and lay their ones, increasing the survival probabilities of their eggs. The host bird of the nest takes care of cuckoo eggs until it realizes and destroy the alien eggs or it leaves the nest.

In the algorithm, every nest has an egg that represents the current solutions, and cuckoos eggs are the new solutions. A cuckoo can lay an egg at a time. Nests are selected by use Levy flight because it allows simulating birds flight. If the cuckoo egg is better than the egg of the chosen nest, it will be laid in the nest; otherwise, the cuckoo egg is discarded. At the other hand, the host birds can leave the nest with a probability p, and move to another location generating a new solution.

3.2 Hybridisation

A property of the optimisation algorithms based on the evolution of populations is that they are good exploring the search space, but they have problems exploiting it. That is, these algorithms can discover scattered areas of the search space efficiently, but they do not consider solutions near to those they found previously.

Algorithm 1 Hybrid Cuckoo Search Algorithm

```
nests ← initialise_nests(number_of_nests)
while not stop_criterion do
    cuckoos ← generate_new_cuckoos(number_of_nests)
    for cuckoo in cuckoos do
        pos ← get_random_position(nests)
        if cuckoo.fitness > nest[pos].fitness then
            nest[pos] ← local_search(cuckoo)
        end
    end
    nests ← abandon(nests)
end
Return best_solution
```

To aboard this problem, the typical approach is to use hybrid algorithms (also known as memetic algorithm). This approach combines both Population-Based Algorithms and Local Search Algorithms, where the former brings the exploration capabilities and the latter provides the exploitation [20].

Algorithm 1 shows the performed hybridisation steps. Firstly, the algorithm initialises nests randomly. So, until the stop criterion has been satisfied, new cuckoos are generated regarding the existing nests. It is performed regarding the levy flight value. When the levy flight function gives a value bigger than 2, Double Bridge Method is used. Otherwise, Two-Opt Method is used. Then, If obtained cuckoo is fitter than a randomly selected nest, the cuckoo lays its egg after enhancing it by the local search method. Finally, each nest has a probability of being abandoned. If it occurs, a new random nest will be generated to replace the abandoned nest.

In this work, we chose two well-known techniques for hybridising the CSA. First, we use a Simulated Annealing for its efficiency in discrete search space. Second, Problem Aware Local Search (PALS) for being a fairly successful technique for solving the DNA-FAP problem.

- Simulated Annealing (SA): Simulated annealing arises from an analogy between the annealing of the solids and optimisation problems. The algorithm begins generating an initial solution. Then, perturbations are applied to it to identify its neighbours. A generated neighbour will be kept if it is fitter than the current solution. Otherwise, it will be kept considering the Boltzmann probability [21].
- Problem Aware Local Search (PALS): is an algorithm designed for DNA fragment assembly problem. PALS improves a single solution iteratively. At every step, the current solution is replaced by a better solution that belongs to its neighbourhood. Candidate solutions are generated by interchange elements of the current solution, to reduce the number of contigs as long as the overlap score stays the same or increases. The interchange of elements that reports betters results is going to replace the current solution [22].

Table 1. Information of datasets

Dataset Acronym	Coverage	Mean of fragment length	Number fragments	Original sequence length	Optimum
x60189_4 (x_4)	4	395	39	3835	11478
x60189_5 (x_5)	5	386	48		14161
x60189_6 (x_6)	6	343	66		18301
x60189_7 (x_7)	7	387	68		21271
m154216_5 (m_5)	5	398	127	10089	38746
m154216_6 (m_6)	6	350	173		48052
m154216_7 (m_7)	7	383	177		55171
j02459_7 (j_7)	7	405	352	20000	116700
bx842596_4 (b_4)	4	708	442	77292	227920

4 Experiments

This section presents the details related to the experiments that have been made.

We carried out several experiments with different instances of DNA-FAP benchmark data sets, which were described in [13].

Table 1 summarizes the instances information with their names, coverage, sizes (*Number of fragments* by instance), original sequence length and the *optimum*, respectively. The instances are ordered by group and number of fragments. A more detailed description of these benchmarks can be found in [13].

We perform 30 independent runs to evaluate the instances. Stop condition for all the algorithms is defined in one hundred of evaluations. Table 2 presents the settings for every variation of the CSA. First, the Canonical CSA is described because it has common parameters to the other variations. Then, the hybridisation adds their specific parameters.

Table 2. Settings for each variation of CSA

Algorithm	Parameter	Value
CSA	Nests Amount	20
	Mutation Operator	Two-Opt/Double Bridge
	Abandon Nest Probability	0.01
CSA+PALS	Local Search Algorithm	PALS
	Local Search Iterations	10
CSA+SA	Local Search Algorithm	SA
	Local Search Iterations	10

Experiments were performed in the Toko cluster (`toko.uncu.edu.ar`) with an AMD Opteron/Epyc processor (64 cores and 128 GB of RAM). The operating system is Ubuntu 18.04 LTS. Algorithms are implemented in jmetalpy [23].

5 Experimental Results

This section presents the experimental results obtained by the proposed Cuckoo algorithm for solving different DNA-FAP instances. We measure the quality of the solution by considering the fitness value achieved.

Table 3. Information of datasets. Accession numbers are used as the name of the instances

Instance	CSA		CSA+PALS		CSA+SA	
	Hit-rate	Avg.	Hit-rate	Avg.	Hit-rate	Avg.
x_4	0	10948.83	**6**	**11350.97**	0	11023.93
x_5	0	13640.10	**8**	**13986.43**	0	13680.63
x_6	0	17674.60	**13**	**18062.07**	0	17240.13
x_7	0	20597.57	0	**20971.43**	0	20170.23
m_5	0	36061.53	0	**38618.30**	0	32801.53
m_6	0	40917.67	**6**	**47974.27**	0	35273.93
m_7	0	47294.77	**1**	**54926.33**	0	40735.37
j_7	0	82453.77	0	**102963.63**	0	64075.57
b_4	0	143290.27	0	**209575.60**	0	104439.67

Table 3 presents the fitness value results for all the used DNA-FAP instances. The first column presents the abbreviation of each instance. Next, Hit-Rate and average fitness over the thirty runs are showed for each algorithm tested. The results in Table 3 show a marked improvement of the performance by introducing PALS as a method to enhance existing solutions in Canonical CSA. CSA+PALS has reached the optimum or is very near to it. In many instances. CSA+PALS has been the only one to achieve optimal value several times. For the largest instances, CSA+PALS has been the only algorithm that has come closest (highest value). The hybridisation of CSA+SA obtains worse results. This result may be related to SA cannot find promising areas of the search space due to the reduced number of evaluation that it performs as a local search method. The improvement is lowest and, at the same time, it consumes evaluations of the CSA.

Figures 1, 2 and 3 present the evolution of the fitness quality for the different proposal of CSA with the three instance: x_4, m_7 and b_4, respectively. In Figure 1, algorithms have similar behaviour, standing out for a slight difference the hybridisation between CSA and PALS. Nevertheless, the difference is noticeable for m_7 and b_4 dataset, as shown in Figure 2 and Figure 3. There, CSA+PALS has obtained significant improvements quite faster than the other approaches. In particular, the SA hybridisation does not improve the results with the evolution of the algorithm.

Now we offer an analysis from the accuracy (quality) between CSA+PALS (that has reached the best results) and different approaches from the literature. Table 4 presents the performance of CSA+PALS algorithm compared against other algorithms used for computing the fitness value. These algorithms are

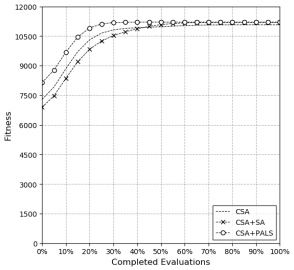


Fig. 1. Evolution of fitness for instance x_4.

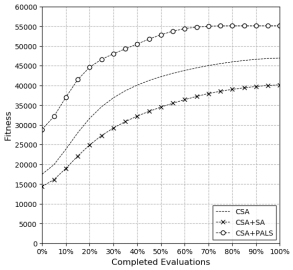


Fig. 2. Evolution of fitness instance m_7.

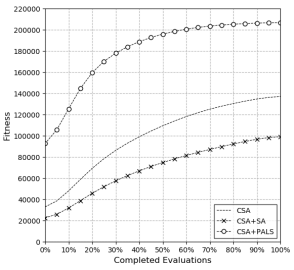


Fig. 3. Evolution of fitness for instance b_4.

Particle Swarm Optimisation with Differential Evolution (PPSO+DE), Queen Bee Evolution Based on Genetic Algorithm (QEGA) [24], Simulated Annealing (SA) [25], Problem Aware Local Search (PALS), SAX [26] and Prototype Optimisation with Evolved Improvement Steps (POEMS) [7].

From Table 4, we can observe that CSA+PALS obtain in general competitive results in all the instances. We can see that our proposal has reached the best solution and the optimum in six of nine cases. With small benchmarks, the quality of the obtained solution for all the algorithms is the same. For medium instances (as m_5, m_6 and m_7) our approach reaches the best solution, and for the largest instances CSA+PALS is below the rest of the algorithms compared.

Table 4. Best results for the 16 benchmark datasets. The "-" symbol indicates that this information is not provided by the corresponding paper.

Instance	CSA +PALS	PPSO +DE	QEGA	SA	PALS	SAX	POEMS
x_4	**11478**	**11478**	11476	**11478**	**11478**	**11478**	**11478**
x_5	**14131**	13642	14027	14027	14021	14027	–
x_6	**18301**	**18301**	18266	**18301**	**18301**	**18301**	–
x_7	21268	20921	21208	**21271**	21210	21268	21261
m_5	**38745**	38686	38578	38583	38526	38726	38610
m_6	**48052**	47669	47882	48048	48048	48048	–
m_7	**55171**	54891	55020	55048	55067	55072	55092
j_7	104474	114381	116222	116257	115320	115301	**116542**
b_4	213422	224797	**227252**	226538	225783	223029	227233

In general terms, the hybridisation of CSA with PALS has allowed the optimum values to be reached in different instances. For those more complex, CSA+PALS has come very near to the optimum value. The collaboration between CSA and PALS allows the efficient exploration and exploitation of the search space.

6 Conclusions

We have proposed a Hybrid Cuckoo Search Algorithm (HCSA) to assemble DNA fragments to reconstruct a genome sequence. We performed the tests over a group of well-known DNA-FAP benchmarks and compared with other assemblers. Promising results were obtained considering the optimal values reached by CSA+PALS. From the analysis of all the results, we can observe that our proposal can achieve the optimum or is very close to it. The preliminary results of our experiments are very promising regarding the effectiveness of the method that we have developed. Indeed, CSA+PALS can be considered a satisfactory assembler, given its high accuracy and its ability to achieve optimal values for DNA fragment assembly problem. In general, CSA+PALS provides the right balance of exploitation and exploration. This is an interesting idea to apply this algorithm for big database of genomes. Besides, we have found that the hybrid model provides a robust model that allows solving instances of different sizes without significant degradation in the quality of the solutions.

As future works, it is interesting to implement a perturbation method that includes characteristics of both PALS and other techniques based in permutation, to increase the speed in that search space is explored. We also plan to contemplate other factors in the analysis of the quality of the solutions, such as amounts of contigs, more evaluations, among others, to improve the accuracy of our proposals.

Acknowledgement

The authors thank Universidad Nacional del Sur (UNS) for grant PGI 24/N042 and be grateful for Universidad Nacional de Cuyo for grant B081.

References

1. P. Pevzner, *Computational Molecular Biology: An Algorithmic Approach.* MIT, 2000.
2. P. Green, "Phrap, version 1.090518, http://phrap.org," 2009.
3. G. G. Sutton, O. White, M. D. Adams, and A. R. Kerlavage, "Tigr assembler: A new tool for assembling large shotgun sequencing projects," *Genome Science and Technology*, vol. 1, no. 1, pp. 9–19, 1995.
4. T. Chen and S. S. Skiena, "A case study in genome-level fragment assembly," *Bioinformatics*, vol. 16, 2000.
5. X. Huang and A. Madan, "Cap3: A DNA sequence assembly program," *Genome research*, vol. 9, no. 9, pp. 868–877, 1999.
6. E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington *et al.*, "A whole-genome assembly of drosophila," *Science*, vol. 287, no. 5461, pp. 2196–2204, 2000.
7. J. Kubalik, P. Buryan, and L. Wagner, "Solving the DNA fragment assembly problem efficiently using iterative optimization with evolved hypermutations," in *Proceedings of GECCO-2010.* ACM, 2010, pp. 213–214.

8. G. Minetti and E. Alba, "Metaheuristic assemblers of DNA strands: Noiseless and noisy cases," in *Proceedings of the IEEE Congress on Evolutionary Computation, CEC2010, Barcelona, Spain, 18-23 July 2010*, 2010, pp. 1–8.
9. P. J. Vidal and A. C. Olivera, "Solving the DNA fragment assembly problem with a parallel discrete firefly algorithm implemented on GPU," *Computer Science and Information Systems*, vol. 15, no. 2, pp. 273–293, 2018.
10. X.-S. Yang and S. Deb, "Cuckoo search via lévy flights," in *2009 World congress on nature & biologically inspired computing (NaBIC)*. IEEE, 2009, pp. 210–214.
11. Y. Chen and N. Wang, "Cuckoo search algorithm with explosion operator for modeling proton exchange membrane fuel cells," *International Journal of Hydrogen Energy*, vol. 44, no. 5, pp. 3075–3087, 2019.
12. T. T. Nguyen and T. T. Nguyen, "An improved cuckoo search algorithm for the problem of electric distribution network reconfiguration," *Applied Soft Computing*, vol. 84, p. 105720, 2019.
13. G. M. Mallén-Fullerton, J. A. Hughes, S. Houghten, and G. Fernández-Anaya, "Benchmark datasets for the DNA fragment assembly problem," *International Journal of Bio-Inspired Computation*, vol. 5(6), pp. 384–394, 2013.
14. M. Pop, "Shotgun sequence assembly," *Adv. Comput.*, vol. 60, pp. 193–248, 2004.
15. N. C. Jones and P. A. P. Preface, *An Introduction to Bioinformatics Algorithms*. Massachusetts Institute of Technology, 2004.
16. R. Parsons, S. Forrest, and C. Burks, "Genetic algorithms, operators, and DNA fragment assembly," *Machine Learning*, vol. 21, no. 1-2, pp. 11–33, 1995.
17. A. B. Ali, G. Luque, and E. Alba, "An efficient discrete PSO coupled with a fast local search heuristic for the DNA fragment assembly problem," *Information Sciences*, vol. 512, pp. 880–908, 2020.
18. M. Allaoui, B. Ahiod, and M. El Yafrani, "A hybrid crow search algorithm for solving the DNA fragment assembly problem," *Expert Systems with Applications*, vol. 102, pp. 44–56, 2018.
19. R. Indumathy, S. U. Maheswari, and G. Subashini, "Nature-inspired novel cuckoo search algorithm for genome sequence assembly," *Sadhana*, vol. 40, no. 1, pp. 1–14, 2015.
20. P. Moscato and C. Cotta, "An accelerated introduction to memetic algorithms," in *Handbook of Metaheuristics*. Springer, 2018, pp. 275–309.
21. D. Delahaye, S. Chaimatanan, and M. Mongeau, "Simulated annealing: From basics to applications," in *Handbook of Metaheuristics*. Springer, 2019, pp. 1–35.
22. A. B. Ali, G. Luque, E. Alba, and K. E. Melkemi, "An improved problem aware local search algorithm for the DNA fragment assembly problem," *Soft Computing*, vol. 21, no. 7, pp. 1709–1720, 2017.
23. A. Benitez-Hidalgo, A. J. Nebro, J. Garcia-Nieto, I. Oregi, and J. Del Ser, "jmetalpy: a python framework for multi-objective optimization with metaheuristics," *arXiv:1903.02915*, 2019.
24. D. E. Knuth, *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*. Redwood City, CA, USA: Addison Wesley, 1998.
25. J. S. Firoz, M. S. Rahman, and T. K. Saha, "Bee algorithms for solving DNA fragment assembly problem with noisy and noiseless data," in *Proceedings of the GECCO'14*. New York, NY, USA: ACM, 2012, pp. 201–208.
26. G. Minetti, G. Leguizamón, and E. Alba, "Sax: a new and efficient assembler for solving DNA fragment assembly problem," in *2012 Argentine Symposium on Artificial Intelligence*, 2012.

Pruebas de Aleatoriedad de Segundo Nivel para Secuencias Binarias Aleatorias

Andrés Francisco Farías[1] - Pablo Marcelo García[2] - Germán Antonio Montejano[3]

[1] DACEFyN – UNLaR (afarias665@yahoo.com.ar)
[2] FCEyN – UNLPam (pablogarcia@exactas.unlpam.edu.ar)
[3] FCFMyN – UNSL (gmonte@unsl.edu.ar)

Resumen. Tradicionalmente, la aplicación de pruebas de aleatoriedad sobre una secuencia binaria aleatoria es de carácter integral, es decir que se verifica la totalidad de la misma. Pero puede suceder que alguna subsecuencia no sea auténticamente aleatoria; ello lleva a plantear la necesidad de tener en cuenta esta situación. Este trabajo considera esta posibilidad y analiza subsecuencias de una secuencia binaria aleatoria principal, mediante Pruebas de Segundo Nivel.

1 Introducción

Para comprobar la aleatoriedad de una secuencia binaria, se realizan pruebas estadísticas, sobre la secuencia completa, las que llamaremos pruebas estadísticas de aleatoriedad de Primer Nivel.

$$\overbrace{0,1,1\,,1,0,1,1,0,0,0,\ldots,1,0,1,0,1}^{secuencia\ \ binaria\ \ aleatoria}$$

Generalmente utilizamos una parte de la secuencia completa y puede suceder que esa fracción no tenga un buen comportamiento aleatorio o el valor p está próximo a 0 o al valor 1, lo que hace sospechar de la calidad aleatoria de la cadena binaria. Por lo que es conveniente realizar pruebas adicionales sobre las subsecuencias, las que llamamos pruebas estadísticas de aleatoriedad de Segundo Nivel.

Para realizar este estudio, se divide la secuencia en varias subsecuencias a las que se las somete a repetidas pruebas de Primer Nivel [1] [2],[3] y [4].

$$\overbrace{0,1,1,1,0,}^{subsecuencia}\ \overbrace{1,1,0,0,0,}^{subsecuencia}\ldots,\ \overbrace{1,0,1,0,1}^{subsecuencia}$$

Luego se practican, sobre las resultados obtenidos de las evaluaciones de las subsecuencias, Pruebas de Bondad de Ajuste: Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises, para determinar la uniformidad de los valores p [5] y [6].

2 Pruebas de Primer Nivel

Las pruebas estadísticas de Primer Nivel son aquellas que se realizan sobre la secuen-

cia binaria completa; se puede utilizar pruebas individuales, grupos o paquetes completos. Para este trabajo se recurre a un conjunto de siete pruebas estadísticas de aleatoriedad indicadas en [7], de las que hablará en detalle más adelante. En la figura 1 se muestra el proceso de aplicación de tests de primer nivel.

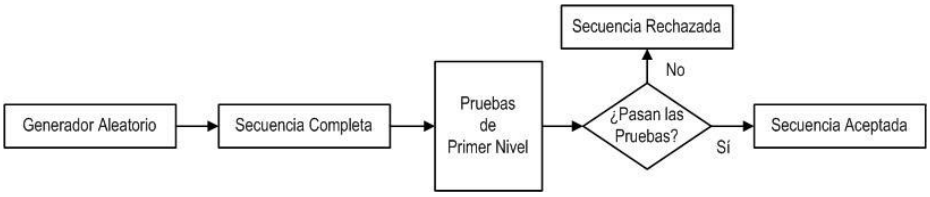


Figura 1. Esquema de las Pruebas de Primer Nivel

3 Pruebas de Segundo Nivel

Como se indicó en el punto anterior, la secuencia principal se divide en subsecuencias iguales, adoptando para este trabajo diez, a las que se les realizarán algunas pruebas estadísticas utilizadas en el Primer Nivel.. El esquema del procedimiento se indica en la figura 2.

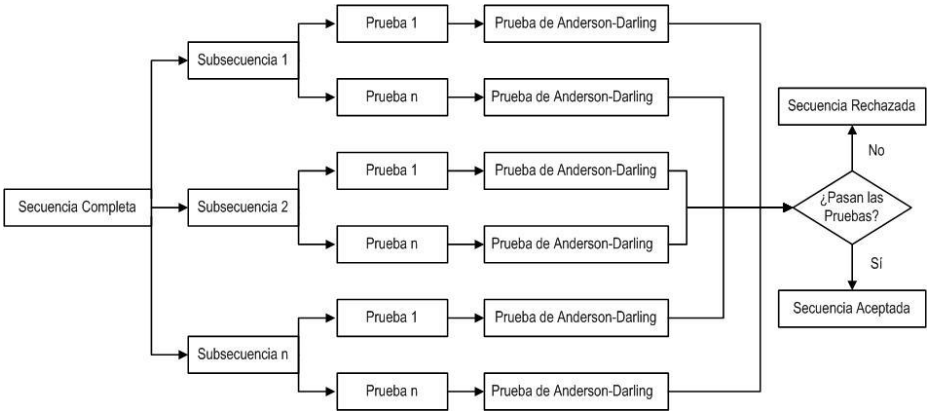


Figura 2. Esquema de las Pruebas de Segundo Nivel

Los resultados de las pruebas de las subsecuencias serán sometidos a Prueba de Bondad de Ajuste, se adopta para este trabajo la Prueba de Anderson-Darling, con alguna de las siguientes distribuciones: Normal, Exponencial, Gamma, Lognormal y Weibull [5] y [6].

4 Prueba de Bondad de Ajuste Anderson-Darling

Esta prueba es una alternativa de la prueba Kolmogorov-Smirnov, y permite detectar

mejor, las diferencias entre las distribuciones acumuladas empírica y teórica.

El procedimiento es el siguiente, con la muestra de n datos se estiman los parámetros de la distribución propuesta, y se realizan los dos ordenamientos de datos siguientes: los datos de menor a mayor: x_i siendo: $i = 1,2,3,\dots,n$ y los datos de mayor a menor: x_{n+1-i}

Luego, se calcula la Probabilidad Acumulada Esperada PAE_{x_i} para cada dato x_i, y la Probabilidad Acumulada Esperada $PAE_{x_{n+1-i}}$, para cada dato x_{n+1-i} .

Posteriormente, se calcula el estadístico de la prueba A^2, el que se ajusta conforme a la distribución de probabilidad teórica propuesta y finalmente se compara con el valor crítico. Si el estadístico de prueba es menor que el valor crítico del punto anterior, se acepta la hipótesis nula, en caso contrario se rechaza.

4.1 Anderson-Darling con Distribución LogNormal

Para una Distribución LogNormal, los momentos muestrales calculados a partir de los datos x_i son:

$$m_1^{'} = \frac{1}{n} \cdot \sum_{i=0}^{n} Ln\, x_i \qquad m_2^{'} = \frac{1}{n} \cdot \sum_{i=0}^{n} (Ln\, x_i)^2 \tag{1}$$

La media y la desviación estándar serían:

$$\mu = m_1^{'} \qquad \sigma = \sqrt{(m_2^{'} - (m_1^{'})^2)} \tag{2}$$

Estadístico Anderson-Darling

$$A^2 = -\sum [(2i-1)\,\{ln\, F_x(x_i) + ln\,[1 - F_x(x_{n+1-i})]\}/n\,] - n \tag{3}$$

Estadístico de Anderson-Darling ajustado:

$$A^* = A^2(1 + \frac{3}{4n} + \frac{9}{4n^2}) \tag{4}$$

4.2 Anderson-Darling con Distribución Normal

Para una Distribución Normal, los momentos muestrales calculados a partir de los datos x_i son:

$$m_1^{'} = \frac{1}{n} \cdot \sum_{i=0}^{n} x_i \qquad m_2^{'} = \frac{1}{n} \cdot \sum_{i=0}^{n} (x_i)^2 \tag{5}$$

La media y la desviación estándar serían:

$$\mu = m_1^{'} \qquad \sigma = \sqrt{(m_2^{'} - (m_1^{'})^2)} \tag{6}$$

Estadístico Anderson-Darling

$$A^2 = -\sum [(2i-1)\,\{ln\, F_x(x_i) + ln\,[1 - F_x(x_{n+1-i})]\}/n\,] - n \tag{7}$$

Estadístico de Anderson-Darling ajustado:

$$A^* = A^2(1 + \frac{3}{4n} + \frac{9}{4n^2}) \tag{8}$$

5 Generador Binario Pseudoaleatorio

Para la realización de la Pruebas de Primer y Segundo Nivel, utilizamos un Generador Binario Pseudoaleatorio [8] , que aparece en la figura 5.

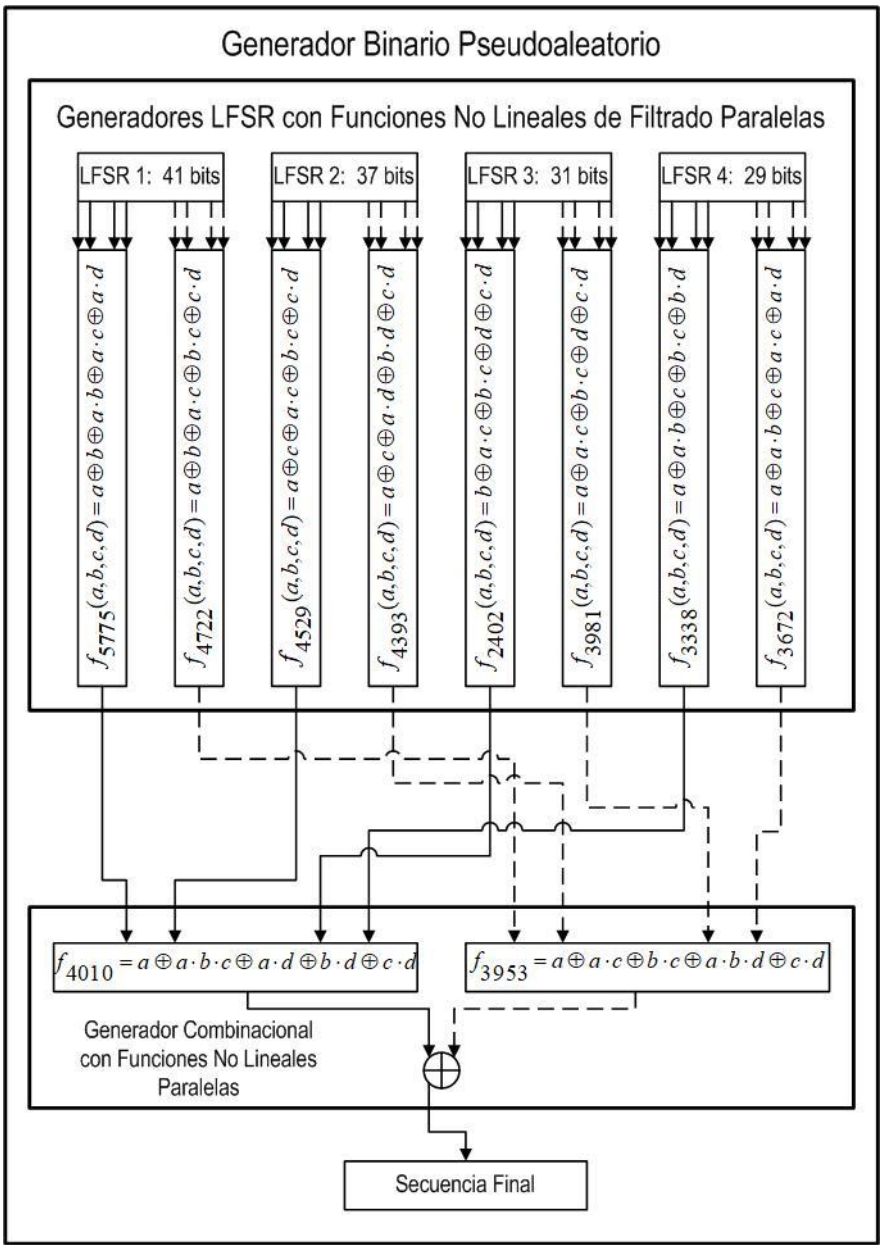


Figura 3. Generador Binario Pseudoaleatorio

6 Realización de las Pruebas de Primer Nivel

6.1 Pruebas Estadísticas de Aleatoriedad

Se optó por un grupo de siete pruebas estadísticas para secuencias binarias, del conjunto de pruebas de la Norma NIST Special Publication 800-22, vistas en el trabajo de Rukhin (et al.) [7]:

Tabla 1. Pruebas Estadísticas

Prueba	Denominación
1	Frecuencias
2	Frecuencias dentro de un bloque
3	Rachas
4	La más larga rachas de unos en un bloque
5	Sumas acumulativas adelante
6	Sumas acumulativas atrás
7	Entropía estimada

6.2 Puesta en Funcionamiento del Generador Binario

El paso siguiente es poner en funcionamiento el dispositivo con 100 claves distintas, generando 100 secuencias de 100.000 bits y a continuación verificar la condición de aleatoriedad de las mismas, con un nivel de significancia: $\alpha = 0{,}01$, los valores p obtenidos deben satisfacer la hipótesis nula:

$$H_0 \rightarrow p_valor > 0{,}01 \tag{9}$$

Debido al gran volumen de procesamiento requerido, se buscó el auxilio de un programa escrito en lenguaje C++, que contenía los algoritmos correspondientes al generador y a las pruebas estadísticas. Es decir que el software obtuvo las secuencias binarias y simultáneamente realizó las pruebas sobre las mismas.

6.3 Interpretación de los resultados

Para interpretar los resultados de las Pruebas de Primer Nivel, se consideran las proporciones de muestras que superan las pruebas individuales y se compara en dos límites: superior e inferior [7], que se determina con la expresión (10):

$$LS, LI = (1 - \alpha) \pm 3 \cdot \sqrt{\alpha \cdot (1 - \alpha/k)} \tag{10}$$

No habiendo valores que queden fuera los de límites, se dan por superadas las pruebas, y se pasa al próximo nivel.

Para el caso de las Pruebas de Segundo Nivel, se repite el mismo procedimiento, utilizando la expresión (10), teniendo en cuenta que en este caso el número de muestras puede ser distinto.Sí las proporciones de muestras que superan las pruebas, caen

dentro de los límites, se acepta el generador.

6.4 Proporción de Secuencias que Pasan las Pruebas

En nuestro caso k = 100 y el nivel de significancia elegido es: α = 0,01, los límites quedan de acuerdo con las ecuación 11 y 12.

$$LS = (1 - 0{,}01) \pm 3 \cdot \sqrt{0{,}01 \cdot (1 - 0{,}01)/100} = 1{,}02 \quad (11)$$

$$LI = (1 - 0{,}01) \pm 3 \cdot \sqrt{0{,}01 \cdot (1 - 0{,}01)/100} = 0{,}96 \quad (12)$$

Se consideran todas pruebas, los resultados se indican en la tabla 2.

Tabla 2. Pruebas Estadísticas

Prueba	Pasan	Proporción	Superior	Inferior	Aceptado
1	99	0,99	1,02	0,96	Sí
2	98	0,98	1,02	0,96	Sí
3	99	0,99	1,02	0,96	Sí
4	97	0,97	1,02	0,96	Sí
5	100	1,00	1,02	0,96	Sí
6	100	1,00	1,02	0,96	Sí
7	100	1,00	1,02	0,96	Sí

En la figura 4 se aprecia mejor el resultado:

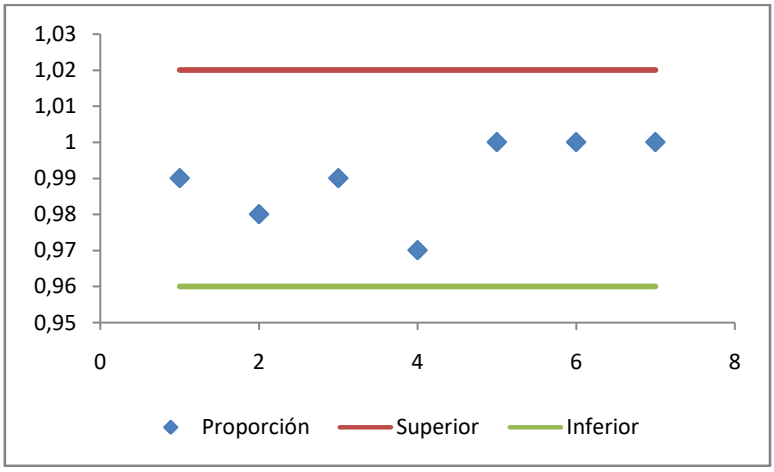


Figura 4. Proporción de muestras que superan las Pruebas

Las secuencias producidas por el generador, son aceptadas porque las proporciones de las muestras que pasan las pruebas están dentro de los límites.

7 Realización de las Pruebas de Segundo Nivel

7.1 Puesta en funcionamiento del Generador

Dividimos la secuencia 100.000 bits en 10 subsecuencias de 10.000 bits a la que le realizamos las Pruebas Segundo Nivel, se optó por un grupo de tres pruebas estadísticas para secuencias binarias, del conjunto de pruebas de la Norma NIST Special Publication 800-22, vistas en el trabajo de Rukhin (et al.) [7]:

Tabla 3. Pruebas Estadísticas

Prueba	Denominación
1	Frecuencias
2	Frecuencias dentro de un bloque
3	Rachas

A la que posteriormenete, en conjunto sometemos a la Prueba de Bondad de Ajuste de Anderson Darling con distribución Normal.

7.2 Prueba de Frecuencias sobre Subsecuencias

Para una sola muestra tenemos, el siguiente resultado de la prueba de Frecuencias:

Tabla 4. Anderson-Darling

Sub.	$Datos$	x_i	x_{n+1-i}	PAE_{x_i}	$PAE_{x_{n+1-i}}$	$2i-1$	B_i
1	0,0549	0,0549	0,8729	0,1052	0,9712	1	-5,8002
2	0,2983	0,0784	0,7188	0,1226	0,9041	3	-13,3304
3	0,7188	0,1141	0,5619	0,1528	0,7583	5	-16,4931
4	0,5619	0,2983	0,4354	0,3765	0,5845	7	-12,9852
5	0,4354	0,3077	0,3576	0,3903	0,4656	9	-14,1058
6	0,8729	0,3576	0,3077	0,4656	0,3903	11	-13,8527
7	0,1141	0,4354	0,2983	0,5845	0,3765	13	-13,1229
8	0,3576	0,5619	0,1141	0,7583	0,1528	15	-6,6376
9	0,0784	0,7188	0,0784	0,9041	0,1226	17	-3,9371
10	0,3077	0,8729	0,0549	0,9712	0,1052	19	-2,6659
							-10,2931

Estadístico de Anderson-Darling:

$$A^2 = -1(C - n) = 0{,}2931 \tag{13}$$

Estadístico de Anderson-Darling ajustado:

$$A^* = A^2\left(1 + \frac{3}{4 \cdot 10} + \frac{9}{4 \cdot 10^2}\right) = 0{,}3217 \tag{14}$$

Comparación con el valor crítico:

$$A^* = 0{,}3217 < A^*_{Crítico} = 1{,}029 \quad (15)$$

Realizado el estudio para veinte muestras nos queda:

Tabla 5. Análisis de muestras

Muestra	$A^*_{calculado}$	Muestra	$A^*_{calculado}$	Muestra	$A^*_{calculado}$
1	0,3217	8	0,4526	15	0,3307
2	0,3202	9	0,2958	16	0,9265
3	0,7779	10	0,6155	17	0,9833
4	0,4346	11	1,0188	18	0,2988
5	0,5614	12	0,4410	19	0,3279
6	0,6534	13	0,3982	20	0,5806
7	0,5603	14	0,3510		

Todas las muestras superan la comparación con el valor crítico:

$$A^*_{Calculado} < A^*_{Crítico} = 1{,}029 \quad (16)$$

7.3 Prueba de Frecuencias dentro de un Bloque sobre Subsecuencias

Para una sola muestra tenemos, el siguiente resultado de la prueba Frecuencias de Dentro de un Bloque:

Tabla 6. Anderson-Darling

Sub.	$Datos$	x_i	x_{n+1-i}	PAE_{x_i}	$PAE_{x_{n+1-i}}$	$2i-1$	B_i
1	0,3778	0,3692	0,7695	0,1121	0,9618	1	-5,4538
2	0,4048	0,3778	0,6737	0,1249	0,8548	3	-12,0294
3	0,6710	0,4048	0,6710	0,1712	0,8501	5	-18,3145
4	0,4173	0,4173	0,5797	0,1960	0,6391	7	-18,5428
5	0,7695	0,4983	0,5589	0,4007	0,5794	9	-16,0257
6	0,4983	0,5589	0,4983	0,5794	0,4007	11	-11,6351
7	0,6737	0,5797	0,4173	0,6391	0,1960	13	-8,6551
8	0,5797	0,6710	0,4048	0,8501	0,1712	15	-5,2522
9	0,5589	0,6737	0,3778	0,8548	0,1249	17	-4,9354
10	0,3692	0,7695	0,3692	0,9618	0,1121	19	-2,9990
							-10,3843

Estadístico de Anderson-Darling

$$A^2 = -1(C - n) = 0{,}3843 \quad (17)$$

Estadístico de Anderson-Darling ajustado:

$$A^* = 0{,}3843\left(1 + \frac{3}{4 \cdot 10} + \frac{9}{4 \cdot 10^2}\right) = 0{,}4218 \quad (18)$$

Comparación con el valor crítico:

$$A^* = 0{,}4218 < A^*_{Crítico} = 1{,}029 \tag{19}$$

Realizado el estudio para veinte muestras nos queda:

Tabla 7. Análisis de muestras

Muestra	$A^*_{calculado}$	Muestra	$A^*_{calculado}$	Muestra	$A^*_{calculado}$
1	0,4218	8	0,2911	15	0,2990
2	0,7180	9	0,3690	16	0,5113
3	0,4277	10	0,3559	17	0,4903
4	0,3966	11	0,3749	18	0,3947
5	0,7371	12	0,3873	19	0,4911
6	1,2237	13	0,3727	20	0,8786
7	0,4704	14	0,3308		

La muestra 6 no supera la comparación con el valor crítico:

$$A^*_{Calculado} = 1{,}2237 > A^*_{Crítico} = 1{,}029 \tag{20}$$

7.4 Prueba de Rachas sobre Subsecuencias

Para una sola muestra tenemos, el siguiente resultado de la prueba de Rachas:

Tabla 8. Anderson-Darling

Sub.	$Datos$	x_i	x_{n+1-i}	PAE_{x_i}	$PAE_{x_{n+1-i}}$	$2i-1$	B_i
1	0,4371	0,0062	0,8719	0,0175	0,8849	1	-6,2087
2	0,7724	0,3214	0,7793	0,1830	0,8013	3	-9,9426
3	0,8719	0,3254	0,7724	0,1871	0,7939	5	-16,2771
4	0,3254	0,4371	0,7557	0,3221	0,7752	7	-18,3796
5	0,6994	0,6103	0,6994	0,5794	0,7057	9	-15,9212
6	0,7793	0,6994	0,6103	0,7057	0,5794	11	-13,3597
7	0,0062	0,7557	0,4371	0,7752	0,3221	13	-8,3631
8	0,3214	0,7724	0,3254	0,7939	0,1871	15	-6,5695
9	0,7557	0,7793	0,3214	0,8013	0,1830	17	-7,2020
10	0,6103	0,8719	0,0062	0,8849	0,0175	19	-2,6577
							-10,4881

Estadístico de Anderson-Darling

$$A^2 = -1(C - n) = 0{,}4881 \tag{21}$$

Estadístico de Anderson-Darling ajustado:

$$A^* = 0{,}4881\left(1 + \frac{3}{4 \cdot 10} + \frac{9}{4 \cdot 10^2}\right) = 0{,}5357 \tag{22}$$

Comparación con el valor crítico:

$$A^* = 0{,}5357 < A^*_{Crítico} = 1{,}029 \tag{23}$$

Realizado el estudio para veinte muestras nos queda:

Tabla 9. Análisis de muestras

Muestra	$A^*_{calculado}$	Muestra	$A^*_{calculado}$	Muestra	$A^*_{calculado}$
1	0,5357	8	0,9478	15	0,3588
2	0,5261	9	0,4770	16	0,6571
3	0,8470	10	0,4758	17	0,2363
4	0,4345	11	0,4366	18	0,2773
5	0,5692	12	0,4714	19	0,4860
6	1,0313	13	0,2767	20	0,6534
7	0,2088	14	0,4716		

La muestra 6 no supera la comparación con el valor crítico:

$$A^*_{Calculado} = 1{,}0313 > A^*_{Crítico} = 1{,}029 \tag{24}$$

7.5 Resultados de las Pruebas de Segundo Nivel

Para el análisis de los resultados, se determina la proporción de muestras que superan las pruebas, y con esos datos se construye un gráfico de puntos, luego se verifica si los mismos caen dentro de los límites superior e inferior, donde k es el número de muestras. En nuestro caso $k = 20$ y el nivel de significancia elegido es: α = 0.01, los límites quedan de acuerdo con las ecuación 25 y 26.

$$LS = (1-\alpha) + 3 \cdot \sqrt{\alpha(1-\alpha)/20} = 1{,}06 \tag{25}$$

$$LI = (1-\alpha) - 3 \cdot \sqrt{\alpha(1-\alpha)/20} = 0{,}92 \tag{26}$$

Se consideran todas pruebas, los resultados se indican en la tabla 10.

Tabla 10. Análisis de muestras

Prueba	Pasan	Proporción	Superior	Inferior	Aceptado
1	20	1,00	1,06	0,92	Sí
2	19	0,95	1,06	0,92	Sí
3	19	0,95	1,06	0,92	Sí

En la figura 5 se aprecia mejor el resultado:

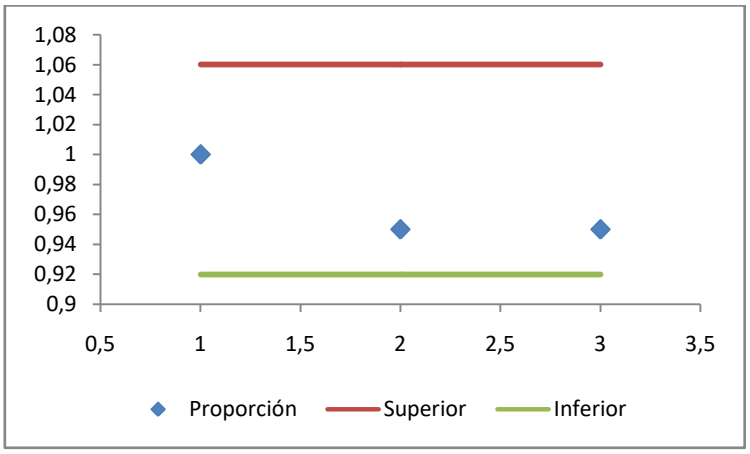


Figura 5. Proporción de muestras que superan las Pruebas

8 Análisis Final

En base a los resultados de las pruebas se realiza una tabla resumen.

Tabla 11. Pruebas de Primer Nivel

Análisis	Pruebas	Resultados
Proporción que pasan las pruebas	Frecuencias	Supera
	Frecuencias dentro de un bloque	Supera
	Rachas	Supera
	La más larga rachas de unos en un bloque	Supera
	Sumas acumulativas adelante	Supera
	Sumas acumulativas atrás	Supera
	Entropía estimada	Supera

Tabla 12. Pruebas de Segundo Nivel

Análisis	Pruebas	Resultados
Anderson Darling Normal	Frecuencias	Supera
	Frecuencias dentro de un bloque	Supera
	Rachas	Supera

En definitiva las secuencias que entrega el generador superan las pruebas de Primer y Segundo Nivel, por lo tanto se pueden considerar pseudoaleatorias.

Conclusiones

El procedimiento explicado en este documento permite incrementar el nivel de seguridad con respecto a los resultados de las Pruebas de Primer Nivel que se practican a una secuencia entregada por un Generador Aleatorio Binario.

En trabajos futuros se podrían incorporar otras pruebas al Segundo Nivel, y también someter esos resultados a la Prueba de Bondad a Ajuste de Anderson-Darling con otras distribuciones tales como Exponencial, Weibull, Normal, Gamma. Por otro lado, la generalización del método exigirá una solución que trabaje sobre todas las subsecuencias posibles; esta construcción muestra características que favorecerían la aplicación de una solución de carácter recursivo.

Tal como se mencionó oportunamente, la generación de bits aleatorios de alta calidad criptográfica reviste un interés significativo. Por ejemplo, el grupo de investigación que produjo este documento, utiliza como insumo este tipo de generadores en sus desarrollos sobre voto electrónico [9].

Referencias

1. Intel Math Kernel Library Vector Statistical Library Notes. Intel Corporation. (2008)
2. L'ecuyer, P., Simard, R.: TestU01: A C Library for Empirical Testing of Random Number Generators. ACM Transactions on Mathematical Software, Vol. 33, No. 4, Article 22. (2007).
3. Accardi, L., Gabler, M.: Statistical Analysis of Random Number Generators. Universit`a di Roma. Brandenburg Technical University Cottbus. (2010).
4. Zhu, S., Ma1, Y., Lin1, J., Zhuang1, J. and Jing, J.: More Powerful and Reliable Second-level Statistical Randomness Tests for NIST SP 800-22. Chinese Academy of Sciences, Beijing, China (2010).
5. Kahya, G.: New Modified Anderson-Darling and Cramer-Von Mises Goodness-Of-Fit Tests for a Normal Distribution with Specified Parameters. Faculty of the School of Engineering of the Air Force institute of Technology Air University (1991)
6. Moreno Quintero, E., Cruz, V.: Modelos de Probabilidad en Transporte e Ingeniería. Usos Comunes y Ajuste de Datos. Publicación Técnica No. 545. (2019).
7. Rukhin, A., Soto, J., Nechvatal, J., Smid, M., Barker, E., Leigh, S., Levenson, M., Vangel, M., Banks, D., Heckert, A., Dray, J., and Vo, S.: A Statistical Prueba Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. National Institute of Standards and Technology (2000).
8. Montejano, G., García, P., Farías, A.: Combinación No Lineal de Generadores LFSR con Funciones No Lineales de Filtrado Paralelas (2020).
9. García, P., Bast, S., Montejano, G.: Recuento y Recuperación de Sufragios en OTP – Vote. Simposio de Informática en el Estado (SIE) del XLIII CLEI (Conferencia Latinoamericana de Informática) y 46° JAIIO (Jornadas Argentinas de Informática e Investigación Operativa). ISSN: 2451-7534. Ps. 38 a 51. UTN (FRC), Córdoba, Argentina (2017).

Avoided traffic jams optimising the scheduling of signalised intersection of Mendoza Downtown and surroundings*

Pablo Javier Vidal[1,2][0000−0001−6502−8010] and Ana Carolina Olivera[1,2][0000−0001−7825−1959]

[1] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Universidad Nacional de Cuyo (UNCuyo). ITIC. Padre Jorge Contreras 1300, M5502JMA Mendoza, Argentina.
[2] Facultad de Ingeniería (UNCuyo), Centro Universitario Edifico 6, M5502KAF, Mendoza, Argentina. {pjvidal,acolivera}@conicet.gov.ar

Abstract. Congestion and pollution are crucial parameters of quality for living in a city. A correct synchronisation of traffic signals cycles can relieve these issues considering the traffic flow in an area. However, the synchronisation of large scale real scenarios is still an open research topic. In this work, an analysis of the use of optimisation algorithms for signalised intersections scheduling of Mendoza Downtown and surroundings is presented. The results confirm that correct traffic lights programming in large scale scenarios applying optimisation algorithms contributes to improving the air quality and traffic flow of vehicles.

Keywords: optimisation algorithms · traffic lights scheduling · Mendoza.

1 Introduction

Mendoza is a city in Argentina and the capital of the province of the same name. Mendoza, like many other urban agglomerations, suffers from traffic jams, pollution and other traffic problems. Different authors argue that correct traffic signals scheduling can provide benefits concerning the waiting time and velocity of vehicles [4, 16]. In particular, in developing countries, investment in intelligent traffic lights infrastructure is still an open topic.

The configuration of traffic signals with many intersections must be coordinated to achieve a more dynamic traffic flow. The scheduling of hundreds of traffic signal is very complex and challenging [15]. To overcome the potentially high computational complexity, it requires the implementation of advanced techniques to improve small and large configuration traffic signals.

In recent years, different meta-heuristics have been used to solve several traffic signal scheduling problems [4–6, 11, 15]. In particular, Genetic Algorithms

* Supported by Universidad Nacional de Cuyo Project Type 1 B081.

(GA) [7] and Simulated Annealing (SA) [10] have proven to obtain better solutions than the existing ones in different scenarios [2, 8]. The exploration capacity of this type of techniques allows scanning the ample search space of the problem and at the same time trying to improve those promising solutions.

This work introduces (i) analysis and comparison of optimisations algorithms considering a large scale real scenario of Mendoza Downtown and surrounding. (ii) The optimisation model contains a more realistic characterisation of the dynamics of the traffic flow that impact in reducing emissions and fuel consumption even without these variables considering in the fitness function.

The manuscript is organised as follows: the Traffic Lights Scheduling Problem is presented in Section 2, and in subsection 2.1 the main references about traffic signals scheduling and optimisations approaches are described. Next, Section 3 exposes the general structure of the optimisation approach, consider the combination of optimisations algorithms and micro-simulation. Section 4 display the results achieved with Genetic Algorithms, Simulated Annealing, random search and the SUMO traffic lights schedule. The analysis of the results is made in Section 5, and conclusions are shown in Section 6.

2 Traffic Lights Scheduling Problem

The programming of the traffic light cycles is defined as the seconds during which a state of lights or phase remains active of the set of states of lights that crosses an intersection. These phases follow the traffic rules and are previously defined, considering the topology of each intersection. For the present study, the phases will always be valid and previously defined, varying only the seconds during which they are active. Furthermore, each intersection has an *offset* that regulates the moment within the phases at which each intersection begins.

For this work, we want to optimise the programming of the traffic light cycles by analysing their impact on the traffic network, considering as the objective function to minimise Z defined in (1) [16].

$$Z = \frac{w}{v} \times 3.6 \tag{1}$$

where, w is the average waiting time in seconds (s) that the vehicles had on the traffic network while v is the average speed of the vehicles (m/s) on the network. Then the constant 3.6 converts from m/s to km/h. Z then expresses the relationship between the wait and the speed such that the longer the wait for stopped vehicles, the greater the congestion. This implies that the speed of the vehicles on the network decreases. In other words, the lower the Z, the better the traffic flow. This means less waiting times on the network and higher vehicle speeds.

Both, the average waiting time and the average speeds, are obtained from the traffic simulation with SUMO [12] of Mendoza scenario. SUMO is a popular micro-simulator in the literature that takes as input a map, programming of the map's traffic lights, vehicles and their routes to simulate the traffic network. As

a result, the simulator obtains relevant information on the efficiency parameters of the traffic network in the different streets of the city, the number of vehicles circulating, among others. From this data, the value of Z is constructed.among others. From this data the value of Z is constructed.

2.1 Related works

Exist, in literature, several works related to optimisations algorithms and traffic lights scheduling problem. [3, 13, 8, 14]. In [3], the author optimises the duration of green lights for traffic in the city of Rouen, France. The purpose of this work is to minimise total waiting time and polluting emissions from vehicles. The proposal improves the traffic light cycle program, minimising most of the variables analysed when compared to the initial program used in the real network. [14] use the same approach multiobjective analysis for three small instances of Montevideo, Uruguay.

In [2], the author proposes the resolution of traffic lights scheduling for only one intersection using GA and SA. The results are promising; however, the main drawback is the small and straightforward instance use to compare the algorithms.

Jovanović et al. [8] develop a method based Bee Colony Optimisation to solve the problem of programming the traffic light cycles. They use as a function of evaluation of the solutions a mono-objective function that considers the total travel time of the vehicles in the traffic network. Comparisons are made with a Simulated Annealing Algorithm. [16] propose an approach to the optimisation of the traffic light programming considering only variables of the vehicular flow. Differential Evolution, combined with a micro-simulator, was used as an optimisation algorithm. Two small instances are used to test the algorithm.

In this work, we applied optimisation algorithms -Genetic Algorithms (GA), Simulated Annealing (SA) and Random Search (RS)- in combination to a micro-simulator [12] for the large scale real scenario of Mendoza Downtown and surroundings [7, 10]. The fitness function contemplates only traffic flow variables for the optimisation process. The results show that traffic lights scheduling that considers the traffic flow can improve vehicle velocity and, at the same time, reduce the greenhouse emission even if these are not considered in the fitness.

3 Optimisation approach

We used a metaheuristic optimisation techniques to optimise scheduling traffic signal intersections in Mendoza scenario to minimise (1). SUMO provides the information to calculus the fitness function for each solution proposed. Here we used two canonical versions of very well suited optimisation algorithms, Genetic Algorithms (GA) and Simulated Annealing (SA) in the optimisation process.

The general optimisation methodology is shown in Fig. 1. An optimisation algorithm generates initial solutions (programming of the traffic lights cycles) that are sent to the simulator for these to be evaluated. SUMO, with the map of

Mendoza, the programming of the traffic light cycles (`programming_traffic_lights.xml`) and the information of the vehicle routes (`routes.rou.xml`) performs the simulation. From the simulator output data, the solutions are evaluated. Next, the algorithm begins a solution improvement process by trying to improve the current solution configurations. Each algorithm applies different disturbance operators to explore different search spaces or exploit solutions close to the current one. This process is repeated until the termination condition is reached.

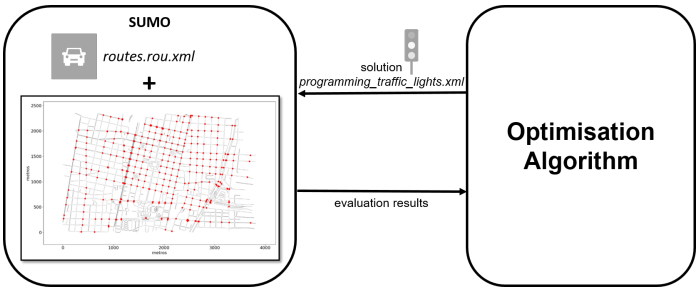


Fig. 1. Scheme of the optimisation approach.

A solution for the traffic lights scheduling problem is codify like a list of integers $[o_1 t_{1_1} t_{1_2} \dots t_{1_l} \dots o_h t_{h_1} \dots t_{h_j}]$, $h = 1$ to 335 where $o_1, \dots, o_h$ are the offset of a signalised intersection $s_1, \dots, s_h$. t_{k_p} corresponds to seconds assigned of each phase $p = 1..q$, q are the number of phases for the junction s_k. The *offset* allows an intersection to start said the program at different moments in its cycle to achieve coordination (green wave) with the other traffic lights on the network and improve traffic flow. Fig. 2 illustrates the SUMO view of a signalised intersection (on the left), the representation inside the vector solution (top right) and the part of the file `programming_traffic_lights.xml` that contains the information of it (bottom right) used in the simulation. Following the recommendation of the Handbook Emission Factors for Road Transport (HBEFA) [9] Green and red phases times are between 5 and 60. Offsets are defined such that $o_i \in [0, 60]$. Yellows are set in 3 seconds to follow the HBEFA guidance.

Explore the solution space of Mendoza scenario is a complex task and exact methods are infeasible. For its resolution in this work use Genetic Algorithms, Simulated Annealing and Random Search. In particular, Random Search is used as a sanity check test.

Genetic Algorithms. GA was proposed by Holland [7] inspired by the theory of natural evolution. The algorithm begins generating an initial usually randomly. The population is evaluated using SUMO. Then, parents are selected based on their fitness. These parents are combined, and children are created. The children can mutate with a low probability. When the children are evaluated, the

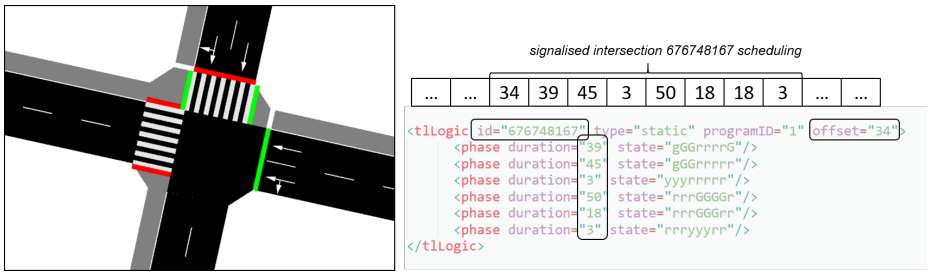


Fig. 2. A signalised intersection view in SUMO. Inside the vector solution and in the file, use to evaluation during the simulation.

selection for the next generation is achieved, a new population is constructed, and the process repeats until the condition stop is reached, and the best solution is reported.

Simulated Annealing. SA is a stochastic technique mimics the physical annealing process [10]. Like the physical annealing begins with a starter solution (traffic lights scheduling random created) and a temperature. The solution is evaluated using SUMO. The temperature is reduced using the α function. The iteration process consists in to generate a neighbourhood of solutions and pick up one to compare to the actual. Suppose the old solution is worst than the new one, the solution is replaced. However, if the new one is worst than the old, a probability number is generated and compare to the energy magnitude $e^{\frac{-\Delta c}{t}}$, where c is the difference between fitness function and t is the current temperature. In this way, the worst solution can be chosen to replace the old. This process is crucial to avoid local minimum. This process is repeated until a condition stop is reached, and the best solution is returned.

Random Search. RS initialises a solution x randomly and evaluates it. Then, until a termination criterion is met, RS repeats the process of creating and evaluating a new solution y. If y is better than x, x is replaced with y; otherwise, x is left unchanged.

Finally, SUMO by default provides the valid combination of states that the traffic light controller can go through inside the maps specification file, and an approximation of interval times for these states. SUMO already incorporates a solver algorithm for the cycle program of traffic lights based on greedy and human knowledge. That solver will be called SUMO to compare to traffic lights cycle programs obtained by metaheuristics.

4 Experiments

This section is devoted to present the instance and the methodology used in the experiments carried out in this work. First, the instance used is explained. Then, a description of each algorithm configuration is given.

The Mendoza scenario covers 10km^2 and contains 335 signalised intersections. Fig. 3 shows the treatment of the map observed until a sumo representation is produced. First, it localised the area to study, then using OpenStreetMap a `.osm` file of streets, directions and traffic lights is obtained, then using SUMO `netconvert` tool a `.net.xml` is generated. The red points on the SUMO representation are the traffic lights to optimise.

The vehicles flowing in the simulation are obtained by observation of Google Traffic Map (https://www.google.com.ar/maps/) between 8 a.m to 9 a.m in business day. The SUMO tools `duarouter` and `route2trips` generated the traffic over the net (`vehicles_routes.rou.xml`).

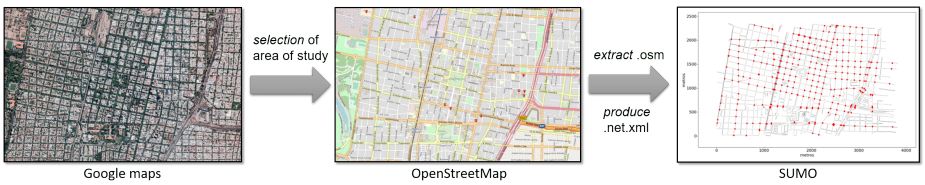


Fig. 3. Google Relieve Map, OpenstreetMap and SUMO map processing.

Table 1 presents the settings for each algorithm tested. GA uses a specific mutation operator that change the value of all the phases of each intersection between a range of [-5,+5] [16].

Table 1. Algorithms configuration.

Algorithm	Parameter	Value
Genetic Algorithm	Population	50
	Selection Operator	Binary Tournament
	Crossover Operator	Single Point Crossover (prob.=0.9)
	Mutation Operator	Traffic Mutation (prob.=0.1)
	Replacement Operator	Elitism
Simulated Annealing	Population	1
	Mutation Operator	Traffic Mutation (prob.=0.2)
Random Search	Population	1

Each algorithm is running 30 times with a stopping condition defined as 5000 fitness evaluations to guarantee a similar exploration of the search space.

Experiments were performed in the Toko cluster (toko.uncu.edu.ar) with an AMD Opteron/Epyc processor (64 cores and 128 GB of RAM). The operating system is Ubuntu 18.04 LTS. Algorithms are implemented in jmetalpy [1].

5 Analysis and Results

This section reports the numerical results of the optimisation approach on Mendoza scenario for traffic lights schedule.

Table 2 presents in column one the name of the optimisation algorithm, columns 2, 3 and 4 show the mean and its standard deviation, and the best fitness obtain for GA, RS, SA and SUMO. SUMO scheduling always attains the same fitness value (standard deviation = 0). Columns 5 and 6 present for each algorithm the runtime consumption (seconds) and its standard deviation.

Genetic Algorithm reaches the mean and best values, and the small standard deviation of fitness follows by SA without count deterministic algorithm of SUMO. Random search obtains the worst results, even poorer than the value obtained with the configuration bring by SUMO. GA can achieve the best fitness values without detriment to runtime. SUMO do only one simulation to obtain the fitness value so consume lower time to the other approaches.

Table 2. Mean, standard deviation and best of the Fitness quality indicator, and the mean time consumption by the algorithms and the standard deviation.

Optimisation	fitness			Time Execution	
Algorithm	mean	std dev.	best	time	std. dev.
GA	20.09	0.34	19.42	145664.24	632.08
SUMO	25.10	0.00	25.10	51.37	0.00
RS	25.12	0.90	23.83	146524.34	450.85
SA	22.04	0.52	21.18	146038.20	682.90

Fig. 4 shows the boxplot for fitness distribution values of GA, SUMO, RS and SA. SUMO generates one fitness value, and it represented by a line. It's evident than GA obtains the best fitness (lower values) than the other approaches and, at the same time, the less dispersion of fitness value. RS has the worst fitness value even worst than SUMO.

Finally, Table 3 shows some indicators of interest to understand how the configuration obtained by each algorithm affect the traffic flow. In the same way, it can observe in other variables such as pollution emissions. Table 3 indicates the mean values for the speed of vehicles (km/h), density (veh/km), occupancy index of the edges of Mendoza in column 2, 3 and 4, respectively. Column 5 presents the total number of cars that arrived at the destination. Columns 6, 7 and 8 show the average of CO_2, NO_x emissions (kg/h) and fuel consumption (l/h) generate by adding the consumption or emission of each vehicle in each simulation step and dividing it by the total number of cars that arrived at their destination. GA gets the best results follows by SA. The exciting thing is that even when more vehicles arrive at their destination (GA), this has a positive impact on fuel consumption and pollutant emissions, which suggests that cars flow better through the traffic network. This positive impact its confirm in density and occupancy where GA has a lower density and occupancy than the rest of the

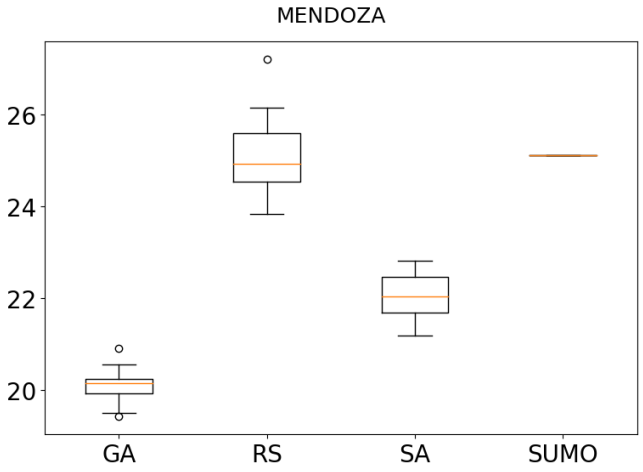


Fig. 4. Boxplot shows GA, SUMO, RS and SA distribution of fitness.

algorithms, which implies that vehicles quickly leave the streets without being stuck for a long time on them.

Table 3. Best fitness average speed, average density and average occupancy of the streets.

Alg.	Indicators						
	avg. speed	avg. density	avg. occupancy	vehicles arrived	CO_2	NO_x	Fuel
GA	20.71	31.02	1.53	686	9119.83	3.94	3923.35
SUMO	19.55	37.86	1.81	642	9878.75	4.28	4250.02
RS	19.77	36.35	1.73	629	10124.29	4.39	4355.64
SA	20.06	32.80	1.59	658	9585.01	4.15	4123.44

Finally, Fig. 5 shows the edge mean occupancy by step simulation, taking into account the best configuration of traffic lights obtained by GA in comparison with the simulation of SUMO configuration. Line blue corresponds to GA, and dotted line red corresponds with SUMO results. Throughout the entire simulation, Genetic Algorithms achieve less occupancy index for the edges. In particular, starting from 300 simulated seconds GA reduces the occupancy index of lanes considerably.

In resume, fitness function can capture characteristics in traffic flow that allow traffic lights scheduling improvement. GA produces traffic lights cycle scheduling that elevate the traffic flow over Mendoza scenario, reduces pollution emissions and fuel consumption without detriment of vehicles speed.

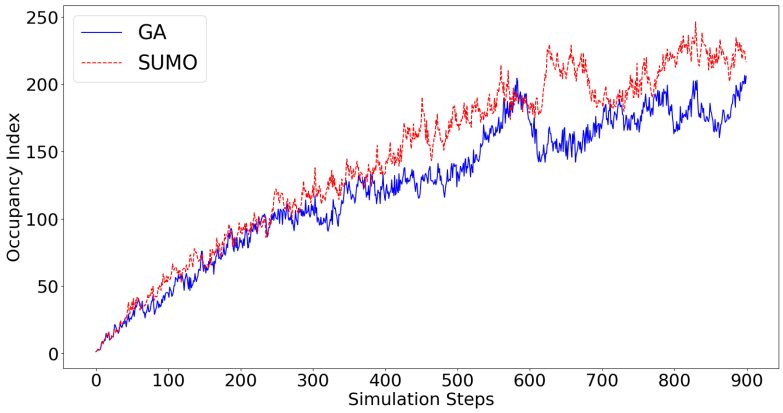


Fig. 5. Evolution of the street occupancy for one solution of GA vs. SUMO.

6 Conclusions

In this work, we analyse the treatment with optimisation algorithms of traffic lights cycle scheduling of a large scale scenario. Mendoza Downtown and surroundings have 335 signalised intersection. The area cover in this work has about ten km^2. Genetic Algorithms, Simulated Annealing, Random Search and the algorithm configuration for traffic lights cycles scheduling of the micro-simulator were comparison. The Genetic Algorithms overcomes the other optimisation techniques, and SUMO proposed schedule. The improvement of the traffic flow produces a reduction of the density of vehicles over the net and at the same time, increase the velocity of the vehicles while reducing the fuel consumption and pollutant emissions.

Acknowledgements. Authors thank CONICET, Argentine. Vidal and Olivera thank Universidad Nacional de Cuyo, Facultad de Ingeniería and Secretaría de Investigación, Internacionales y Posgrado for grant Type 1 B081.

References

1. Benitez-Hidalgo, A., Nebro, A.J., Garcia-Nieto, J., Oregi, I., Del Ser, J.: jmetalpy: a python framework for multi-objective optimization with metaheuristics. arXiv:1903.02915 (2019)
2. Burvall, B., Olegård, J.: A comparison of a genetic algorithm and simulated annealing applied to a traffic light control problem: A traffic intersection optimization problem (2015)
3. Damay, N.: Multiple-objective optimization of traffic lights using a genetic algorithm and a microscopic traffic simulator. Master's thesis, KTH, School of Computer Science and Communication (CSC) (2015)

4. Gao, K., Zhang, Y., Su, R., Yang, F., Suganthan, P.N., Zhou, M.: Solving traffic signal scheduling problems in heterogeneous traffic network by using meta-heuristics. IEEE Transactions on Intelligent Transportation Systems **20**(9), 3272–3282 (2019)
5. Gao, K., Zhang, Y., Zhang, Y., Su, R., Suganthan, P.N.: Meta-heuristics for bi-objective urban traffic light scheduling problems. IEEE Transactions on Intelligent Transportation Systems **20**(7), 2618–2629 (2019)
6. Gao, K., Zhang, Y., Sadollah, A., Su, R.: Optimizing urban traffic light scheduling problem using harmony search with ensemble of local search. Applied Soft Computing **48**, 359 – 372 (2016). https://doi.org/10.1016/j.asoc.2016.07.029
7. Holland, J.H.: Genetic algorithms. Scientific American (Jul 1992)
8. Jovanović, A., Nikolić, M., Teodorović, D.: Area-wide urban traffic control: A bee colony optimization approach. Transportation Research Part C: Emerging Technologies **77** (2017)
9. Keller, M.: Handbook of Emission Factors for Road Transport (HBEFA) 3.1. quick reference. Technical Report INFRAS. John Wiley & Sons (2010)
10. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220 4598**, 671–680 (1983)
11. Li, X., Sun, J.: Signal multiobjective optimization for urban traffic network. IEEE Transactions on Intelligent Transportation Systems **19**(11), 3529–3537 (2018)
12. Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wießner, E.: Microscopic traffic simulation using SUMO. In: The 21st IEEE International Conference on Intelligent Transportation Systems. IEEE (2018), https://elib.dlr.de/124092/
13. Olivera, A.C., García-Nieto, J.M., Alba, E.: Reducing vehicle emissions and fuel consumption in the city by using particle swarm optimization. Applied Intelligence **42**(3), 389–405 (2015)
14. Péres, M., Ruiz, G., Nesmachnow, S., Olivera, A.C.: Multiobjective evolutionary optimization of traffic flow and pollution in Montevideo, Uruguay. Applied Soft Computing **70**, 472–485 (sep 2018). https://doi.org/10.1016/j.asoc.2018.05.044
15. Thaher, T., Abdalhaq, B., Awad, A., Hawash, A.: Whale optimization algorithm for traffic signal scheduling problem. In: Jain, L.C., Peng, S.L., Alhadidi, B., Pal, S. (eds.) Intelligent Computing Paradigm and Cutting-edge Technologies. pp. 167–176. Springer International Publishing, Cham (2020)
16. Vidal, P.J., Olivera, A.C.: Management of urban traffic flow based on traffic lights scheduling optimization. IEEE Latin America Transactions **17**(1), 102–110 (Jan 2019). https://doi.org/10.1109/TLA.2019.8826701

Analítica de Negocio en la Gestión de Vuelos Gubernamentales: Una Revisión Sistemática de la Literatura

María Inés Lund[1,2][0000-0002-1687-7267], Silvina I. Migani[1,2], Alejandro Riveros[1] y Susana B. Ruiz[3,2]

[1] Departamento de Informática, FCEFN, Universidad Nacional de San Juan, Argentina
[2] Instituto de Informática, FCEFN, Universidad Nacional de San Juan, Argentina
[3] Departamento de Geofísica y Astronomía, FCEFN, Universidad Nacional de San Juan, Argentina

mlund@iinfo.unsj.edu.ar,
{silvina.migani, alejandroriveros09, sbruizr}@gmail.com

Abstract. La Dirección Provincial Aeronáutica depende de la Secretaría General de la Gobernación de la provincia de San Juan de la República Argentina y se encarga de gestionar los recursos necesarios para cumplimentar vuelos oficiales, sanitarios y de fumigación, tratando de maximizar el servicio ofrecido a la comunidad y ajustándose al presupuesto otorgado. Usar métodos, procedimientos y tecnologías propias de la Analítica e Inteligencia de Negocio (Business Analytics, Business Intelligence) le permitiría optimizar el uso de recursos y planificar con menos incertidumbre.
El objetivo del presente trabajo fue realizar una revisión sistemática de la literatura para identificar métodos, técnicas y herramientas propias de Business Intelligence y Business Analytics utilizados en el área gubernamental, con el fin de administrar, operar y mantener eficientemente los recursos aeronáuticos del Estado.
El estudio realizado ha puesto en evidencia la escasa bibliografía en relación a la temática abordada, y así descubrir un vacío en el conocimiento científico dentro de este dominio concreto.

Keywords: Gestión de vuelos gubernamentales - Analítica de negocio – Revisión sistemática de la literatura.

1 Introducción

Hoy en día, las organizaciones que pretenden lograr mantener o aumentar el éxito de sus negocios, necesitan fundamentalmente conocimiento sobre ellos mismos, sobre la oferta y las estrategias comerciales de sus competidores y de las preferencias de sus clientes (actuales y potenciales). El nivel de competencia global y la variabilidad de los mercados son tan altos que ese conocimiento es un factor vital para cualquier tipo de organización, sea pública, privada, pequeña o grande.

La Inteligencia y Analítica de Negocios (BI: Business Intelligence y BA: Business Analytics respectivamente) son términos que refieren al uso de métodos, procedimientos y tecnologías que permiten reunir datos (pasados y presentes, internos y externos) y aplicar técnicas analíticas de extracción de conocimiento a partir de estudios descriptivos, predictivos y prescriptivos que apoyen y favorezcan la toma de decisiones.

El caso bajo estudio, la Dirección Provincial Aeronáutica (DPA) que depende de la Secretaría General de la Gobernación de la provincia de San Juan (SGG), se encarga de gestionar los vuelos realizados con las aeronaves de la provincia (aviones y helicópteros) para dar servicio a la comunidad. Los vuelos pueden ser con fines oficiales, sanitarios o de fumigación. En la asignación de recursos es fundamental considerar la urgencia y la prioridad de los vuelos, que en orden decreciente son: sanitarios, oficiales y de fumigación.

La DPA presenta actualmente la necesidad de administrar eficientemente recursos (humanos, materiales y financieros) planificando vuelos, licencias y capacitaciones de pilotos, mantenimiento interno y externo de aeronaves, etc., para lograr que la asignación de recursos sea más productiva y eficiente, minimizando gastos y maximizando la satisfacción de sus usuarios (funcionarios o simples miembros de la comunidad). Un punto de partida para generar propuestas, ante la problemática planteada, es investigar sobre antecedentes de trabajos realizados dentro del ámbito aeronáutico. Este trabajo tiene como objetivo exponer los hallazgos encontrados en el área, en relación a BI y BA, que surgieron de la aplicación de una de revisión sistemática de la literatura.

2 Metodología: Revisión Sistemática de la Literatura

2.1 Identificación de las RQS (Preguntas de investigación)

RQ1. Los datos que se usan para el análisis ¿qué origen tienen? ¿Incluyen datos provenientes de redes sociales? ¿Qué tipo de datos son?
RQ2. ¿Qué algoritmos, métodos o técnicas fueron utilizados?
RQ3. ¿Cuáles son los tipos de resultado que presentan los estudios encontrados?
RQ4. ¿Presentan resultados de aplicar técnicas o recursos para la descripción o visualización de la información? ¿Cuáles?
RQ5. ¿Especifican alguna herramienta/software de análisis y procesamiento de los datos? ¿Cuáles?

2.2 Búsqueda literaria

El objetivo en esta etapa es reunir evidencias sobre el tema planteado, respondiendo a las preguntas de investigación definidas. El proceso se realiza siguiendo los lineamientos propuestos por Petersen et al. [1] y luego la revisión literaria de los artículos seleccionados [2].

Las bibliotecas digitales que se consideraron fueron Scopus [3], ScienceDirect [4], IEEE [5] y ACM [6].

2.3 Definición del String de Búsqueda

Se realizó una lectura previa de algunos artículos vinculados a la temática, a fin de descubrir distintos términos que podían abarcar información relacionada al objetivo de la búsqueda.

Las palabras involucradas (semillas) en las cadenas de búsqueda usadas fueron:

- *official / government / federal*
- *flight / aircraft / airplane / aeronautics*
- *business intelligence / business analytics / datawarehouse / datawarehousing / datamining/ data mining / analytics/ decision support*

Dado que cada biblioteca tiene su propia lógica y estructura, se muestra a continuación las cadenas de búsqueda aplicadas en cada una de ellas.

Scopus:
(TITLE-ABS-KEY ("official flight") OR TITLE-ABS-KEY ("government flight") OR TITLE-ABS-KEY ("federal flight") OR TITLE-ABS-KEY ("official flights) OR TITLE-ABS-KEY ("government flights") OR TITLE-ABS-KEY (federal flights") OR TITLE-ABS-KEY ("official aircraft") OR TITLE-ABS-KEY ("government aircraft") OR TITLE-ABS-KEY ("federal aircraft") OR TITLE-ABS-KEY ("federal airplane") OR TITLE-ABS-KEY ("government airplane") OR TITLE-ABS-KEY ("official airplane") OR TITLE-ABS-KEY ("federal aeronautics") OR TITLE-ABS-KEY ("government aeronautics") OR TITLE-ABS-KEY ("official aeronautics")) AND (TITLE-ABS-KEY ("business intelligence") OR TITLE-ABS-KEY ("business analytics") OR TITLE-ABS-KEY ("datawarehouse") OR TITLE-ABS-KEY ("data mining") OR TITLE-ABS-KEY ("analytics") OR TITLE-ABS-KEY ("decision support")).

IEEE:
"All Metadata": ((official OR government OR federal) AND (flight OR aircraft OR airplane OR aeronautic) AND ("business intelligence" OR "business analytic*" OR datawarehous* OR "data mining" OR analytic* OR "data warehous*" OR "desicion support")).

ACM:
[[Abstract: official] OR [Abstract: government] OR [Abstract: federal]] AND [[Abstract: flight*] OR [Abstract: aircraft] OR [Abstract: airplane] OR [Abstract: aeronautic*]] AND [[Abstract: "business intelligence"] OR [Abstract: "business analytic"] OR [Abstract: "business analytics"] OR [Abstract: datawarehous*] OR [Abstract: "data mining"] OR [Abstract: datamining] OR [Abstract: analytic*] OR [Abstract: "data warehouse"] OR [Abstract: "data warehousing"] OR [Abstract: datawarehous*] OR [Abstract: "desicion support"]].

Science Direct (Elsevier): Por limitaciones del buscador, se realizaron 2 búsquedas, cuyos resultados fueron acumulados.
(official OR government OR federal) (flight OR aircraft OR airplane OR aeronautic) ("business intelligence" OR "business analytic" OR "datawarehous" OR "data mining") in Title, abstract or author-specified keywords.
(official OR government OR federal) (flight OR aircraft OR airplane OR aeronautic) (analytic OR "data warehous" OR "desicion support") in Title, abstract or author-specified keywords.

2.4 Selección

En este proceso participaron 4 revisores (los autores del presente trabajo). Cada uno realizó la búsqueda, según los criterios establecidos, en una biblioteca y analizó ligeramente si los artículos recuperados eran consistentes con el objetivo perseguido. Es así que algunas publicaciones se rechazaron en esta fase.
Los artículos seleccionados fueron plasmados en una tabla por biblioteca (tabla 2). Una vez finalizado este proceso las 4 tablas fueron unidas en una única tabla (tabla 3), indicando la biblioteca fuente (eliminando los trabajos duplicados). A posteriori, se procedió a distribuir equitativamente los trabajos a cada revisor, para realizar la revisión literaria y descartar aquellos hallazgos que no satisficieran los criterios de selección definidos. Considerando por un lado, que los trabajos fueran asignados a un revisor diferente del que lo preseleccionó inicialmente y por otro, que cada trabajo fuese analizado por 2 revisores, para minimizar resultados sesgados. Esto provocó que algunos trabajos tuviesen calificaciones diferentes, para lo cual se utilizaron las reglas de decisión definidas (tabla 1). El resultado de este proceso de selección se muestra en la tabla 4.

Tabla 1. Tabla de combinaciones de reglas de decisión [1]

		R2		
		Include	Uncertain	Exclude
R1	Include	A	B	D
	Uncertain	B	C	E
	Exclude	D	E	F

Siendo R1 y R2 cada revisor, las acciones a tomar en cada caso: A: Se incluirá, B: Se incluirá, C: Se volverá a revisar, D: Se volverá a revisar, E: Se excluirá, F: Se excluirá.

Criterios de Inclusión
- Artículos cuyo foco esté relacionado con el objetivo de la revisión.
- Artículos que estén publicados a partir del año 2000.
- Trabajos pertenecientes al área Ingeniería y Ciencias de la Computación.
- Revisiones bibliográficas.

Criterios de Exclusión
- Trabajos duplicados.
- Trabajos fuera del intervalo de los años considerados.
- Hallazgos a los que no se tenga acceso al abstract, al menos.

- Prefacios o Introducciones de libros, revistas o proceedings.
- Trabajos que no tengan que ver con el objeto de la búsqueda.

2.5 Extracción de datos y Clasificación

Los artículos o hallazgos encontrados en las búsquedas fueron volcados a tablas, como la que se muestra a continuación, siguiendo los criterios de inclusión y exclusión previamente mencionados.

Tabla 2. Recuperación de artículos seleccionados por biblioteca

Nro	Título	Autor/es	Año	Tipo de Artículo	Publisher	Abstract
1						
2..						

Luego de realizar la búsqueda por biblioteca y selección inicial, los resultados fueron unidos en una tabla como la siguiente, para poder realizar la segunda selección, extracción y catalogación, por medio de la revisión literaria de cada hallazgo.

Tabla 3. Artículos seleccionados y catalogados

Nro	Decisión (si/no/duda)	+columnas de la tabla 2	Técnica aplicada	Objeto/fin	Source	Rev1	Rev2
1							
2..							

Finalmente, con los hallazgos seleccionados luego de los 2 procesos previos, se preparó una tabla como la siguiente

Tabla 4. Artículos resultantes de la revisión literaria

Nro	+ columnas tabla 2 + source	keywords	Tipos de datos	Teorías - técnicas	Software - herramientas	Tipo de visualización	Tipo de resultado
1							
2..							

2.6 Amenazas a la validez

Las amenazas a la validez que pueden presentarse son:
Descriptivas. Referidas a la posibilidad de que observaciones expuestas no sean descriptas correcta y objetivamente. Este problema podrá reducirse publicando y exponiendo a la comunidad científica, el trabajo realizado.
Teóricas. El riesgo de no poder lograr lo que se intentó realizar. Esta situación puede deberse a:

- No se seleccionaron la totalidad de los estudios primarios relacionados al objetivo planteado.
- La calidad de los trabajos seleccionados.
- Los datos extraídos no responden claramente las preguntas planteadas.
- El sesgo que puedan cometer los revisores en la selección de los trabajos.

Para mitigar este problema participaron varios revisores en la selección de artículos, se aplicó la tabla de decisiones presentada en la sección 2.4 (tabla 1), y se previó pedir asesoramiento a expertos en caso de ser necesario.

Interpretativas. No poder alcanzar conclusiones razonables en base a los datos extraídos. Nuevamente aquí puede influir el sesgo del investigador al interpretar los datos. Por ello una medida paliativa es la participación de varios investigadores, describir detalladamente el proceso de investigación seguido; entre otras.

3 Trabajo realizado

Cada revisor realizó la búsqueda en una biblioteca asignada. La fecha de búsqueda fue el 17/02/2020. Algunas publicaciones se rechazaron en una primera selección, debido a no presentar ninguna relación con el objetivo de la búsqueda. Los resultados se muestran en la tabla 5.

Tabla 5. Total de publicaciones recuperadas, discriminando las rechazadas inicialmente y las que fueron aceptadas para su análisis

Biblioteca	Total recuperadas	Rechazadas iniciales	Total extraído para análisis
SCOPUS	4	0	4
ACM	54	49	5
IEEE	76	47	29
SCIENCE DIRECT	25	19	6
Total de publicac.	159	115	44

Las publicaciones aceptadas para el primer análisis se recuperaron en una tabla, respetando las indicaciones mencionadas en la tabla 2.

Análisis de los hallazgos seleccionados

El total de publicaciones extraídas para su análisis (última columna de la tabla 5) se ordenaron alfabéticamente y se distribuyeron equitativamente, entre los revisores, de tal manera, que cada una de ellas fuera analizada por dos revisores. En los casos de discrepancias se aplicó la tabla de decisiones del punto 2.4 (tabla 1).

Este proceso de revisión fue muy exhaustivo ya que en una primera revisión no se detectó ningún resultado estrechamente vinculado al objetivo, con lo cual se decidió revisar nuevamente los 44 artículos, y se extrajeron 10 publicaciones que se consideraron parcialmente o débilmente relacionadas, como se muestra en la tabla 6.

Tabla 6. Publicaciones seleccionadas de la Revisión literaria realizada

Biblioteca	Cant.publicaciones analizadas	Cant.publicaciones rechazadas	Cant.publicaciones relacionadas	Cant.publicaciones consideradas útiles por las técnicas utilizadas
SCOPUS	4	4	0	0
ACM	5	2	0	3
IEEE	29	24	0	5
SCIENCE DIRECT	6	4	0	2
Tot. Publicaciones	44	34	0	10

Se detalla a continuación, en la tabla 7, por biblioteca, los hallazgos débilmente relacionados, considerados para el presente trabajo:

Tabla 7. Hallazgos débilmente relacionados

Biblioteca	Cantidad de hallazgos	Referencias
ACM	3	[7], [8], [9]
IEEE	5	[10], [11], [12], [13], [14]
SCIENCE DIRECT	2	[15], [16]

Se puede observar que en la biblioteca Scopus no se encontraron publicaciones relacionadas o débilmente relacionadas con el objetivo de la búsqueda.

Se presenta en la tabla 8 los hallazgos según el tipo de publicación

Tabla 8. Hallazgos según tipo de publicación

Tipo de publicación	Cantidad de hallazgos	Referencias
Conference Article	6	[7], [8], [10], [11], [12], [13]
Journal Article	3	[9], [14], [16]
Book chapter	1	[15]

Se advierte que la mayoría de los hallazgos corresponden a artículos de conferencia, seguidos por los de Journal.

Respecto a los años de las publicaciones, en la siguiente tabla se muestra los hallazgos por año.

Tabla 9. Hallazgos según año de publicación

Año	2009	2013	2014	2015	2016	2017	2018	2019
Cantidad	1	1	1	1	2	1	2	1
Referenc.	[15]	[13]	[12]	[9]	[8], [16]	[11]	[7], [14]	[10]

Se puede observar que si bien se consideró la búsqueda desde el año 2000, recién en el año 2009 se encuentra una publicación y después, desde el 2013 en adelante, entre 1 y 2 en forma constante.

Respecto al país de origen, se ha considerado el correspondiente a los autores, en caso de ser de distintos países, se considera el del primer autor, o si son varios autores, se considera el origen de la mayoría de ellos.

Tabla 10. Hallazgos por país que realiza la investigación

País	Cantidad de hallazgos	Referencias
India	4	[9], [10], [11], [13]
Estados Unidos	2	[7], [15]
Brasil	1	[14]
Eslovaquia	1	[12]
Malaysia	1	[16]
Singapur	1	[8]

La tabla 10 muestra que la mayoría de los estudios (4) corresponden a investigaciones en India; EEUU aporta tres artículos, y Brasil, Eslovaquia, Malaysia y Singapur, 1 cada uno. Es de destacar que tres (3) de los estudios de India utilizan datos extraídos de fuentes de EEUU y en los estudios de Eslovaquia, Malaysia y Singapur también se utilizó la misma fuente de datos. Sólo dos (2) artículos indican que toman datos de otras fuentes. Este escenario podría dar pie para emprender un análisis de la situación aeronáutica de cada país. Es de destacar que India ha tenido un importante aumento en el tráfico doméstico de pasajeros en los últimos años, y también ha crecido en el desarrollo de equipamiento e infraestructura, la cantidad de profesionales informáticos y científicos de los datos, y EEUU es quien más soporte público de datos ofrece.

4 Resultados: respondiendo a las RQs planteadas

4.1 RQ1. Los datos que se usan para el análisis, ¿qué origen tienen? ¿incluyen datos provenientes de redes sociales? ¿qué tipo de datos son?

Tabla 11. Hallazgos según las características de los datos

Clasificación	Sub-clasificación	Cant.hallazgos	Referencias
Origen de los datos	Organismos públicos	7	[7], [8], [10], [11], [12], [13], [15]
	Organismos privados	2	[9], [14]
	Mixtos	1	[16]
Datos fuente relativos a	Tráfico aéreo	2	[10], [11]
	Accidentes	3	[12], [13], [16]
	Alertas	1	[14]
	Mediciones sensores	2	[7], [15]
	Datos de mercado	1	[8]
	otros	2	[9], [16]
Tamaño de la muestra	Menos de un año	3	[9], [11], [16]
	Un año	2	[7], [15]
	Más de un año	5	[8], [10], [12], [13], [14]
	Una sola	5	[7], [10], [11], [12], [13]

Cant. de fuentes de datos	Más de una	5	[8], [9], [14], [15], [16]

En base a los trabajos encontrados se puede observar que, en general, los datos analizados provienen de organismos públicos de EEUU, y el objetivo principal que persiguen es reducir accidentes aéreos. Por ello analizan fundamentalmente datos relativos a accidentes, tráfico aéreo y mensajes de alerta a la tripulación. Sólo un artículo analiza datos de mercado, y ninguno incluye análisis de datos provenientes de redes sociales.

Considerando la cantidad de fuentes usadas, la cantidad de trabajos se distribuyeron equitativamente, entre una y más de una fuente. Por otra parte, las muestras analizadas mayormente abarcan datos de más de un año. No se ha podido analizar otra medida relativa al tamaño de la muestra dado que, en general, no se mencionan en los trabajos. Inicialmente se concibió analizar el tipo de datos fuente utilizado en las publicaciones; sin embargo, esa característica no era informada en las evidencias encontradas.

4.2 RQ2. ¿Qué algoritmos, métodos o técnicas fueron utilizados? (Se listan según las tareas que realizan cada proceso o técnica)

Tabla 12. Hallazgos según técnicas o métodos utilizados

Tarea	Método / Técnica	Cant.hallazgos	Referencias
Descubrimiento de grupos, patrones y reglas. Modelizado Descriptivo y/o predictivo	Clusterización	1	[8]
	Modelos generadores de reglas	2	[8], [14],
	Proceso de jerarquía analítica	2	[9], [16]
Modelizado predictivo o prescriptivo	Clasificador mediante consulta a expertos	1	[16]
	Métodos estadísticos y numéricos, de regresión	5	[7], [8], [10], [11], [16]
	Árboles de Decisión	5	[11], [12], [13], [14], [15]
	Generadores de reglas	2	[8], [13]
	Basados en redes neuronales	2	[7], [15]

Como se puede observar en la tabla 12, los métodos y técnicas más empleados resultan de aplicar métodos estadísticos y numéricos de regresión (5 trabajos), como también árboles de decisión (5 artículos) para plantear modelos predictivos. Les siguen la aplicación de modelización por medio de minería de reglas de asociación, de redes neuronales, mediante el descubrimiento de patrones utilizando reglas de asociación o mediante el proceso de jerarquía analítica, estando presentes en dos (2) artículos cada uno.

Se advierte que el tema del pre-procesamiento y tratamiento de la Información de grandes volúmenes de datos de diferente naturaleza, no es abordado. Esto es, los artícu-

los no mencionan del empleo de técnicas especiales para filtrado, detección y eliminación de espurios, el rellenado de datos ausentes, transformación de datos y/o reducción de la dimensionalidad, que son propias de minería de datos.

4.3 RQ3. ¿Cuáles son los tipos de resultado que presentan los estudios encontrados?

Tabla 13. Hallazgos según tipo de resultado

Tipo de resultado	Cant. hallazgos	Referencias
Descriptivo	1	[16]
Predictivo	9	[7], [8], [10], [11], [12], [13], [14], [15], [16]
Prescriptivo	2	[9], [16]

En los hallazgos mostrados en la tabla 13 se observan resultados que van más allá de lo puramente descriptivo, mediante la aplicación de técnicas exploratorias. La mayoría de los estudios presentan resultados predictivos (9) y algunos prescriptivos (2). Esto manifiesta que se centran en problemas de estudio que atañen a BA.

4.4 RQ4. ¿Presentan resultados de aplicar técnicas o recursos para la descripción o visualización de la información? ¿cuáles?

Tabla 14. Hallazgos según técnicas o recursos de visualización

Técnicas / Recursos	Descriptor	Cant.hallazgos	Referencias
Descriptor estadístico / Tabla resumen	Tablas	10	[7], [8], [9], [10], [11], [12], [13], [14], [15], [16]
Técnicas simples de visualización	Gráfico de barras	4	[7], [8], [10], [11]
	Gráfico de dispersión	1	[11]
	Diagrama de caja y bigote	1	[11]
	Gráfico de líneas	3	[7], [8], [15]
Técnicas de visualización multivariable	Gráfico de superficie 3D	1	[8]
	Mapa de calor	1	[10]
Otras técnicas de visualización	Curvas caract.de operador-receptor (curvasROC)	1	[10]
	Gráfico de estructuras jerárquicas (AHP) (árboles para toma de decisiones)	2	[9], [16]

Como se puede notar en la tabla 14, todos los hallazgos usan tablas para mostrar sus resúmenes estadísticos y otros descriptores. Las técnicas de visualización simple se utilizaron en 9 trabajos. Y en menor medida, las técnicas de visualización multivariable (2) y otras técnicas de visualización (3).

4.5 RQ5. ¿Especifican alguna herramienta / software de análisis y procesamiento de los datos? ¿cuáles?

Tabla 15. Hallazgos según herramienta o software de análisis y procesamiento

Herramienta / software	Cant. hallazgos	Referencias
Scikit - Learn Machine Learning Toolbox	1	[10]
Python	1	[11]
R	1	[11]
STATISTICA by StatSoft.	1	[15]
Microsoft Azure Machine Learning Studio	1	[11]
Expert Choice software	1	[16]
SPSS Clementine 10.1	1	[12]
WEKA	2	[13], [14]
Sin especificar	3	[7], [8], [9]

De los hallazgos considerados en la tabla 15 y teniendo en cuenta detalles de las herramientas o software empleados, se puede observar que 7 (siete) de las publicaciones especifican al menos una herramienta o software de análisis y procesamiento de los datos. De ellos, 2 (dos) utilizan WEKA, y 5 (cinco) emplean paquetes específicos de procesamiento estadístico. Además, hay 3 (tres) artículos en los que no se indica explícitamente las herramientas o software utilizados durante la investigación.

5 Conclusiones

Al comenzar el trabajo, cuando se hicieron las búsquedas preliminares con el fin de descubrir vocablos, conceptos, sinónimos, etc., se vislumbró que había poca bibliografía científica relativa a la temática de estudio en el dominio elegido. De hecho, esto influyó en la definición de las cadenas de búsqueda, mostradas en la sección 2.3. La revisión literaria realizada permitió corroborar esa percepción original. De hecho, no se encontró ningún trabajo científico totalmente inherente a la temática propuesta. Consecuentemente, se puede afirmar que se ha encontrado un nicho de investigación interesante y por ende, justificaría emprender una investigación profunda y original. Esta situación también desencadena interrogantes: ¿por qué hay tan poco publicado?, ¿por qué no se investiga?, ¿qué pasa con los datos, no se registran?, ¿no son de fácil acceso?, ¿se está buscando en repositorios incorrectos?.

Sin embargo se encontraron trabajos muy interesantes que emplean técnicas y herramientas de BI y de BA en problemáticas aeronáuticas, fundamentalmente relacionados a la prevención de accidentes. Varios de ellos analizan datos públicos, correspondientes a distintas aerolíneas, provenientes del Sistema de Datos de Accidentes / Incidentes de la Administración Federal de Aviación FAA (Federal Aviation Administration) de los EEUU. Estos hallazgos fueron considerados en esta revisión para poder dar sustento al presente trabajo y poder dar respuesta a las preguntas de investigación planteadas.

Referencias

[1] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Inf. Softw. Technol.*, vol. 64, pp. 1–18, 2015.

[2] B. Kitchenham and S. Charters, *Guidelines for performing Systematic Literature Reviews in Software Engineering*. 2007.

[3] Scopus, "Scopus - bibliographic database (ELSEVIER)," *www.scopus.com*. [Online]. Available: www.scopus.com.

[4] "ScienceDirect.com | Science, health and medical journals, full text articles and books." [Online]. Available: https://www.sciencedirect.com/. [Accessed: 09-Apr-2020].

[5] IEEE, "IEEE Xplore Digital Library," *http://ieeexplore.ieee.org*. [Online]. Available: http://ieeexplore.ieee.org.

[6] ACM, "ACM Digital Library," *http://dl.acm.org*. [Online]. Available: http://dl.acm.org.

[7] V. M. Janakiraman, "Explaining aviation safety incidents using deep temporal multiple instance learning," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 406–415.

[8] B. An, H. Chen, N. Park, and V. S. Subrahmanian, "MAP: Frequency-based maximization of airline profits based on an ensemble forecasting approach," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, vol. 13-17-August-2016, pp. 421–430.

[9] A. Shanmugam and T. Paul Robert, "Ranking of aircraft maintenance organization based on human factor performance," *Comput. Ind. Eng.*, vol. 88, pp. 410–416. 2015.

[10] N. Chakrabarty, "A data mining approach to flight arrival delay prediction for American airlines," in *IEMECON 2019 - 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference*, 2019, pp. 102–107.

[11] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, "A statistical approach to predict flight delay using gradient boosted decision tree," in *ICCIDS 2017 - International Conference on Computational Intelligence in Data Science, Proceedings*, 2018, vol. 2018-January, pp. 1–5.

[12] A. Lukáčová, F. Babič, and J. Paralič, "Building the prediction model from the aviation incident data," in *SAMI 2014 - IEEE 12th International Symposium on Applied Machine Intelligence and Informatics, Proceedings*, 2014, pp. 365–369.

[13] A. B. A. Christopher and S. A. Alias Balamurugan, "Data mining approaches for aircraft accidents prediction: An empirical study on Turkey airline," in *2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology, ICE-CCN 2013*, 2013, pp. 739–745.

[14] P. Barros Soares, O. Mendes Silva Junior, R. A. Souza Fernandes, and J. Henrique Bidinotto, "Data Mining-Based Analysis of Alert Messages of Executive Aircraft," *IEEE Lat. Am. Trans.*, vol. 16, no. 10, pp. 2561–2566, Oct. 2018.

[15] A. Stolzer, "Data Mining for Aviation Safety," in *Handbook of Statistical Analysis and Data Mining Applications*, Elsevier, 2009, pp. 377–389.

[16] S. A. Delbari, S. I. Ng, Y. A. Aziz, and J. A. Ho, "An investigation of key competitiveness indicators and drivers of full-service airlines using Delphi and AHP techniques," *J. Air Transp. Manag.*, vol. 52, pp. 23–34, Apr. 2016.

Ontology of Cardiovascular and Respiratory Diseases

William da Rosa Fröhlich[1], Sandro José Rigo[1], Clarissa Almeida Rodrigue[1], María Eugenia Ruiz[2]

[1]University of Vale do Rio dos Sinos (UNISINOS)
[2]National University of Córdoba
william_r_f@hotmail.com | rigo@unisinos.br | clarissa.ar@gmail.com | eugenia.ruiz@unc.edu.ar

Abstract. The medical diagnoses are essential in the work of health professionals, but often the task of collecting information for diagnosis demands a long time and it can be performed by others professionals that work in the health center. In this context, this article aims to present the building of an ontology designed to serve as a knowledge-base, relating diseases and their symptoms, for that is made a brief description the current scenario and an overview of others related works. In sequence, a careful description of how the ontology is structured, finally, an explanation of the functioning of the ontology with its diseases and related symptoms. Using this knowledge-base, it is possible apply to developing a diagnostic system, with focusing in a friendly and simple interface to use, so that the diagnostic process can be performed by any person previous trained, with the objective of assisting the health professional and make the screening process more dynamic.

Keywords: Ontology, Medical Diagnoses, Knowledge-Base.

1 Introduction

In the current scenario of the health centers, combining the easy access to technology and advances in the study of intelligent systems, the use of artificial intelligence applied to health has grown. Much researches emphasize the need to use computing in medicine and public health. Artificial Intelligence in medicine has the function of analyzing a large volume of data, following algorithms defined by specialists in the field, with the propose to solve simple, or even advanced, medical problems [10]. In the past, the nurses assumed several roles that nowadays are delegated to other professionals, allowing nurses to focus on more specific functions. Changes are common in several fields, such as social, scientific, technological and educational, although they have brought some limitations to the exercise of the profession [12].

Lopes and Higa [13] highlight that for nurses to make a clinical judgment, they need to use cognitive skills and knowledge, in addition to performing adequate data collection of the basic condition so that the evaluation is possible. Based on this, several tools emerged to facilitate the process and, among them, computers represented a

fundamental resource that can be used during the data collection and organization. From the utilization of the computers, information retrieval time is decreased and the memory increased in capability, thus enabling a broader diagnosis, taking into account more details, and in less time [10]. In view of the facts presented, the proposal arose for the elaboration of an ontology focused on relating the symptoms, from the most basic to the most advanced, of cardiovascular and respiratory diseases.

The objective of the present work is to elaborate an ontology that serves as a knowledge-base for the development of an intelligent diagnostic system, so that it can be operated by any trained professional of the health center, capable of asking questions to the patient. The ontology lists all the symptoms and lifestyle that are related to heart or respiratory diseases in order to make a more accurate diagnosis. This article is divided in others five topics, the second part is made a brief description of the general concepts about ontologies, in sequence are presented some related works, describing applications in the health area. In the fourth and fifth topics, respectively, describe building of the ontology and the reasoning applied, as well as, the analysis of the ontology. Finally, it is presented the conclusion about the article and some suggestions to next works.

2 General Concepts

In this session, it will be presented basic concepts about ontologies, for a better understanding of the development process used. According to Almeida [2], there are several techniques available nowadays that can be used as a way to organize information and knowledge, with ontologies being one of the main ones. This is a subject of study in many fields, such as Philosophy, Computer Science and Information Science. In the area of computing, ontologies are part of the field of artificial intelligence, more specifically called knowledge representation, referencing a structure of concepts represented by a logical vocabulary. The term ontology [12] has several definitions, often contradictory depending on the author.

Ontologies work in order to classify hierarchical structures into branches of classes and subclasses, in which the highest level is called an entity or individual. The presence of context in the set of technologies is referred to as the Semantic Web, which consists of a combination of facts and rules to compose a knowledge-base [12]. As explained by Ramalho [13], ontologies transform contents with little, or no standard, into a standardized structure, through languages that can process this data computationally, however, due to the different possible combinations in structured languages.

There are restrictions and axioms to enable concepts, relationships and properties of domain, in which, axioms are responsible for providing the facts through rules [16]. Silva [15] details the basic components, with the lowest level component in the hierarchy being the classes, also called concepts, which can be used to describe a task, function, action or a reasoning process, in turn the taxonomy is used to organize the classes.

Other important components, fundamental in ontologies, are the relationships and functions, with the objective of establishing relationships between classes and other basic components, relating how a component belongs to the domain and other attributes.

Silva [15] describes axioms as a way to model true sentences, and may impose restrictions, corrections and deductions. Finally, the instances, also known as entities, have attributes that are relevant properties and instances of a concept, being the highest level of the hierarchy assembled by ontologies. There are many languages [14] that can be used to build ontologies, but the one that stands out is the Web Ontology Language (OWL), recommended by the World Wide Web Consortium (W3C), the main standardization organization of the World Wide Web, an international consortium with 450 members to establish standards for the creation and interpretation of content for the Web.

Roque [14] further defines that OWL is projected as a complex and rich form of knowledge representation based on computational logic. Its main objective is to standardize ontologies as a way to facilitate sharing among specialists in the field of interest, allowing reuse. The development of ontologies can be divided in seven stages [12], the first stage is to define the scope of the ontology, determining which area of knowledge will be served, has the purpose of answering which doubts and which people will benefit from the development. In the sequence, it is important to consider the reuse of another existing knowledge-base, as it is often possible to refine or expand an ontology already developed in part, very important also when it is necessary to interact with other applications.

In the third step, must be evaluated in the development is the number of important terms, listing all the terms that will be declared and what is their relevance to the project. Another important step is to define the classes, as a way of structuring the hierarchy, which can be: 1 – Top-down (from top to bottom) starting with defining basic concepts and starting with the most specific ones; 2 – Bottom-up (starting from the bottom up) starting with the definition of the most specific classes and directing them towards the classes in more general concepts. The fifth step is made to define which connector properties will be applied, in function of the classes alone are not capable of providing information, for this reason, the connector properties are used, with the objective of interconnecting the classes, attaching one to the other [12].

Noy and McGuinness [12] define that the sixth stage is responsible for establishing the characteristics of the connectors, which may have different characteristics with different types of values, which may be Boolean, string, integers or any allowed values. As a last step, it is important to establish which are the individual instances, responsible for representing the set of elements of the classes.

3 Literature Review

In this session, some current works developed using ontologies applied to health will be presented, which can serve to understand how ontologies help in this area. González [8] presents the development of an ontology with approximately 13,000 axioms, based on studies of human genomes, with the objective of making the diagnosis of rare diseases. With the study carried out, the analyzes confirmed the relationship and the possibility of making diagnoses in this model.

The health care for the population is a constant challenge in public management [17], especially when referring to the integration of different systems for a continuous care system. The final result of the ontology presents a method of systematizing problems of epistemological overlap and the intermediate model are direct contributions of the research, also it is evident that clinical terminologies are not always prepared for use in the digital context and the Semantic Web.

Another interesting project developed using ontologies deals with electronic data, used in hospitals, which are responsible for managing tasks and alerting patients, in this context, an ontology for an alert system was developed. The developed ontology is responsible for identifying, evaluating and triggering alerts for health professionals and the results obtained evaluated by health professionals indicate that the use of ontology contributes to the definition and issuance of alerts at times, however, the quality of the hardware directly affects the performance of the system [7].

Finke, Filice and Kahn [6] present the development of an ontology that relates the diagnosis of human diseases, phenotypes and radiological diagnosis. The developed project confirmed the possibility of using mappings that can be used to support the reasoning of automated diagnosis, data mining and knowledge discovery. Some literatures [1, 16] point out the importance of the advanced in the smart systems applied to the health, especially when aligned with internet of things (IoT). Akhtar et al. [1] presented a system bringing together ontologies and the internet of things, with the goal to develop an intelligent decision support system, a smart device measures the signals, in sequence is made a previous diagnosis and allows automatically sharing the patient's critical status with their doctor.

Alsubaei, Abuhussein and Shiva [3] use a different term to define the IoT applied to health, Internet of Medical Things (IoMT), in their article, it is presented a security recommendation tool, focusing mainly in remains the patient data confidential and secure. The ontology that was developed has context sensitive rules to allow reasoning, thereby building a recommendation system. The tool can classify threats and automatically recommends the actions for the security control required for each threat

Kim and Lee [9] developed a system for diagnosis using an ontology and point out that an intelligent health diagnosis technique may provide reliable diagnostic to users, using a technique automatically generates, as Human Disease Diagnosis Ontology. The

experimental results of their work showed a good accuracy over existing symptom checkers and related approaches, with the goal to provide an appropriate healthcare action. Based in these works presented, it is possible state that the ontologies applied to health have been widely used, in different ways, either using the internet of things or automating processes within health centers.

4 Knowledge-base

In this session, will be presented the stages of the development of ontology. The ontology editor and knowledge manager Protégé was chosen for the development of the ontology, due to the easy and agile graphical interface, widely used for this purpose. The language chosen is OWL, even though it has more complexity in development, as a knowledge-base purpose, it is a more complete language for the project context.

Following the recommendations of Noy and McGuinness [13] for the development of an ontology, the first step is to determine the scope of the project, with this, it is defined that the purpose of the work is to develop a knowledge-base that relates symptoms and pre-existing diseases with cardiovascular and respiratory diseases, in order to provide an intelligence base for the development of a software interface that can be made available to health professionals to facilitate the process of diagnosing diseases of these origins.

The second part, of the seven stages, of the development of an ontology, consists in analyzing other existing ontologies, however, when it was analyzed the knowledge-bases available in the NCBO BioPortal repository, none were found that were within the defined scope, all the available ones had different objectives or out of the defined focus. However, when analyzing the structures of ontologies, part of the idea of how to structure hierarchies and entities, as well as defining the properties responsible for establishing relationships that was used.

The third step of the project was to perform the list of terms that are applied in the ontology, in the context of the developed project, in this stage it is defined which diseases will be present in the knowledge-base and relating the symptoms for each of the comorbidities. The relationship pointed out 47 symptoms, six physiological characteristics and three habits, which through their combinations of symptoms, present 21 possible diagnoses.

Of the possible diagnoses, there are twelve cardiovascular diseases, namely angina, arterial fibrillation, atrial flutter, bradycardia, Brugada syndrome, cardiac insufficiency, endocarditis, murmur, long QT syndrome, myocardial infarction, tachycardia and ventricular extrasystole [3, 5], and nine respiratory diseases, such as asthma, bronchitis, flu, pharyngitis, pneumonia, cold, rhinitis, sinusitis and tuberculosis [4].

In the next step, after defining the terms, using the top-down process, first defining the basic classes and then defining the specific classes, the basic classes chosen were Diseases, Lifestyle, Physiological and Symptoms, adding within those classes the previously defined terms, as shown in Fig. 1.

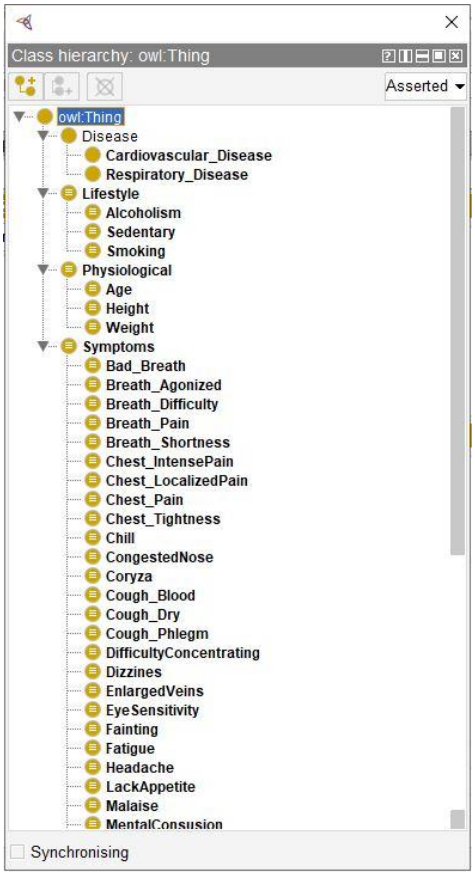


Fig. 1. Class Hierarchy.

In the fifth stage of the project, the connector properties are defined, responsible for making the relationship between classes and instances, the properties were created following the idea of establishing a relationship of "is..." and "is... of", as shown in Fig. 2. For step six, data properties were defined, created mostly as Boolean variables, as they are analyzed with reference to the presence or absence of a certain symptom.

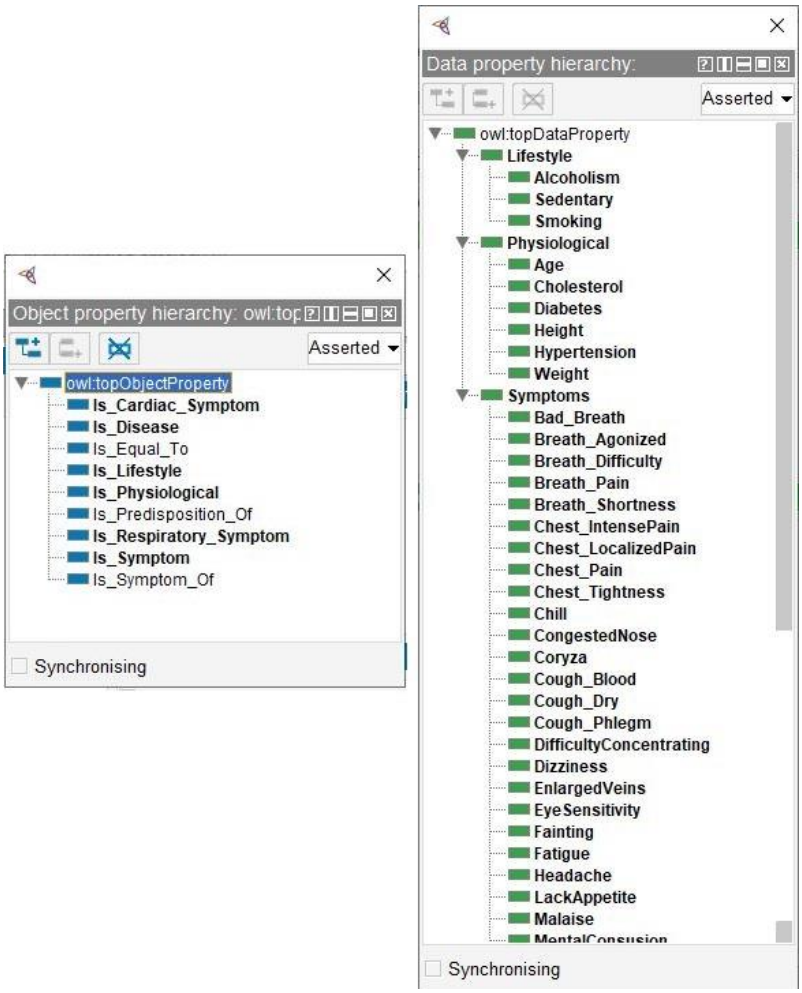


Fig. 2. Relationship of Object Ownership and Data Ownership.

Finally, as the last stage of the project, the instances that are analyzed are established, in the context of the developed ontology, the instances are the point of interest of the analysis, the diseases. The instances created were divided into two main classes, separated between cardiovascular and respiratory diseases (Fig. 3).

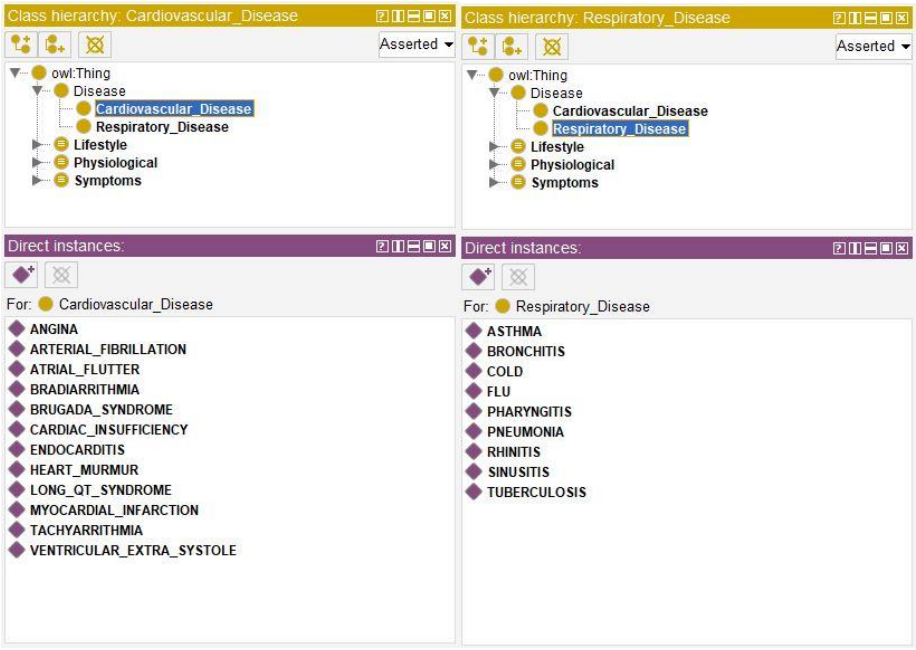


Fig. 3. Instances.

5 Ontology Analysis

The developed ontology presents a knowledge-base of cardiovascular and respiratory diseases and the respective symptoms. As a way of making it as complete as possible, all classes and instances were duly related with the aim of making the diagnosis as accurate as possible. Each instance is related to a class by the connector properties specific to its own type of class and as a way of defining each type of data.

The data property for each symptom is also defined and listed in the instance itself (Fig. 4). As a way of making the knowledge-base complete, the symptoms are defined both as a class, to provide a relationship with the other classes, as a data property, in order to assign the necessary value for the diagnosis of the disease to be evaluated.

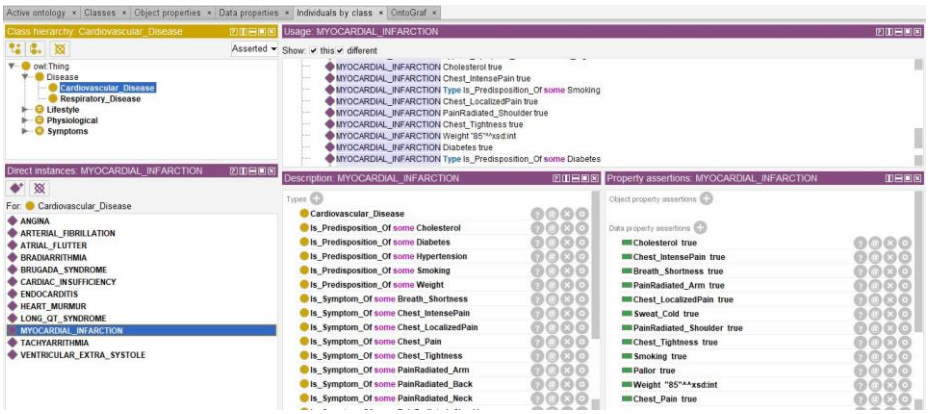


Fig. 4. Example of the Relationships Established for the Instance “Myocardial Infarction”.

As a way to carry out the ontology validation, the function “reasoner” is used to check the classes and structure of the developed knowledge-base, which did not return alerts or errors. When accessing the Individuals tab, it is possible to check the presence of inferences made by the software itself about information that was not represented explicitly.

6 Conclusion

The ontology developed aims to provide an intelligent database for the development of an interface for interaction with the user, helping the health professionals, as a way to automate and streamline the diagnostic process, considering all possible diagnoses related to symptoms cited by the patient. By the query tests performed, with the SPARQL language, it is possible to verify that the knowledge-base is efficient, the data access was satisfactory and with the possibility of developing an application and interface.

As a proposal for the continuation of this work, it is the elaboration of an interface of interaction with the user that can use the developed ontology. The idea of the project sequence is the implementation using the Python programming language, so that it can be implemented and tested, for possible necessary adjustments. Another point of continuation for this work is the expansion of the knowledge-base, adding others symptoms and diseases, such as hemorrhagic fevers and respiratory syndromes, thus also expanding the number of classes and instances.

References

1. Akhtar, S. M., Nazir, M., Saleem K., Haque, H. M. U and Hussain, I.: An Ontology-Driven IoT based Healthcare Formalism. In: International Journal of Advanced Computer Science and Applications. Vol. 11, No. 2, (2020).
2. Almeida, M. B.: Uma abordagem integrada sobre ontologias: Ciência da Informação, Ciência da Computação e Filosofia. In: Perspectivas em Ciência da Informação, Vol. 19, No. 3, P. 242-258, (2014). DOI: doi.org/10.1590/1981-5344/1736
3. Anderson, K. M. et al.: Cardiovascular disease risk profiles. In: American Heart Journal. USA. Vol. 121, No. 2. Pg. 293-298, (1991). DOI: doi.org/10.1016/0002-8703(91)90861-B
4. Bousquet J., Dahl R., Khaltaev N.: Global Alliance against Chronic Respiratory Diseases. In: European Respiratory Journal. Switzerland. Vol. 29. Pg. 233-239, (2007). DOI: doi.org/10.1111/j.1398-9995.2007.01307.x
5. Einstein, H. A.: Cardiologia - Doenças e Sintomas. In: Hospital Albert Einstein, Sociedade Beneficente Israelita Brasileira. https://www.einstein.br/especialidades/cardiologia/doencas-sintomas/, last accessed 2020/08/13
6. Finke, M. T., Filice, R. W. and Kahn, C. E.: Integrating ontologies of human diseases, phenotypes, and radiological diagnosis. In: Journal of the American Medical Informatics Association, 26(2), Pg. 149–154, (2019). DOI: 10.1093/jamia/ocy161
7. Freitas, A. L. S., Freitas, A. L. S. e Teixeira, H. M. P.: Sistema de Alerta Baseado em Ontologia para Lousa Eletrônica em um Hospital Público, Revista Brasileira de Computação Aplicada. 11, 3 (set. 2019), 99-109 (2019). DOI: https://doi.org/10.5335/rbca.v11i3.8552.
8. González, J. A. M.: Aplicación de técnicas de NLP y Mapeo ontológico en el Ámbito del Diagnóstico Clínico. In: Escuela Politécnica Superior, Universidad Autónoma de Madrid. Es. (2019).
9. Kim, G. and Lee, D.: Intelligent Health Diagnosis Technique Exploiting Automatic Ontology Generation and Web-Based Personal Health Record Services. In: 2019 IEEE. Vol. 7. Pg. 9419-9444. (2019). DOI: 10.1109/ACCESS.2019.2891710
10. Lobo, L. C.: Inteligência Artificial e Medicina. In: Revista Brasileira de Educação Médica. Brasília. Vol. 41, No. 2 (2017). DOI: doi.org/10.1590/1981-52712015v41n2esp
11. Lobo, L. C.: Inteligência artificial, o Futuro da Medicina e a Educação Médica. In: Revista Brasileira de Educação Médica. Brasília. Vol. 42, No. 3 (2018). DOI: dx.doi.org/10.1590/1981-52712015v42n3RB20180115EDITORIAL1.
12. Lopes, M. H. B. M., Higa, S.: Desenvolvimento de um sistema especialista para identificação de diagnósticos de enfermagem relacionados com a eliminação urinária. In: Revista Brasileira de Enfermagem, Vol. 58, No. 1 (2005). DOI: doi.org/10.1590/S0034-71672005000100005.
13. Noy, N. F., McGuinness, D. L.: Ontology development 101: A guide to creating your first ontology. In: Stanford Knowledge Systems Laboratory. (2001).
14. Ramalho, R. A. S.: Ontologias como instrumento de representação: reflexões interdisciplinares. In: I EnReDo – Encontro de Representação Documental. São Paulo, Brasil. (2017) http://www.telescopium.ufscar.br/index.php/enredo/enredo/paper/viewFile/98/93, last accessed 2020/08/13
15. Roque, B. A. S.: Desenvolvimento de uma Ontologia Especifica para o Cenário de Inovação do Estado de São Paulo. In: Centro Universitário Eurípides de Marília. São Paulo, Brasil. (2017). http://hdl.handle.net/11077/1644, last accessed 2020/08/14
16. Sharma, N. and Aggarwal, R. K.: Ontology as a Tool to Enable Health Internet of Things Viable 5G Communication Networks. In: Scrivener Publishing LLC. (2020). https://doi.org/10.1002/9781119641391.ch14

17. Silva, A. R. L. et al.: Ontologia como Representação do Conhecimento: aplicação no curso de Formação Continuada em Tecnologias Educacionais na Web. In: Revista Novas Tecnologias na Educação. Porto Alegre, RS. Vol. 12, No. 1 (2014). DOI: doi.org/10.22456/1679-1916.50274
18. Teixeira, L. M. D., Almeida, M. B.: Princípios ontológicos no suporte a terminologias clínicas: método e ontologia para reorganização da Classificação Internacional de Doenças. In: Ciência da Informação, Brasília, DF, Vol. 48 No. 1, p.94-112. (2019) http://revista.ibict.br/ciinf/article/view/4291/4101

Consideraciones Éticas de los Datos en Business Intelligence

María del Carmen Becerra[1,2], María Romagnano[1,2], Alicia Aballay[1,2], Adriana Sarmiento[1,2]

[1]Universidad Nacional de San Juan. FCEF y N

[2] Proyecto: "Aporte de los Sistemas de Inteligencia de Negocios a los Sistemas de Información Organizacionales para la Toma de Decisiones". Resol.N°0591-20-R-CICITCA

mcbecerra2008@gmail.com, prof.alicia@gmail.com, maritaroma@iinfo.unsj.edu.ar, adriva2005@yahoo.com.ar

Abstract. Este trabajo analiza la inclusión de la ética de datos como una estrategia para lograr una gestión eficiente del manejo de los mismos en el contexto de Business Intelligence (BI), todo ello estructurado de manera crítica en todo el ciclo de vida de los datos desde su recolección, análisis, gestión y posteriores usos. Se afirma hoy más que nunca, la prudencia es inteligencia y el camino para la eficacia en la toma de decisiones de las organizaciones. Se combina la experiencia de diferentes áreas de los integrantes del proyecto para detectar los problemas éticos que surgen por la aplicación de BI y se elabora una estrategia jurídica y tecnológica interdisciplinar para satisfacer las expectativas de desarrollo de los individuos, las empresas y la sociedad en general. Se trata de identificar y aplicar los principios éticos para orientar la tarea de los profesionales de datos, como sujetos que asumen su responsabilidad y deben desarrollar prácticas para responsabilizarse a sí mismos y a sus compañeros por los estándares compartidos. Todo ello inmerso en una cultura organizacional orientada hacia una revisión ética, consistente, eficiente y procesable cuya transparencia en el tratamiento de datos no solo minimiza los riesgos, sino que contribuye a la confianza pública.

Keywords: Reflexión ética. Problemas éticos de BI. Responsabilidad Social. Estrategias técnico jurídicas

1 INTRODUCCIÓN

Actualmente las organizaciones necesitan manejar grandes volúmenes de datos para la toma de decisiones e incluso para sobrevivir en el mundo de los negocios. Existe un espacio común de datos que demandan una estrategia de desarrollo, se enfrentan con problemas jurídicos y de seguridad, por lo que se impone una reflexión sobre la ética de la cultura de los datos. Definida de manera simple, la ética son los principios de comportamiento basados en ideas de lo correcto y lo incorrecto. Estos se basan en conceptos tales como la equidad, el respeto, la responsabilidad, la integridad, la calidad, la confiabilidad, la transparencia y la confianza. La ética en el tratamiento de los datos tiene que ver con la forma en la que obtenemos, almacenamos, administramos, usamos y disponemos de los datos de manera que estén alineados con estos principios. Se torna indispensable llevar un control de datos, definiendo una cultura organizacional basada en principios éticos donde se identifique al sujeto moralmente responsable.

En este trabajo se parte de una reflexión ética sobre el manejo de los datos que tratan las organizaciones y que debería estar presente en todo el ciclo de vida del dato: captación, gestión, privacidad y uso. Ello debe estar implícito en (BI) [1], porque la

Ética de los Negocios [2], ha adquirido especiales caracterizaciones, dado que su objetivo no es otro que el de dotar a la empresa de eficacia en la toma de decisiones estratégicas. La ética de datos no se restringe a la recopilación o la gobernanza de estos, sino que también incluye la forma en que se interpretan los datos y se realizan las acciones dentro de la organización para alcanzar un mayor éxito en los resultados. BI es el conjunto de técnicas, métodos, estrategias y herramientas que permiten utilizar los datos y la información generada a partir de ellos, correspondientes al período deseado (mes, trimestre, semestre, año, etc.) y determinar en base a ella el estado actual de la organización, respecto a sus clientes, competidores, proveedores, el mercado en sí, etc.; y en base a ello tomar una decisión. Estos datos e información son de vital importancia para la organización y la posibilidad de convertirlos en conocimiento para el proceso de toma de decisiones de una manera informada y aceptada, lo que es un gran beneficio.

Las plataformas de BI modernas han simplificado el análisis de datos para muchas personas, y hay más roles en el tratamiento de datos personales que deberían respetar los principios éticos de los datos [3].Además, se puede considerar también como una gran ventaja competitiva de la organización respecto al resto de organizaciones que no aplican BI y, por ende, toman decisiones basadas en opiniones personales o ideas, sin tener un respaldo en concreto [4]; es un medio para optimizar la gestión de la información, eje fundamental de los procesos de negocio actuales. Los elementos en que se sustenta la conceptualización de inteligencia de negocios son los sistemas de información, los mecanismos de innovación y los procesos de toma de decisiones. En cada uno de ellos se implementan estrategias que pueden llevar a la organización a adquirir conocimiento y a mejorar la manera en que este incrementa el valor de los productos y servicios que se ofrecen [5]. Así como lo afirma (Rodríguez, Miguel Ángel) "Un aspecto importante de la comprensión actual de la ética se refiere a cuestiones de responsabilidad individual y social con respecto al impacto de nuestras elecciones a la luz de la influencia de la ciencia y la tecnología. Si bien las tecnologías de la información y la comunicación abren las puertas a nuevas posibilidades tecnológicas y científicas, también actúan como catalizador de un encuentro sin precedentes con la alteridad, asegurando a través de los medios digitales la colisión masiva de sistemas éticos y cosmovisiones culturales hasta ahora cerrados "[6].

En este trabajo nos centraremos en los problemas éticos que mundialmente deben afrontar los profesionales de datos en un contexto de BI, para poder evaluar si los hechos están presentados de manera clara o no y si los límites de los datos han sido comprendidos de manera correcta y son adecuados para la pregunta o no. A medida que va aumentando el número de personas que trabajan en datos, la ética de los datos se vuelve una pieza fundamental en las iniciativas relacionadas con el conocimiento de los mismos y cambia la manera en que se abordan los datos en contextos personales, profesionales y sociales.

2 OBJETIVOS

Los objetivos de este trabajo, hacen referencia a lineamientos generales del proyecto de investigación "Aporte de los Sistemas de Inteligencia de Negocios a los Sistemas de Información Organizacionales para la toma de decisiones", y se dividen en:

1. **Objetivo General**: Determinar la aplicabilidad de los principios éticos en el manejo de los datos en el contexto de BI.
2. **Objetivos Específicos o Particulares**:
   - Especificar qué tipo de problemas éticos, legales y de seguridad, surgen como consecuencia de la aplicación y desarrollo de BI y BA en los sistemas de información.
   - Identificar al sujeto moral responsable en cuanto a la toma de decisiones en un contexto individual, empresarial y social.
   - Dejar planteada una reflexión ética para el uso coherente de datos por parte de los Profesionales de datos
   - Definir estrategias técnico-jurídicas en los Sistemas de Información Organizacionales basadas en principios éticos para la toma de decisiones en BI.

3 REFLEXIÓN ÉTICA

La ética empresarial se refiere a una reflexión sobre los valores que encierra una decisión. Para la European Business Ethics Network (EBEN), se trata de una reflexión sobre las prácticas de negocios que implican normas y valores de los individuos, de la empresa y de la sociedad. "El comportamiento de una organización debería basarse en los valores de honestidad, equidad e integridad. Estos valores implican la preocupación por las personas, animales y medio ambiente, y un compromiso de tratar el impacto de sus actividades y decisiones en los intereses de las partes interesadas" (ISO 26000, 2010) [7]. La persona actúa conforme a la ética siempre y cuando elija el bien. En la empresa no es sino la aplicación concreta de los principios objetivos de la ética, para que quienes trabajan en el mundo de los intercambios económicos, adquieran los hábitos operativos adecuados que les faciliten alcanzar la felicidad, meta última de cualquier persona. (Fernández, Aguado, p.42) [8]. Considerar a las personas como fines, no como recursos, apuesta en serio por la confianza como principio esencial y propone la consecución de un equilibrio sinérgico entre valores generadores de salud económica, ética y emocional (Salvador García, y Shimon Dolan) [9].

Como se muestra en la Fig.1, desde antiguo se planteó una taxonomía sobre la reflexión Ética, Colmarejo Fernández [10] menciona la ética de los negocios como un movimiento mundial por el cual las corporaciones ofrecen autorregularse. La ética de la información empresarial aborda las consideraciones informativas de la difusión de bienes y servicios, incluida la información como una mercancía, y proporciona orientación ética en el análisis del uso de bienes y servicios, incluido el discurso sobre el impacto que tienen en la sociedad. La Ética de la información empresarial también aborda las preocupaciones sobre la gestión de la información e incluye el sub-campo de ética de la información organizativa, representado por el Centro para la ética de la información empresarial (CBIE). La ciber-ética se ocupa de analizar los modos en que internet está condicionando la comunicación social entre individuos y organizaciones incluye a su vez la ética de las redes sociales, la ética de IoT e incluso la ética aplicada a la robótica. Se menciona también la ética de la computación y la ética de la información en investigación biométrica. Respecto a la ética de los sistemas de información, se la define como aquella que tiene por objeto la representación,

organización, clasificación, recuperación, disponibilidad, preservación, circulación y uso de la información en relación con el medio digital sistemas de información [11].

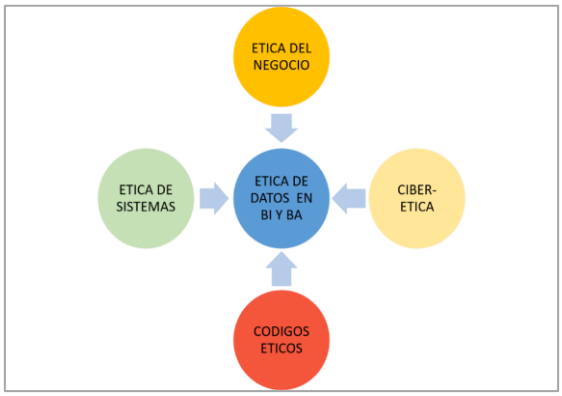


Fig. 1. Taxonomía de la reflexión ética aplicada a BI y BA (Fuente: Colmarejo Fernandez)

Conforme la taxonomía descrita, la ética aplicada a la gestión masiva de datos tiene su origen en la ética de los negocios, de ella también deviene la Ética aplicada a BI. A medida que las fuentes de datos se vuelven más complejas, diversas y numerosas, la administración de datos es más importante que nunca en las implementaciones de BI modernas; las empresas recurren a la selección de datos para superar desafíos relacionados con la administración de datos y la gobernanza que trae aparejado el acceso más generalizado a los datos. La selección de datos incluye la forma en que una organización captura, limpia, define y alinea datos diversos. Ese proceso crea un puente entre los datos y los usos en la vida real.

Las organizaciones ya están invirtiendo millones de dólares en tecnologías que integran las definiciones de datos con las herramientas de análisis para ayudar a analizar los datos y, así, eliminar las ambigüedades en los equipos y las instituciones. Como consecuencia, los procesos y las herramientas de selección de datos (cómo catálogos de datos y gobernanza semántica) convergen con las plataformas de BI para vincular datos con el contexto de negocios.

Cuando las organizaciones desarrollan sistemas de información, sus líderes tienen la responsabilidad adicional de ser buenos administradores de estos sistemas [12]. Establecer políticas para el acceso justo al sistema es una parte importante de la administración. Ésta requiere, además, que los profesionales encargados de la gestión informática monitoreen el nivel de integración de sus sistemas en la infraestructura de la sociedad. A medida que el nivel de adopción cambia, es probable que las responsabilidades éticas de la organización o grupo también cambien [13]. El monitoreo continuo de la forma en la cual la sociedad está usando un sistema permitirá que la organización sea consecuente con sus obligaciones éticas. Cuando no existen normas de cuidado apropiadas, los profesionales tienen el deber de garantizar que se desarrollen y se cumplan [14].

4 PROBLEMAS ETICOS

Los problemas que trae consigo la tecnología y el manejo masivo de datos, como nueva estrategia empresarial, no sólo han sido tratados por las normas de seguridad informática como los estándares tecnológicos que nos ofrece la familia ISO 27.001[15] sino que han determinado la aplicación de las leyes de protección de la vida privada y otras relacionadas con otras conductas disvaliosas como es la discriminación, pero la falta de compendios internacionales sobre protección de datos, determina que surjan múltiples problemas jurídicos y de seguridad.

Ante esta disyuntiva hay que revalorizar el papel de la ética que va más allá de la ley, porque BI no sólo dan lugar a nuevas formas de hacer negocios, sino que son un instrumento para el desarrollo sostenible. Los problemas éticos planteados por la tecnología no se circunscriben a la posibilidad de que ésta sea mal utilizada, sino que se convierta en un medio para que las personas logren fines moralmente condenables, en función de que como se expresó anteriormente, la tecnología no es responsable de los fines, no tiene autoconciencia y únicamente puede realizar aquello para lo que fue diseñada, evidenciando entonces que el único responsable de los usos, abusos y efectos de la tecnología es quien la crea y la utiliza[16]. En las últimas dos décadas, las organizaciones dedican una parte importante de su tiempo y de sus recursos económicos y humanos a la obtención, proceso, aplicación y proyección de información. Por esta razón, la información juega un papel decisivo en ellas y se convierte en su principal patrimonio intangible [17].

El problema se presenta cuando no se dispone de la información pertinente, necesaria para tomar las mejores decisiones, de una manera puntual [18]. Además, los directivos deben contar con herramientas adecuadas para la explotación y análisis de los datos que les permitan obtener el conocimiento necesario en el proceso de toma de decisiones estratégicas. Entonces, se podría pensar que el problema comienza cuando los ejecutivos no toman buenas decisiones debido a que no cuentan con información calificada y herramientas tecnológicas que les permitan la explotación de la misma para tomar decisiones bien fundamentadas e informadas. Por todo ello las Organizaciones deberían estructurar un buen Gobierno de datos, más allá del cumplimiento normativo, para ganarse la confianza de los consumidores la clave es la comunicación, la honestidad y transparencia.

Organizaciones de todo tipo y tamaño recolectan, almacenan, procesan e intercambian grandes volúmenes de información con el deseo de crear nuevos productos y servicios a partir de los datos; se busca generar mayores beneficios, reduciendo los costes, para consolidar y reforzar la fidelidad de los clientes y ofrecer auténticas ventajas competitivas. Pero al mismo tiempo, se enfrentan a importantes retos tales como gestionar adecuadamente los datos, garantizar su privacidad, o cumplir con regulaciones y estándares internos y externos [19]. El gran reto que enfrentan las organizaciones es conciliar las demandas de la sociedad con las expectativas internas, debiendo analizar la forma de conducir las personas y los procesos, para anticiparse a fenómenos de conducta no ética en el seno de las mismas.

Como se muestra en la fig.2, dichas tecnologías son producto del trabajo intelectual del ser humano, y son desarrolladas con propósitos benéficos para las organizaciones y la sociedad. Sin embargo, las personas tienen distintas intenciones y propósitos y

valiéndose de sus conocimientos y de las herramientas tecnológicas diseñadas para auxiliar a la gestión, pueden perseguir fines considerados moralmente condenables (Discriminación social, económica, de acceso a la información, abuso de control, fenómeno de dominio de la opinión pública, etc.). Lo anterior conduce a replantear la concepción ética y su instrumentación en los sistemas de gestión empresarial, para prevenir malas prácticas empresariales, que repercuten en la reputación de la empresa y en la consecuente pérdida de confianza, que constituyen uno de los problemas más graves en el mundo de negocios. La ética de la información comprende las cuestiones sobre ética que se relacionan con el proceso de la digitalización, la comunicación de sus resultados y la utilización adecuada de la información generada. Los códigos de ética de información tratan todo lo relacionado con el uso y mal uso de la información, tales como, la propiedad intelectual, acceso a la información libre o restringida, censura, uso de información de las instituciones gubernamentales, la intimidad y confidencialidad, integridad de los datos, flujo internacional de información, entre otros [20]. También, la ética profesional realiza el análisis de cómo se emplean los principios éticos en decisiones y acciones de los profesionales de la información.

Fig. 2. Problemas éticos en el manejo de datos en BI (Fuente de Elaboración Propia)

La ética del manejo de datos es compleja, pero se centra en conceptos tales como el impacto que tiene en las personas la calidad de los datos debido a que estos representan características de las personas y se utilizan para tomar decisiones que les afectan. El uso indebido de los datos afecta negativamente a las personas y a las organizaciones, por lo que es imperativo evitarlo; y finalmente los datos tienen valor económico per se, más allá del cumplimiento de leyes y de regulaciones internas y externas, los profesionales de la gestión y explotación de los datos deben reconocer que existen razones éticas para proteger la información y garantizar que no se utilice indebidamente.

Los pilares sobre los que debe articularse la confianza son la calidad y precisión de los datos, seguridad y protección de los mismos. Algunos de estos aspectos están regulados en el Reglamento Europeo de Protección de Datos (RGPD)que regula todos los eslabones y en este sentido la Ley de Protección de Datos Argentina regula las funciones del responsable del tratamiento de los datos personales y fija los principios fundamentales para la recolección de Datos personales como son; Licitud, calidad, pertinencia, confidencialidad y seguridad.

5 RESPONSABILIDAD SOCIAL DE LA EMPRESA

Actualmente, el concepto de responsabilidad social se entiende como sinónimo de compromiso de la empresa con todos los que tienen interés directo en su funcionamiento: accionistas, empleados, clientes, proveedores, competidores y las comunidades donde operan. Una sociedad con sólidos valores morales y éticos es la clave para minimizar la corrupción.

La responsabilidad social, que implica aplicar un enfoque basado en el respeto por los todos los sectores involucrados en la actividad de la organización, tiene en la ética un aliado estratégico, que supera el cumplimiento de la ley. La realización de eventos intersectoriales, pueden encender el interés por la aplicación de valores, que permitan beneficios mutuos por objetivos económicos y sociales comunes. De lo anterior se desprende, que los directivos de las organizaciones pueden erradicar la corrupción mediante planificación y control basados en estándares éticos sujetos a evaluación. Además, se debe tomar conciencia por parte de los miembros, del sentido y trascendencia del respeto a reglas éticas, elaborando y comunicando declaraciones con las creencias y valores imprescindibles para servir a la comunidad con su actividad presente y futura.

La responsabilidad de la empresa se concibe de forma acotada, no es con la sociedad en abstracto que la empresa se obliga, únicamente lo hace con la comunidad en la que participa de manera directa. Se precisa entonces, a propósito de la vinculación de los temas de conducta y gestión empresarial, el establecimiento del perfil ético propio de cada organización, dentro de un programa de ética empresarial que contribuya para orientar la actuación de sus integrantes y crear su identidad, en función del cumplimiento, la gestión de riesgos, el valor agregado y, por tanto, la mejora en su reputación ante la sociedad [21].

Bajo este enfoque, se asume a la ética empresarial como una cultura, en la cual las actitudes de los integrantes de una empresa u organización son éticamente aceptables y se apegan a los lineamientos y políticas establecidos, y donde el comportamiento de los directivos sirve de ejemplo y genera un entorno de "valores compartidos" para reducir los conflictos entre las convicciones morales individuales y el logro de los objetivos económicos de la organización, motivando a la vez a los empleados a comportarse éticamente [22]. Este esquema de valores compartidos implica que los empleados se identifiquen con la organización y, que quienes se encargan de la gestión asuman un comportamiento ético. Es por esto que las organizaciones actuales deben ser particularmente exigentes al seleccionar a su personal e implementar acciones más allá de los sistemas de vigilancia y control o las orientaciones vertidas en los Códigos de Conducta, ya que el mundo empresarial actual requiere de personas responsables y de alta moralidad, cuyo comportamiento se sitúe dentro de los límites éticos, con capacidad para hacer juicios de valor, tomar decisiones y trabajar en equipos multidisciplinarios con la finalidad de lograr, a un mismo tiempo, los objetivos corporativos, la satisfacción de necesidades de la sociedad y el desarrollo individual de los empleados [23].

Un programa de gestión y gobierno de los datos es la primera pieza sobre la que sustentar una adecuada explotación de la información, considerando los datos y posterior información inferida a partir de ellos como activos empresariales valiosos. Los datos y la información han de ser gestionados de manera cuidadosa, como cualquier

otro activo, asegurando la calidad, seguridad, integridad, disponibilidad, uso efectivo, el cumplimiento de regulaciones internas y externas, y por ende los aspectos éticos derivados del uso de los datos. Para alcanzar este objetivo las compañías deben implementar proyectos de gobierno de los datos, un conjunto de políticas y procedimientos que combinados establecen los procesos que supervisan el uso y gestión de los datos para transformarlos en un activo estratégico, con el objetivo de llevar a nuestra compañía a un nivel superior de "madurez en el uso de la información," mejorar la calidad de los datos y solucionar los posibles inconsistencias, gestionar el cambio en relación con el uso de los datos y cumplir con regulaciones y estándares internos y externos.

Según Idalberto Chiavenato: "La cultura organizacional es un conjunto de valores, creencias, tradiciones y modos de ejecutar las tareas que, de manera consciente o inconsciente, cada organización adopta y acumula con el tiempo, y que condiciona fuertemente el pensamiento y el comportamiento de sus miembros" [24]. Se requiere entonces, la generación de condiciones que eviten la comisión de actos deshonestos, fundamentada en una propuesta no sólo de mecanismos de control interno, sino que ahonde en aspectos formativos en Ética, que incentiven la integridad de los miembros de las organizaciones, mediante actividades de reflexión sobre las consecuencias derivadas de conductas inapropiadas. La corrupción se extiende con facilidad cuando se dan ocasiones propicias con beneficios altos y bajo riesgo. "La oportunidad viene dada habitualmente por la concentración de poder de decisión de una persona u organismo, sin el control suficiente, el beneficio deriva de los ingresos que pueda obtener el corrupto y los riesgos derivan de la detección de la corrupción y su castigo" [25].

Si se incorpora en los principios de la empresa el concepto de bien común, se puede hablar de ética social. Es imprescindible que los colaboradores se consideren intérpretes de una aspiración común y participantes que tienen un papel valioso en el logro de la misma. El bien común es un conjunto de condiciones de la vida social que posibilitan a los hombres el logro de los objetivos individuales y organizacionales en total armonía El empresario debe favorecer el progreso técnico, el espíritu de innovación, la iniciativa, la motivación de cuantos integran la organización.

Si bien se persiguen objetivos económicos, no se puede tener una visión reduccionista de la empresa, centrada en el logro de resultados cuantificables en términos económico-financieros, sino que se debe poner la actividad al servicio del hombre. "Se deben tener en cuenta necesidades materiales y las exigencias intelectuales, morales, espirituales de todo hombre, de todo grupo de hombres; así el empresario, siguiendo sus métodos y leyes propias, servirá de forma eficiente contribuyendo al desarrollo del bien común" [26].

Según Doménec Melé "La ética de la virtud no solo marca lo que no se debe hacer, sino que apunta a la excelencia. En otras palabras, no basta con resistir el entorno, sino que es necesario luchar por cambiar la situación. Empresarios, directivos y empleados pueden plantearse diversas posibilidades de acción como medidas preventivas" [27].

En el esquema propuesto se enuncian las siguientes iniciativas: Implementación de códigos de conducta empresariales, mediante declaraciones de cumplir con leyes, valores y normas de la organización; establecimiento de criterios generales, con políticas concretas para que se respeten los valores a la hora de contratar un proveedor; relacionarse con clientes; establecer contactos con funcionarios públicos, entre otros.

Es importante que la empresa forme a sus empleados dándole los criterios morales y jurídicos para entender en qué consiste el problema de la corrupción, cuáles son las razones para rechazarla y cuáles son las consecuencias que se derivan de este flagelo y su influencia en el entorno social externo.
A través de la cooperación social se perfeccionan y actualizan las normas éticas, por una dependencia recíproca entre los hombres que facilita crear lazos y formular objetivos comunes. Sin el concepto de un verdadero bien común en el cual los colaboradores se consideren intérpretes de una aspiración común y participantes interesados en la misma, no podemos hablar de una ética social. La política orientada hacia los valores que se rige por metas morales y que se somete a controles razonables, despliega la capacidad de crear el equilibrio necesario entre los intereses recíprocos de los diferentes actores, partícipes de la actividad. Bien sea por la presión interna en las organizaciones, por la demanda social o por las influencias políticas que establecen leyes o disposiciones reglamentarias, las empresas han asumido que, sin ética, no hay negocio. Para Argandoña, es más un proceso que un resultado. El autor la concibe como un modo de operar que se naturalice al interior de las empresas, maximizando el beneficio del conjunto [25]. Surge una mayor conciencia social, donde los consumidores demandan cada vez más comportamiento socialmente responsable y se percibe una sensibilidad mayor hacia el impacto que puede generar la actividad de la empresa en el medio y en la calidad de vida de la comunidad en la que opera [2].

6 ESTRATEGIAS TECNICO JURIDICAS

Se deberán tener en cuenta los lineamientos generales cuando los conocimientos derivados de los datos puedan afectar la dignidad humana, los usos posteriores de los conjuntos de datos y de las herramientas analíticas que los acompañan considerando su historial para la toma de decisiones, propiciando un uso coherente que garantice la privacidad y seguridad de los datos personales, formulando una estrategia que considere las demandas cambiantes del medio.
Como se muestra en la figura 3, los Profesionales de datos pueden tratar de mitigar los impactos de la exclusión llevando a cabo buenas prácticas, tratando de maximizar la transparencia mediante la elaboración de prácticas éticas que prevean el respeto de la dignidad humana y con el uso coherente de los datos para evitar la discriminación en todas sus formas.
Asimismo, se deberían desarrollar prácticas para identificar al sujeto responsable y prácticas de control interno, promoviendo la realización de auditorías en forma rutinaria y transparente para la gobernanza ética de los datos; la reducción de diferencias que surgen luego de los procesos de recopilación, correlación y predicción de datos mediante el respeto de la Dignidad de la Persona; responsabilizarse a sí mismo y a sus pares en el cumplimiento de los estándares.
Si bien son importantes los planes, los principios éticos son el pilar en el que se debe apoyar la organización para fijar políticas y también los objetivos a largo plazo. Incorporar la transparencia en el diseño de prácticas de recolección de datos, teniendo cuidado de recolectar sólo los lícitos, pertinentes, y no excesivos; todo ello acompañado de prácticas de rendición de cuentas y prácticas de auditoria.

Las prácticas colaborativas deben ser planeadas para incrementar la transparencia y deben ser conocidas por todos y revisadas regularmente. Entonces, los directivos deben prestar atención a las condiciones de funcionamiento de la organización y del mercado a la luz de normas y valores que se plasman en Códigos de Conducta. El objetivo es prever todo hecho que pueda relacionarse con comportamiento no ético, identificando partícipes y analizando la aplicación de sanciones. Minimizar los riesgos y tener en cuenta la resiliencia a la hora de evaluarlos.

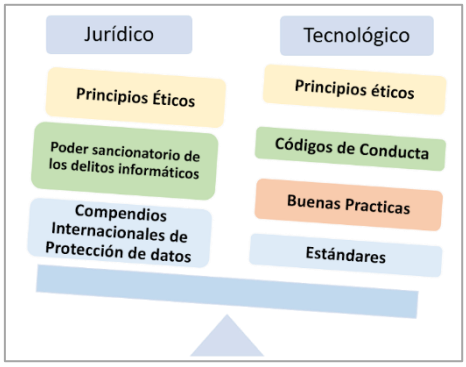


Fig. 3. Estrategia Técnico Jurídica para el manejo de datos en BI. Fuente: de Elaboración Propia

El mercado facilita la transmisión e intercambio de información y el cálculo de los resultados en términos de precios y cantidades. Si los individuos se prestan a prácticas que no respetan las normas establecidas, respecto a sus intenciones o en cuanto a los productos que ofrecen, surgen imperfecciones en los procesos que perjudican a las personas de las organizaciones intervinientes. Sus efectos influyen en el normal funcionamiento del mercado, provocando desequilibrios, afectando la rentabilidad, pero también genera acciones nocivas, las cuales se traducen en la violación de derechos como el de propiedad, la libertad de asociación, los derechos laborales, y el cumplimiento de las leyes relativas a las transacciones económicas, perjudicando a las organizaciones.

En base a lo anterior, se puede afirmar que los comportamientos no éticos acarrean altos costos para las organizaciones ya que, si no se detectan, se puede producir un aprendizaje en los partícipes. Se requiere fomentar una cultura basada en valores y monitorear las políticas y programas de responsabilidad social y calidad para evitar el surgimiento de cambios indeseados que dificulten el logro de objetivos. El gran reto que enfrentan las organizaciones es conciliar las demandas de la sociedad con las expectativas internas, debiendo analizar la forma de conducir a las personas y a los procesos, para anticiparse a fenómenos de conducta no ética en el seno de las mismas.

Las organizaciones de la segunda década del siglo XXI deben asumir los riesgos que impone el mundo global, generando condiciones para el desarrollo del BI con las herramientas que permitan su desarrollo y sostenibilidad. Las personas son los artífices de la actividad organizacional, no solo por su inteligencia y habilidades, sino por sus valores y la contribución de internet y de las tecnologías aplicadas a los negocios, como BI, en un marco ético.

7 CONCLUSIONES

El comportamiento organizacional deberá encauzarse entonces hacia el desempeño responsable donde la conducta se regule por normas éticas. El esfuerzo motivacional será eficaz si se analiza la organización para determinar sus pautas culturales y conocer las creencias en las que está enraizado el comportamiento de los miembros y comprender los ejes que posibilitan el logro de objetivos.

Además, se puede establecer el camino a seguir si se enfrentan cambios hacia nuevos enfoques de negocio, basados en la búsqueda de una visión compartida, objetivos de mejora, responsabilidad social, desarrollo sostenible. Para todos los partícipes, la presencia de un modelo de negocio que aspire a la realización de metas involucrando a las personas con valores compartidos, propicia la competitividad, creando el ambiente adecuado para el modelo de BI, con la inserción de nuevas formas de realizar negocios, con mejoras en forma continua y metodologías que aseguren lograr los objetivos tendientes al desarrollo sostenible.

La gestión ética de la información debe realizarse para los datos correspondientes a todos los niveles y unidades, gestionando con personas calificadas los procesos organizacionales. Se debe analizar la necesidad de crear áreas o centros de sistemas, considerando que se conforman grupos humanos alrededor de nuevas actividades del área de sistemas como la Inteligencia de Negocios o más conocida como BI.

Señalar la visión organizacional y gestionar personas y elementos materiales para alcanzarla se realiza con efectividad, cuando la organización decide adoptar formas de conducción relacionados con la ética, lo cual se puede canalizar a través de la responsabilidad social y sus conceptos, adoptando principios en un sistema de valores, que se constituye en el credo de la organización.

Los modelos de BI requieren el cuidado del comportamiento humano alrededor de tecnologías, creando un clima de armonía, enmarcado en valores compartidos, trabajando en equipo para así lograr satisfacer los requerimientos actuales y futuros de la sociedad.

Referencias

[1] Ahumada Tello, E y Perusquia Velazco J.M. (2015). Business intelligence: Strategy for competitiveness development in technology-based firms. Disponible en www.sciencedirect.com, marzo de 2020.

[2] Debeljuch, Patricia (2009). Ética Empresarial, en el núcleo de la Estrategia Corporativa, Cengage Learning, Buenos Aires.

[3] Acosta Velázquez S., Pedraza, A. (s/d). Replanteamiento de la Ética para su implementación en los sistemas de gestión empresarial. Disponible en https://www.uaeh.edu.mx/scige/boletin/tepeji/n8/a17.html

[4] López Benítez, Y. (2018). Business Intelligence. ADGG102PO. IC Editorial. 1° Edición. Andalucía – España. Vol. 1, 151 pág. ISBN: 978-84-9198-467-2.

[5] Cegielski, C., Jones-Farmer, L. (2016). Knowledge, skills, and abilities for entry-level business analytics positions: A multi-method study. Decision Sciences Journal of Innovative Education, Vol. 4(1), enero de 2016, Pp.91–118. Disponible en https://doi.org/10.1111/dsji.12086.

[6] Rodríguez, Miguel Ángel (2008). Ética de los negocios en un mundo global. Disponible en http://www.scielo.org.mx/pdf/fn/v20n40/v20n40a7.pdf

[7] ISO 26001 (2017). ¿Sistema de Gestión de la Responsabilidad Social? Disponible en https://responsabilidadsocialuniversitaria.wordpress.com/2017/04/13/iso-26001-sistema-de-gestion-de-la-responsabilidad-social/
[8] Fernández Aguado, Javier (2001). Ética en los Negocios, Ariel Sociedad Económica, Madrid, 2001, p.42
[9] Salvador García, y Shimon Dolan (2006). La Dirección por Valores, Mc. Graw-Hill, Barcelona, 2006.13-09.
[10] Colmarejo Fernández, Rosa (2018). Ética aplicada a la Gestión de datos masivos. Revista de la Universidad de Granada. Disponible en https://revistaseug.ugr.es/index.php/acfs/article/view/6553.
[11] Pérez Porto, Gardey, J. A. (s/d). Definición de código de ética. Disponible en https://definicion.de/codigo-de-etica.
[12] Código ACM. (s/d). Código de Ética y Conducta Profesional de ACM. Association for Computing Machinery. Disponible en https://www.acm.org/about-acm/code-of-ethics-in-spanish.
[13] Neif, Jane, S. E. (2006). Ética informática en la sociedad de la información. Disponible en https://produccioncientificaluz.org/index.php/rvg/article/view/10441
[14] Pérez Porto, Gardey, J. A. (s/d). Definición de código de ética. Disponible en https://definicion.de/codigo-de-etica/
[15] Normas ISO. Disponible en https://www.normas-iso.com/iso-27001/
[16] Naciones Unidas. Biblioteca Digital. Disponible en https://digitallibrary.un.org/record/79050?ln=es
[17] Arrubias Urrutia, A. (2000). La Importancia de la Comunicación Interna en la Empresa. En Revista Latina de Comunicación Social. Número 27. marzo de 2000.p.6
[18] Kielstra, Paul Denis.Mc Gauley, ed. (2007). Economía Intelligence Unit.
[19] Becerra, María del Carmen (2016). Integración de estándares para la gestión de identidad y privacidad. WSI - II Workshop de seguridad informática CACIC 2016.
[20] Naciones Unidas. Guidelines for the Regulation of Computerized Personal Data. Disponible enhttps://www.accenture.com/t20160629T012639Z__w__/us-en/_acnmedia/PDF-24/Accenture-Universal-Principles-Data-Ethics.pdf
[21] Acosta Velázquez S., Pedraza, E (2017). Replanteamiento de la ética en la gestión empresarial. Disponible en https://repository.uaeh.edu.mx/revistas/index.php/tepexi/article/view/2388
[22] Empresal(2006).Ética Empresarial. Disponible en www.fundemas.org
[23] Martínez, H. (2005). La ética empresarial al comienzo de un nuevo milenio. Una aproximación bibliográfica y pedagógica. Facultad de Ciencias Económicas y Administrativas, Cuadernos de Administración no. 2, abril 2005, Editorial Universidad Javeriana
[24] Chiavenato, Idalberto (2014). Introducción a la Teoría General de la Administración, octava edición, Ed. Mc. Graw Hill-Educación, México, 2014.
[25] Argandoña, A. (1999). "Las empresas ante la corrupción." Documento de investigación" N°392, IESE, Barcelona, 1999, p.3.
[26] Bernardo, J. Ética de los Negocios. Editorial Machi Grupo Editor. 1993.. ISBN: 9505372396
[27] Doménec Melé, "Sobornos y extorsiones en los negocios", Documento de Investigación 239, IESE, Barcelona, 1993, p 10-12.

Privacy Preserving Multi Party Computation for Data-Analytics in the IoT-Fog-Cloud Ecosystem

Julio Lopez-Fenner[1], Samuel Sepúlveda[1], Luiz Fernando Bittencourt[2], Fabio Moreira Costa[3], and Nikolaos Georgantas[4]

[1] Departamento de Ciencias de la Computación e Informática,
Centro de Estudios en Ingeniería de Software
Universidad de La Frontera, Temuco, Chile
{julio.lopez, samuel.sepulveda}@ufrontera.cl
[2] Universidade Estadual de Campinas (UNICAMP), Brasil
bit@ic.unicamp.br
[3] Universidade Federal de Goiás, Brasil
[4] Inria Paris, France

Abstract. In this paper, we propose an architecture for privacy preserving protocols in an IoT-Fog-Cloud ecosystem computing hierarchy. We consider the paradigms of Fog and Edge computing, together with a multi-party computation mechanism that enables secure privacy-preserving data processing in terms of exchanged messages and distributed computing. We discuss the potential use of such an architecture in a scenario of pandemics where social distancing monitoring and privacy are pivotal to manage public health yet providing confidence to citizens.

Keywords: Secure Multi Party Computation · Internet of Things · Edge/Fog computing · Security and Trust

1 Introduction

The Internet of Things (IoT) is bringing countless ways of collecting and processing data as more and more smart devices and sensors (IoT devices, for short) are deployed and connected to the Internet.

IoT devices are scattered at the edge of the network. They produce vast amounts of heterogeneous data that does not generate value by itself by uploading it to a - say - database [27] but needs computing power to be either further processed in order to generate a desired output or to produce an immediate response for a data query, may be even to produce intermediate data results that need to be further acquired or with higher levels of processing.

Examples of these requirements can be easily found in smart agriculture domains [9], healthcare, industry and energy, transportation, and urban infrastructures, to name a few [22].

In this context, privacy concerns upon data being transmitted to and from the cloud arise naturally, as IoT devices and their data constitutes major targets

for security and privacy attacks, including data breaching, data integrity, and data collusion [33]. Moreover, very often the data collected - if leaked and not protected by suitable encryption - can breach the user's privacy and be used for non authorized virtual profiling [14].

On the other hand, however, data collected from IoT devices can also be used for the common good and social profitability, as knowledge gathered from those data sets can help in improving people's quality of life, for example, by offering personalised services, precise and up-to-date information about people's surroundings and interests, and information for improving planning and efficiency of cities as well as for policy making [36]. Therefore, there is ample need for the design and implementation of mechanisms for both, enabling and preserving privacy that still allows significant IoT data to be collected, distributed, and processed in a useful manner.

Currently, cloud computing has been used to store and process data collected by IoT devices. This well-established computing paradigm is very convenient for data processing applications that require large processing capacity, and it fits well to computing problems that use data collected from many sources regardless of their location. However, as IoT devices are located and scattered at the very edge of the network, they can easily be brought to consuming data from neighbouring counterparts, which then would not require further transfer to a centralised cloud for processing [38]. In this scenario, Fog Computing [13] has emerged to bring and distribute computing capacities closer to the edge. In this computing paradigm, micro data centres, or cloudlets, are deployed throughout the network (e.g. at access points or cellular base stations) to bring computing capacities closer to the IoT devices. Combined, all three layers of edge IoT devices, fog devices, and cloud computing, can provide a hierarchical ubiquitous-like infrastructure for processing IoT data. Notwithstanding, privacy concerns are still present in this scenario and should be addressed. Computational models for SMPC that fit the proposed hierarchical structure in terms of resource allocation need to ensure the privacy of individual users (parties) [17].

In order to present our approach, we introduce next some concepts and discuss challenges in the study of the three components of the IoT-Fog-Cloud ecosystem with regard to privacy preserving algorithms and protocols considering secure multi-party interactions. Then, in section 4 we introduce our proposed architecture and discuss its instantiation. Section 5 presents preliminary results towards such an architecture implementation and in section 3 we discuss some available literature on these topics. Section 6 brings conclusions and remarks.

2 Background

2.1 IoT-Fog-Cloud ecosystem

The Internet of Things (IoT) consists of 'zillions' of (inter-) connected devices that can include sensors/actuators with communication capabilities, i.e., virtually any object with embedded micro-controller and antennae, which can be

highly heterogeneous at different levels, as for example energy consumption requirements, communications protocols, computing capacity, and mobility.

Therefore, IoT devices management becomes intrinsically challenging in terms of data communication and processing, while it requires significant processing and knowledge extraction capabilities in order to provide relevant insight and, hence, constitutes an ideal field for developing and testing of Big Data tools. The challenge is precisely to transform gathered data into actual information knowledge by (secure) privacy-preserving processing. Examples of data collected by IoT devices with such requirements are: location-based services, security cameras, personalised preference tracking mechanisms (e.g. for advertisement), medical data collected by body sensors or by professional equipment at hospitals and clinics, see for example [29].

Cloud computing represents nowadays a largely adopted computing paradigm servicing a variety of applications, providing dynamic configuration possibilities such as elasticity and pay-per-use. Resource virtualization allows providers to share slices of computing resources among users through the use of virtual machines and containers, where they appear to be logically isolated for each tenant. On-demand provisioning/de-provisioning, elasticity, ubiquitous access, lower upfront investments with reduced capital expenditures and faster time to market, are a few properties offered by cloud computing.

Clouds can partially fulfil applications requirements for IoT but may fall short for low latency, for example. Fog computing can be combined with the cloud to provide an ecosystem that furnishes full potential for IoT application deployments, as it introduces a hierarchy of computing capacity (fog nodes, cloudlets or micro data centres) between the edge and the cloud [13]. This hierarchy can provide a wider range of services that may not be supported solely by the cloud. Thus, a fog computing infrastructure can attend applications with a variety of Quality of Services (QoS) by using different hierarchical levels to meet latency and computing requirements.

A consequence of bringing processing closer to the edge with fog computing, as illustrated in Figure 1, is to reduce total bandwidth used in the network along the path between edge and cloud, as data can be aggregated at the edge before being transferred to the cloud. This computing hierarchy provided by the IoT-Fog-Cloud ecosystem enables thus processing to occur closer to where the data is being generated. This is useful for applications that depend upon very low delays or response times, but also for applications that only need data from constrained or specific geographic locations, as in this scenario the data does not need to be transferred all the way up to the cloud. Such a scenario demands the implementation of privacy-preserving mechanisms so that data shared from said IoT devices are not unintentionally available to third-parties.

Examples of such applications include real-time traffic estimation, autonomous cars traffic control, collision avoidance in non-signalised road intersections, localisation-based risk assessment, monitoring in disaster scenarios, actuators in dynamically changing environments, and so on.

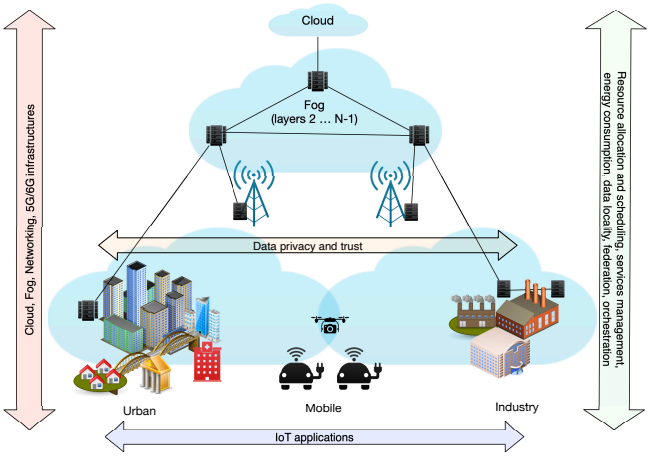


Fig. 1: Overview of the IoT-Fog-Cloud infrastructure (adapted from [4]).

We are interested in the modelling and design of efficient processing data protocols for applications in a privacy-preserving context. We posit that Secure Multi-Party Computation (SMPC) [16,19], understood as the joint computation of a desired function in terms of an appropriate splitting and distribution among a number of parties under the constraint that all data has to remain private, naturally furnishes a setting for privacy-preserving data processing and correctness:

- Privacy, in the sense that no party learns anything new other than its prescribed jointly computed output; and
- correctness, in the sense that parties can trust that the computed quantities are indeed correct.

This is useful, e.g., for the distributed evaluation of trust, as in [7], where parties or players compute a joint confidence level by combining their mutual degrees of trust. See the survey in [37] (and the references therein) for a nice description of the role of SMPC in the IoT field.

In view of potential applications, technologies should also consider, at least, designs providing Confidentiality, Integrity and Availability (CIA triad) [21]. Other requirements may be further imposed, for example on the side of independence of inputs, fairness, etc. (see [16]).

2.2 Privacy preservation in the IoT

Secure multiparty computations (SMPC) allows n players to compute together the output of some function, using private inputs without revealing them to other parties and without gaining access to knowledge entrusted to another party.

Notice that while SMPC protocols have been present in the literature for more than thirty years, their application to a data streams in a hierarchy of IoT

devices (in terms of computational power), which are continuously producing data worth analysing while preserving privacy, can be considered as a relatively recent development [15], in contrast to the more typical reception, storage and analysis of source data by a third-party centralised server that is not under the administrative control of the smart environment and hence potentially untrustworthy.

In the proposed architecture, the server should not obtain access to the raw data of the sources; it only fulfils management and orchestration purposes to carry out SMPC computations that are executed by the sources themselves [19]. This enables privacy-preserving data processing while the sources can be dynamic. The role of an edge-IoT gateway, understood as middle-ware between IoT devices and the cloud that facilitates computations and is a communication gateway as well [24], may be further extended to be able to receive queries regarding the outcomes of the computations still without revealing any private information about the sources.

Performance and resource consumption characteristics of SMPC methods are key limitations for their application in resource constrained environments. A detailed assessment of these characteristics is carried out in [20]. The authors reach the conclusion that SMPC methods can be applied, in practice, in Intranet environments, but with limitations in Internet settings. This shows the need for new SMPC-based solutions where the resource environment performs in close collaboration with privacy-preserving mechanisms.

3 Related work

A set of requirements for privacy preservation mechanisms in the context of distributed and ubiquitous environments, such as in the IoT, was presented in [1]. Among those requirements, they highlight the importance of ensuring privacy of the outputs of SMPC, which may be as important as the privacy of the inputs (as normally targeted in SMPC approaches), since data profiling might be used to de-anonymize the inputs or obtain sensitive information about the parties.

We address this issue to some extent, as the outputs at a lower level of the hierarchy are handled as inputs to the immediate upper level, and thus are subject to the usual privacy requirement. It remains to assess the effects of such privacy handling in relation to the use of the intermediate outputs for answering localised queries in a multi-layered fog architecture.

Some architectures for IoT assume that infrastructural elements in fog and edges are trusted entities [26]. Thus, to preserve data privacy, those architectures adopt data anonymization techniques such as differential privacy [8] based on a trust curator model [28]. In an untrusted component model, which is the center of this proposal, infrastructural components are not assumed as trusted entities so they should not access sensible data. This model usually adopts techniques of secure multi-party computation to preserve data privacy.

In IoT, the adoption of those techniques differs in terms of the mechanism for privacy preservation, resource usage adequacy and the architectural-style

for IoT platforms. Privacy preservation can be accomplished by differential privacy [8], homomorphic encryption (partial or full) [10] and secure multi-party computation. Adoption of those techniques have a direct impact on resource usage adequacy for IoT, but also in terms of requirements of processing power and exchanged messages among devices. In particular, they produce a huge increase in the length and number of messages, so they should be adapted (e.g. [2]) to be applied in large-scale scenarios of IoT. Section 2.2 above discusses relevant approaches for privacy preservation in the IoT.

The architectural-style of an IoT platform may be exploited to satisfy resource limitations (e.g., processing power) or to decrease the number of disseminated messages, by transferring processing from devices or servers to intermediary components. Approaches of privacy preservation for IoT assume architectures such as cloud-based [2] platforms, mixed-mesh [34] networks, edge networks [11], fog [30], and mixed architectures with trusted entities [1]. In particular, fog and edge-fog architectures allow multiple levels of processing and they are more adaptable to the dynamicity and diversity of IoT devices. We hypothesize that fog-edge architectures allow the development of efficient privacy preservation techniques, in terms of exchanged messages and distributed processing power.

4 Privacy towards the Edge

This section describes an overview for the proposed privacy-preserving multi-party computation architecture for IoT using fog computing.

4.1 Hypotheses and Objectives

In order to develop a framework of theoretical nature and a test pilot for the hierarchical architecture in layers: from IoT to Edge and from Edge to Fog computing, in which multi party protocols can be proposed, we first consider current techniques and methods for Fog/Edge technologies and Multi party computations.

Since cloud computing cannot rely exclusively upon cryptographic protocols alone [35] and, furthermore, the data collected by cyber-physical (IoT) sensors is known to be securely (and privately) processed using zero-knowledge proofs, as in [12], we propose to undertake our approach of Privacy Preserving Multi Party Computation for Data Analytic in the Iot-Edge-Fog ecosystem by first establishing a hierarchy of devices, as in Figure 2 and then to follow the lines proposed by [32] and [23].

4.2 Architectural Principles

At the edge, we propose that IoT devices are grouped, either statically or dynamically, to be considered as data generators for a specific application or user query. Grouping is based on criteria such as resource allocation (at the fog),

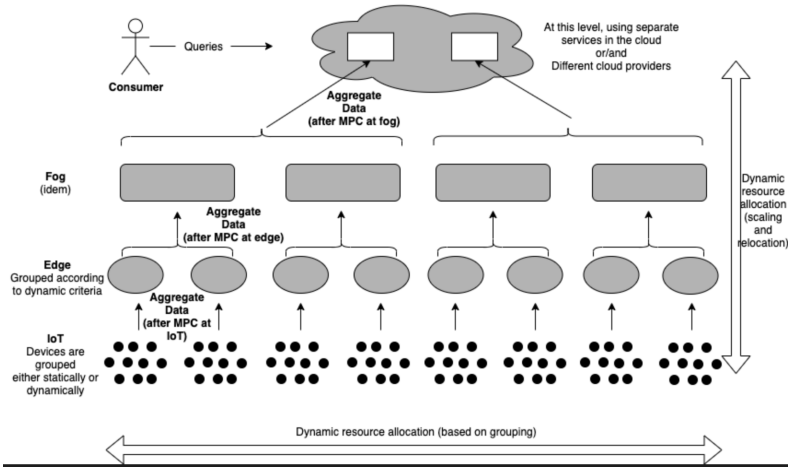


Fig. 2: Edge/Fog Computing hierarchy together with secure multi-party computation brings resource allocation challenges.

scalability and geographic location. These devices are assumed to be capable of executing SMPC in the context of the group which they currently belong to. The output of such computation may either be used immediately, to answer localised queries posed by users/consumers, or sent up (to services running in the same or a different micro-datacenter in the fog) as intermediate outputs that in turn are used in conjunction with outputs from other SMPC computations performed by other IoT device groups. This pattern of dynamically grouping SMPC parties that receive inputs from SMPC groups located underneath in the hierarchy repeats at the fog and cloud levels to produce intermediate outputs and/or to answer user queries.

To illustrate how this works, we consider the current pandemic scenario: Consider an end-user who is willing to know how safe it is to go to a specific place. Assume we can come up with a safety metric based on several pieces of information from the neighbourhood and/or place where this user is willing to go. A combination of data from IoT devices at the location of interest can be used to compute such safety metric and respond to the user query. However, it is clear that such information can often include personal data that raises privacy concerns, as for example medical conditions of people in that area, real-time and accurate information about people's location, data on how those people interact or have interacted on that location, information collected from cameras about the amount of people wearing masks, and so on. A privacy-preserving mechanism using multi-party computation would need to first choose sources to obtain all that data to be processed, and this is an example of how devices are to be grouped into sets that are able to generate data in response to a query. Then, the data set obtained from those devices can be first processed using multi-party computation protocols, and then be further moved to the closest cloudlet/micro

data center at the fog level for further processing. With privacy already in place, it remains to aggregate the data and compute the necessary safety metrics of interest to that query.

Note that more complex scenarios, with more data or data from multiple geographic locations, can involve multiple cloudlets at the fog level or even involve the cloud to provide enough capacity or enough data sets to allow a proper computation of the desired metric.

To enable SMPC in the discussed scenario, many challenges arise. First, a candidate set of SMPC protocols must be identified, studied, and potentially adapted for the proposed infrastructure. Second, the composition of IoT device groups should be enabled by distributed computing mechanisms. Third, once these groups have been defined for a specific application or query, data processing should occur in an efficient manner throughout the system through proper resource allocation and scheduling mechanisms.

4.3 Architecture Instantiation

It is known that effective mechanisms addressing stacked fog layers (cloudlets) constitutes still a challenge and open issue [4], since a universal standard of multi-layered fog computing architecture has not yet been adopted. We address this issue by considering data privacy aspects as inherent to a multi-layered fog architecture, in which resource allocation mechanisms are needed to efficiently process privacy-sensitive information.

Indeed, Fog computing technology enables the deployment of cloud-like computing services at the edge of the network. Fog-aware mechanisms and procedures to implement efficient resource management for applications with different levels of latency requirements remains a non trivial challenge [18]. Several efforts have addressed this body of work to provide a latency-driven fog computing environment for IoT applications using only one fog layer (i.e. *cloudlets*, *micro data centres*, or *fog nodes*) between the edge and the cloud (as illustrated in the proposed fog-cloud topology in Figure 1) [31,3,25].

Next, it is common knowledge that some architectures [26] for IoT assume that infrastructural elements in fog and edges are trusted entities. Thus, to preserve data privacy, those architectures may adopt data randomisation techniques such as differential privacy [8] based on a trust curator model [28]. But, in a complete untrustworthy component model, which is at the centre of our considerations, infrastructural components are not necessarily assumed to be trusted entities, so they should not access sensible data. We propose hence that our model should adopt techniques of secure multi-party computation to preserve data privacy.

Adoption of these techniques will have a direct impact on resource usage adequacy for IoT, in terms of processing power and exchanged messages among devices. In particular, they produce a huge increase in both the length and the number of messages, so they should be adapted to large-scale scenarios with IoT.

Finally, for the deployment of these protocols, end-to-end automated systems for deploying large-scale SMPC protocols between end users, called MPSaaS (acronym for system-as-a-service) have been designed and tested [2].

5 Preliminary Results and Discussion

In the field of smart homes with IoT, secure communications between smart devices endowed with low computational capabilities and a home gateway are frequently established via unsecured wireless communication channels. In [5] we explored secure data transfer generated by IoT using standard Diffie-Hellman secret sharing which can be applied via buffering to higher throughput communications. Our assumptions were that at setup both the user and the gateway have enough processing power to perform - say - one time secured RSA encrypted communication, hence relaxing the need for a trusted secure server outside the domain, and that the protocol should at least be secure for a range of known attacks, as replay or DoS attacks. In turn, the approach failed impersonation attacks via Man in the Middle (MiM) so that a possible way out may lie in the implementation of SMPC as proposed in the architecture proposed here, since authentication would require a consensus (for example defined using Shamir's secret schem) to be build upon trusted devices. Hence, as mentioned above, joint computation of trust among parties can be performed by matrix multiplication, as in [6], where we proposed an *ad hoc* Strassen-Winograd multiplication protocol that use a combination of partial homomorphic encryption schemes and additive masking techniques together with a (novel) schedule for the location and encryption layout of all intermediate computations, such that privacy is preserved. It turned out that the asymptotic communication volume and computational time was reduced from $O(n^3)$ to $O(n^{2.81})$.

Thus, the main challenge now is to incorporate the developed protocols to the proposed computing hierarchy. This involves the prior study of performance of devices from the edge to the cloud to evaluate the possible ways of allocating resources for protocol processing and also assess communication costs. As a second step, evaluating the performance of the protocols in a real-world scenario, considering requirements as the amount of data inputs and needed response time (e.g. in the pandemics scenario described in Section 4), is key for establishing limitations and pinpoint where research must be focused to enable performance-effective multiparty computation at the edge.

This involves developing optimization algorithms for dynamic grouping or clustering, as presented in Figure 2, which should consider multiparty computation aspects both in terms of privacy and response times: a too large set of IoT devices may impair performance for response time, while a too small set may result in poor output for the application requiring data collection and processing. To evaluate these aspects in the platform, we plan to assess metrics such as response time when compared to a centralized cloud deployment and accuracy of deployed algorithms for grouping optimization. Also, networking metrics will be collected to evaluate how/if the framework can reduce data traffic while

maintaining processing capacity and accuracy at the same level as a centralized system. Lastly, power consumption of devices using multi-party computation can be evaluated to learn if the framework can be used in practice without impairing users' experience with mobile devices.

6 Conclusions and future work

This work presented the main concepts and initial findings about the design and evaluation of privacy preserving protocols in an IoT-Fog-Cloud ecosystem computing hierarchy. We briefly discussed how this can be applied in a scenario as the present times of pandemics, but also covered general aspects that need attention for a generalized application of privacy-preserving mechanisms at the edge of the network.

We considered the paradigms of Fog and Edge computing, together with a multi-party computation mechanism to enable secure privacy-preserving data processing at the edge. The proposed architecture demands further developments in terms of understanding performance limitations and also in terms of resource allocation for data collection and processing of multiparty computation protocols, which can be the focus for further research. Also, the formalization of the proposal, with the modeling of inputs and outputs for multiparty computation at the edge, is a key future work to enable proper evaluation of efficiency and limitations of the proposal.

Acknowledgments

This work was partially funded by the São Paulo Research Foundation (FAPESP), grants #2018/23126-3 and #2015/24494-8, CAPES, and CNPq, Brazil, grants 432943/2018-8 and 309562/2019-8. Partial support from Universidad de La Frontera, Vicerrectoría de Investigación grant DI20-0060 is also acknowledged.

References

1. Ankele, R., Küçük, K.A., Martin, A., Simpson, A., Paverd, A.: Applying the trustworthy remote entity to privacy-preserving multiparty computation: Requirements and criteria for large-scale applications. In: 2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld). pp. 414–422 (2016)
2. Barak, A., Hirt, M., Koskas, L., Lindell, Y.: An end-to-end system for large scale p2p mpc-as-a-service and low-bandwidth mpc for weak participants. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. pp. 695–712 (2018)
3. Bittencourt, L., Diaz-Montes, J., Buyya, R., Rana, O., Parashar, M.: Mobility-Aware Application Scheduling in Fog Computing. IEEE Cloud Computing **4**(2), 26–35 (March 2017)

4. Bittencourt, L., Immich, R., Sakellariou, R., Fonseca, N., Madeira, E., Curado, M., Villas, L., DaSilva, L., Lee, C., Rana, O.: The internet of things, fog and cloud continuum: Integration and challenges. Internet of Things **3-4**, 134 – 155 (2018), http://www.sciencedirect.com/science/article/pii/S2542660518300635
5. Díaz Arancibia, J., Ferrari Smith, V., López Fenner, J.: On-The-Fly Diffie-Hellman for IoT. In: 2019 38th International Conference of the Chilean Computer Science Society (SCCC). pp. 1–5. IEEE (2019)
6. Dumas, J.G., Lafourcade, P., Lopez Fenner, J., Lucas, D., Orfila, J.B., Pernet, C., Puys, M.: Secure multiparty matrix multiplication based on strassen-winograd algorithm. In: International Workshop on Security. pp. 67–88. Springer (2019)
7. Dumas, J.G., Lafourcade, P., Orfila, J.B., Puys, M.: Dual protocols for private multi-party matrix multiplication and trust computations. Computers & security **71**, 51–70 (2017)
8. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) Theory of Cryptography. pp. 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
9. Elijah, O., Rahman, T.A., Orikumhi, I., Leow, C.Y., Hindia, M.N.: An overview of internet of things (iot) and data analytics in agriculture: Benefits and challenges. IEEE Internet of Things Journal **5**(5), 3758–3773 (2018)
10. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing. p. 169–178. STOC '09, Association for Computing Machinery, New York, NY, USA (2009)
11. Ghinita, G., Kalnis, P., Khoshgozaran, A., Shahabi, C., Tan, K.L.: Private queries in location based services: Anonymizers are not necessary. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. p. 121–132. SIGMOD '08, Association for Computing Machinery, New York, NY, USA (2008)
12. Goldwasser, S., Micali, S., Rackoff, C.: The knowledge complexity of interactive proof systems. SIAM Journal on computing **18**(1), 186–208 (1989)
13. Group, O.C.A.W., et al.: Openfog reference architecture for fog computing. OPFRA001 **20817**, 162 (2017)
14. Hossain, E., Khan, I., Un-Noor, F., Sikander, S.S., Sunny, M.S.H.: Application of big data and machine learning in smart grid, and associated security concerns: A review. IEEE Access **7**, 13960–13988 (2019)
15. Kerschbaum, F.: Privacy-preserving computation. In: Annual Privacy Forum. pp. 41–54. Springer (2012)
16. Lindell, Y.: Secure multiparty computation for privacy preserving data mining. In: Encyclopedia of Data Warehousing and Mining, pp. 1005–1009. IGI Global (2005)
17. Liu, W., Luo, S.s., Wang, Y.b., Jiang, Z.t.: A protocol of secure multi-party multi-data ranking and its application in privacy preserving sequential pattern mining. In: 2011 Fourth International Joint Conference on Computational Sciences and Optimization. pp. 272–275. IEEE (2011)
18. Mahmud, R., Ramamohanarao, K., Buyya, R.: Latency-Aware Application Module Management for Fog Computing Environments. ACM Trans. Internet Technol. **19**(1), 9:1–9:21 (Nov 2018)
19. von Maltitz, M., Bitzer, D., Carle, G.: Data Querying and Access Control for Secure Multiparty Computation . In: 2019 IFIP/IEEE International Symposium on Integrated Network Management. Washington, DC, USA (2019)
20. von Maltitz, M., Carle, G.: A performance and resource consumption assessment of secret sharing based secure multiparty computation. In: Data Privacy Management Workshop '18. Springer International Publishing (2018)

21. von Maltitz, M.L.: Secure and Privacy-preserving Services Based on Secure Multiparty Computation. Network Architectures and Services NET (2019)
22. Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M.: Deep learning for iot big data and streaming analytics: A survey. IEEE Communications Surveys & Tutorials **20**(4), 2923–2960 (2018)
23. Pallas, F., Raschke, P., Bermbach, D.: Fog computing as privacy enabler. IEEE Internet Computing (2020)
24. Papcun, P., Kajati, E., Cupkova, D., Mocnej, J., Miskuf, M., Zolotova, I.: Edge-enabled iot gateway criteria selection and evaluation. Concurrency and Computation: Practice and Experience p. e5219 (2020)
25. Peterson, L., Anderson, T., Katti, S., McKeown, N., Parulkar, G., Rexford, J., Satyanarayanan, M., Sunay, O., Vahdat, A.: Democratizing the Network Edge. SIGCOMM Comput. Commun. Rev. **49**(2), 31–36 (May 2019)
26. Pinto, S., Gomes, T., Pereira, J., Cabral, J., Tavares, A.: Iioteed: An enhanced, trusted execution environment for industrial iot edge devices. IEEE Internet Computing **21**(1), 40–47 (2017)
27. Raafat, H.M., Hossain, M.S., Essa, E., Elmougy, S., Tolba, A.S., Muhammad, G., Ghoneim, A.: Fog intelligence for real-time iot sensor data analytics. IEEE Access **5**, 24062–24069 (2017)
28. Rao, F.Y., Bertino, E.: Privacy techniques for edge computing systems. Proceedings of the IEEE **107**(8), 1632–1654 (2019)
29. Reddy, K.U.K., Shabbiha, S., Kumar, M.: Design of High Security Smart Health Care Monitoring System using IoT. International Journal of Emerging Trends in Engineering Research **8**(6), 2259–2265 (June 2020)
30. Saleem, A., Khan, A., Malik, S.U.R., Pervaiz, H., Malik, H., Alam, M., Jindal, A.: Fesda: Fog-enabled secure data aggregation in smart grid iot network. IEEE Internet of Things Journal pp. 1–1 (2019)
31. Shah-Mansouri, H., Wong, V.W.S.: Hierarchical Fog-Cloud Computing for IoT Systems: A Computation Offloading Game. IEEE Internet of Things Journal **5**(4), 3246–3257 (2018)
32. Sousa, P.R., Antunes, L., Martins, R.: The present and future of privacy-preserving computation in fog computing. In: Fog Computing in the Internet of Things, pp. 51–69. Springer (2018)
33. Tao, H., Bhuiyan, M.Z.A., Abdalla, A.N., Hassan, M.M., Zain, J.M., Hayajneh, T.: Secured data collection with hardware-based ciphers for iot-based healthcare. IEEE Internet of Things Journal **6**(1), 410–420 (2018)
34. Tonyali, S., Akkaya, K., Saputro, N., Uluagac, A.S., Nojoumian, M.: Privacy-preserving protocols for secure and reliable data aggregation in iot-enabled smart metering systems. Future Generation Computer Systems **78**, 547–557 (2018)
35. Van Dijk, M., Juels, A.: On the impossibility of cryptography alone for privacy-preserving cloud computing. HotSec **10**, 1–8 (2010)
36. Vorakulpipat, C., Rattanalerdnusorn, E., Thaenkaew, P., Hai, H.D.: Recent challenges, trends, and concerns related to iot security: An evolutionary study. In: 2018 20th International Conference on Advanced Communication Technology (ICACT). pp. 405–410. IEEE (2018)
37. Yan, Z., Zhang, P., Vasilakos, A.V.: A survey on trust management for internet of things. Journal of network and computer applications **42**, 120–134 (2014)
38. Zhang, Y., Ren, J., Liu, J., Xu, C., Guo, H., Liu, Y.: A survey on emerging computing paradigms for big data. Chinese Journal of Electronics **26**(1), 1–12 (2017)

Privacy Preserving Multi Party Computation for Data-Analytics in the IoT-Fog-Cloud Ecosystem

Julio Lopez-Fenner[1], Samuel Sepúlveda[1], Luiz Fernando Bittencourt[2], Fabio Moreira Costa[3], and Nikolaos Georgantas[4]

[1] Departamento de Ciencias de la Computación e Informática,
Centro de Estudios en Ingeniería de Software
Universidad de La Frontera, Temuco, Chile
{julio.lopez, samuel.sepulveda}@ufrontera.cl
[2] Universidade Estadual de Campinas (UNICAMP), Brasil
bit@ic.unicamp.br
[3] Universidade Federal de Goiás, Brasil
[4] Inria Paris, France

Abstract. In this paper, we propose an architecture for privacy preserving protocols in an IoT-Fog-Cloud ecosystem computing hierarchy. We consider the paradigms of Fog and Edge computing, together with a multi-party computation mechanism that enables secure privacy-preserving data processing in terms of exchanged messages and distributed computing. We discuss the potential use of such an architecture in a scenario of pandemics where social distancing monitoring and privacy are pivotal to manage public health yet providing confidence to citizens.

Keywords: Secure Multi Party Computation · Internet of Things · Edge/Fog computing · Security and Trust

1 Introduction

The Internet of Things (IoT) is bringing countless ways of collecting and processing data as more and more smart devices and sensors (IoT devices, for short) are deployed and connected to the Internet.

IoT devices are scattered at the edge of the network. They produce vast amounts of heterogeneous data that does not generate value by itself by uploading it to a - say - database [27] but needs computing power to be either further processed in order to generate a desired output or to produce an immediate response for a data query, may be even to produce intermediate data results that need to be further acquired or with higher levels of processing.

Examples of these requirements can be easily found in smart agriculture domains [9], healthcare, industry and energy, transportation, and urban infrastructures, to name a few [22].

In this context, privacy concerns upon data being transmitted to and from the cloud arise naturally, as IoT devices and their data constitutes major targets

for security and privacy attacks, including data breaching, data integrity, and data collusion [33]. Moreover, very often the data collected - if leaked and not protected by suitable encryption - can breach the user's privacy and be used for non authorized virtual profiling [14].

On the other hand, however, data collected from IoT devices can also be used for the common good and social profitability, as knowledge gathered from those data sets can help in improving people's quality of life, for example, by offering personalised services, precise and up-to-date information about people's surroundings and interests, and information for improving planning and efficiency of cities as well as for policy making [36]. Therefore, there is ample need for the design and implementation of mechanisms for both, enabling and preserving privacy that still allows significant IoT data to be collected, distributed, and processed in a useful manner.

Currently, cloud computing has been used to store and process data collected by IoT devices. This well-established computing paradigm is very convenient for data processing applications that require large processing capacity, and it fits well to computing problems that use data collected from many sources regardless of their location. However, as IoT devices are located and scattered at the very edge of the network, they can easily be brought to consuming data from neighbouring counterparts, which then would not require further transfer to a centralised cloud for processing [38]. In this scenario, Fog Computing [13] has emerged to bring and distribute computing capacities closer to the edge. In this computing paradigm, micro data centres, or cloudlets, are deployed throughout the network (e.g. at access points or cellular base stations) to bring computing capacities closer to the IoT devices. Combined, all three layers of edge IoT devices, fog devices, and cloud computing, can provide a hierarchical ubiquitous-like infrastructure for processing IoT data. Notwithstanding, privacy concerns are still present in this scenario and should be addressed. Computational models for SMPC that fit the proposed hierarchical structure in terms of resource allocation need to ensure the privacy of individual users (parties) [17].

In order to present our approach, we introduce next some concepts and discuss challenges in the study of the three components of the IoT-Fog-Cloud ecosystem with regard to privacy preserving algorithms and protocols considering secure multi-party interactions. Then, in section 4 we introduce our proposed architecture and discuss its instantiation. Section 5 presents preliminary results towards such an architecture implementation and in section 3 we discuss some available literature on these topics. Section 6 brings conclusions and remarks.

2 Background

2.1 IoT-Fog-Cloud ecosystem

The Internet of Things (IoT) consists of 'zillions' of (inter-) connected devices that can include sensors/actuators with communication capabilities, i.e., virtually any object with embedded micro-controller and antennae, which can be

highly heterogeneous at different levels, as for example energy consumption requirements, communications protocols, computing capacity, and mobility.

Therefore, IoT devices management becomes intrinsically challenging in terms of data communication and processing, while it requires significant processing and knowledge extraction capabilities in order to provide relevant insight and, hence, constitutes an ideal field for developing and testing of Big Data tools. The challenge is precisely to transform gathered data into actual information knowledge by (secure) privacy-preserving processing. Examples of data collected by IoT devices with such requirements are: location-based services, security cameras, personalised preference tracking mechanisms (e.g. for advertisement), medical data collected by body sensors or by professional equipment at hospitals and clinics, see for example [29].

Cloud computing represents nowadays a largely adopted computing paradigm servicing a variety of applications, providing dynamic configuration possibilities such as elasticity and pay-per-use. Resource virtualization allows providers to share slices of computing resources among users through the use of virtual machines and containers, where they appear to be logically isolated for each tenant. On-demand provisioning/de-provisioning, elasticity, ubiquitous access, lower upfront investments with reduced capital expenditures and faster time to market, are a few properties offered by cloud computing.

Clouds can partially fulfil applications requirements for IoT but may fall short for low latency, for example. Fog computing can be combined with the cloud to provide an ecosystem that furnishes full potential for IoT application deployments, as it introduces a hierarchy of computing capacity (fog nodes, cloudlets or micro data centres) between the edge and the cloud [13]. This hierarchy can provide a wider range of services that may not be supported solely by the cloud. Thus, a fog computing infrastructure can attend applications with a variety of Quality of Services (QoS) by using different hierarchical levels to meet latency and computing requirements.

A consequence of bringing processing closer to the edge with fog computing, as illustrated in Figure 1, is to reduce total bandwidth used in the network along the path between edge and cloud, as data can be aggregated at the edge before being transferred to the cloud. This computing hierarchy provided by the IoT-Fog-Cloud ecosystem enables thus processing to occur closer to where the data is being generated. This is useful for applications that depend upon very low delays or response times, but also for applications that only need data from constrained or specific geographic locations, as in this scenario the data does not need to be transferred all the way up to the cloud. Such a scenario demands the implementation of privacy-preserving mechanisms so that data shared from said IoT devices are not unintentionally available to third-parties.

Examples of such applications include real-time traffic estimation, autonomous cars traffic control, collision avoidance in non-signalised road intersections, localisation-based risk assessment, monitoring in disaster scenarios, actuators in dynamically changing environments, and so on.

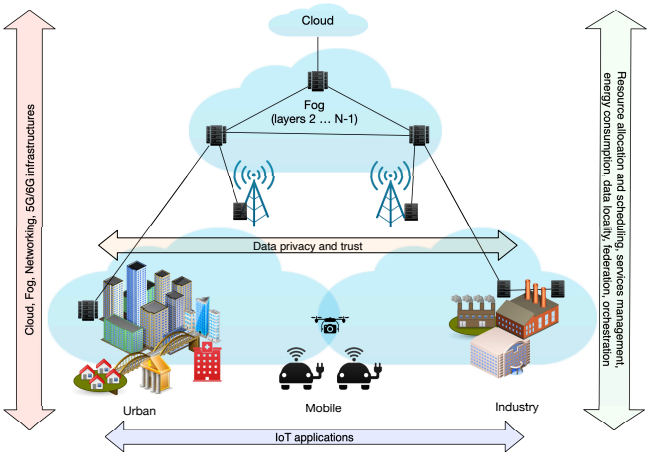


Fig. 1: Overview of the IoT-Fog-Cloud infrastructure (adapted from [4]).

We are interested in the modelling and design of efficient processing data protocols for applications in a privacy-preserving context. We posit that Secure Multi-Party Computation (SMPC) [16,19], understood as the joint computation of a desired function in terms of an appropriate splitting and distribution among a number of parties under the constraint that all data has to remain private, naturally furnishes a setting for privacy-preserving data processing and correctness:

- Privacy, in the sense that no party learns anything new other than its prescribed jointly computed output; and
- correctness, in the sense that parties can trust that the computed quantities are indeed correct.

This is useful, e.g., for the distributed evaluation of trust, as in [7], where parties or players compute a joint confidence level by combining their mutual degrees of trust. See the survey in [37] (and the references therein) for a nice description of the role of SMPC in the IoT field.

In view of potential applications, technologies should also consider, at least, designs providing Confidentiality, Integrity and Availability (CIA triad) [21]. Other requirements may be further imposed, for example on the side of independence of inputs, fairness, etc. (see [16]).

2.2 Privacy preservation in the IoT

Secure multiparty computations (SMPC) allows n players to compute together the output of some function, using private inputs without revealing them to other parties and without gaining access to knowledge entrusted to another party.

Notice that while SMPC protocols have been present in the literature for more than thirty years, their application to a data streams in a hierarchy of IoT

devices (in terms of computational power), which are continuously producing data worth analysing while preserving privacy, can be considered as a relatively recent development [15], in contrast to the more typical reception, storage and analysis of source data by a third-party centralised server that is not under the administrative control of the smart environment and hence potentially untrustworthy.

In the proposed architecture, the server should not obtain access to the raw data of the sources; it only fulfils management and orchestration purposes to carry out SMPC computations that are executed by the sources themselves [19]. This enables privacy-preserving data processing while the sources can be dynamic. The role of an edge-IoT gateway, understood as middle-ware between IoT devices and the cloud that facilitates computations and is a communication gateway as well [24], may be further extended to be able to receive queries regarding the outcomes of the computations still without revealing any private information about the sources.

Performance and resource consumption characteristics of SMPC methods are key limitations for their application in resource constrained environments. A detailed assessment of these characteristics is carried out in [20]. The authors reach the conclusion that SMPC methods can be applied, in practice, in Intranet environments, but with limitations in Internet settings. This shows the need for new SMPC-based solutions where the resource environment performs in close collaboration with privacy-preserving mechanisms.

3 Related work

A set of requirements for privacy preservation mechanisms in the context of distributed and ubiquitous environments, such as in the IoT, was presented in [1]. Among those requirements, they highlight the importance of ensuring privacy of the outputs of SMPC, which may be as important as the privacy of the inputs (as normally targeted in SMPC approaches), since data profiling might be used to de-anonymize the inputs or obtain sensitive information about the parties.

We address this issue to some extent, as the outputs at a lower level of the hierarchy are handled as inputs to the immediate upper level, and thus are subject to the usual privacy requirement. It remains to assess the effects of such privacy handling in relation to the use of the intermediate outputs for answering localised queries in a multi-layered fog architecture.

Some architectures for IoT assume that infrastructural elements in fog and edges are trusted entities [26]. Thus, to preserve data privacy, those architectures adopt data anonymization techniques such as differential privacy [8] based on a trust curator model [28]. In an untrusted component model, which is the center of this proposal, infrastructural components are not assumed as trusted entities so they should not access sensible data. This model usually adopts techniques of secure multi-party computation to preserve data privacy.

In IoT, the adoption of those techniques differs in terms of the mechanism for privacy preservation, resource usage adequacy and the architectural-style

for IoT platforms. Privacy preservation can be accomplished by differential privacy [8], homomorphic encryption (partial or full) [10] and secure multi-party computation. Adoption of those techniques have a direct impact on resource usage adequacy for IoT, but also in terms of requirements of processing power and exchanged messages among devices. In particular, they produce a huge increase in the length and number of messages, so they should be adapted (e.g. [2]) to be applied in large-scale scenarios of IoT. Section 2.2 above discusses relevant approaches for privacy preservation in the IoT.

The architectural-style of an IoT platform may be exploited to satisfy resource limitations (e.g., processing power) or to decrease the number of disseminated messages, by transferring processing from devices or servers to intermediary components. Approaches of privacy preservation for IoT assume architectures such as cloud-based [2] platforms, mixed-mesh [34] networks, edge networks [11], fog [30], and mixed architectures with trusted entities [1]. In particular, fog and edge-fog architectures allow multiple levels of processing and they are more adaptable to the dynamicity and diversity of IoT devices. We hypothesize that fog-edge architectures allow the development of efficient privacy preservation techniques, in terms of exchanged messages and distributed processing power.

4 Privacy towards the Edge

This section describes an overview for the proposed privacy-preserving multi-party computation architecture for IoT using fog computing.

4.1 Hypotheses and Objectives

In order to develop a framework of theoretical nature and a test pilot for the hierarchical architecture in layers: from IoT to Edge and from Edge to Fog computing, in which multi party protocols can be proposed, we first consider current techniques and methods for Fog/Edge technologies and Multi party computations.

Since cloud computing cannot rely exclusively upon cryptographic protocols alone [35] and, furthermore, the data collected by cyber-physical (IoT) sensors is known to be securely (and privately) processed using zero-knowledge proofs, as in [12], we propose to undertake our approach of Privacy Preserving Multi Party Computation for Data Analytic in the Iot-Edge-Fog ecosystem by first establishing a hierarchy of devices, as in Figure 2 and then to follow the lines proposed by [32] and [23].

4.2 Architectural Principles

At the edge, we propose that IoT devices are grouped, either statically or dynamically, to be considered as data generators for a specific application or user query. Grouping is based on criteria such as resource allocation (at the fog),

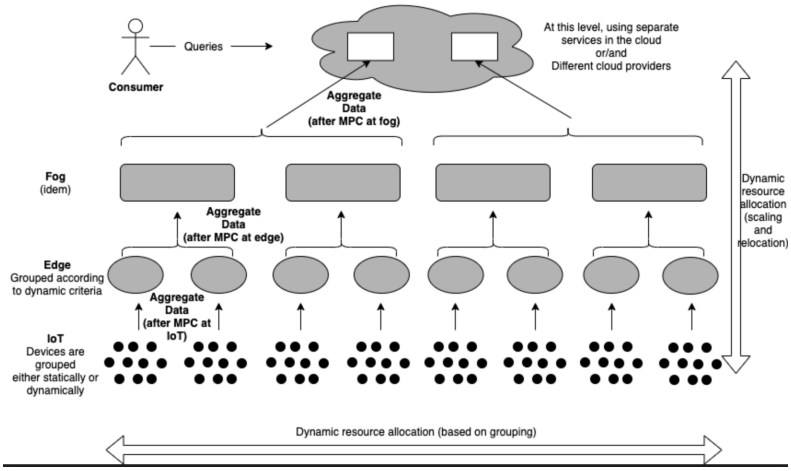


Fig. 2: Edge/Fog Computing hierarchy together with secure multi-party computation brings resource allocation challenges.

scalability and geographic location. These devices are assumed to be capable of executing SMPC in the context of the group which they currently belong to. The output of such computation may either be used immediately, to answer localised queries posed by users/consumers, or sent up (to services running in the same or a different micro-datacenter in the fog) as intermediate outputs that in turn are used in conjunction with outputs from other SMPC computations performed by other IoT device groups. This pattern of dynamically grouping SMPC parties that receive inputs from SMPC groups located underneath in the hierarchy repeats at the fog and cloud levels to produce intermediate outputs and/or to answer user queries.

To illustrate how this works, we consider the current pandemic scenario: Consider an end-user who is willing to know how safe it is to go to a specific place. Assume we can come up with a safety metric based on several pieces of information from the neighbourhood and/or place where this user is willing to go. A combination of data from IoT devices at the location of interest can be used to compute such safety metric and respond to the user query. However, it is clear that such information can often include personal data that raises privacy concerns, as for example medical conditions of people in that area, real-time and accurate information about people's location, data on how those people interact or have interacted on that location, information collected from cameras about the amount of people wearing masks, and so on. A privacy-preserving mechanism using multi-party computation would need to first choose sources to obtain all that data to be processed, and this is an example of how devices are to be grouped into sets that are able to generate data in response to a query. Then, the data set obtained from those devices can be first processed using multi-party computation protocols, and then be further moved to the closest cloudlet/micro

data center at the fog level for further processing. With privacy already in place, it remains to aggregate the data and compute the necessary safety metrics of interest to that query.

Note that more complex scenarios, with more data or data from multiple geographic locations, can involve multiple cloudlets at the fog level or even involve the cloud to provide enough capacity or enough data sets to allow a proper computation of the desired metric.

To enable SMPC in the discussed scenario, many challenges arise. First, a candidate set of SMPC protocols must be identified, studied, and potentially adapted for the proposed infrastructure. Second, the composition of IoT device groups should be enabled by distributed computing mechanisms. Third, once these groups have been defined for a specific application or query, data processing should occur in an efficient manner throughout the system through proper resource allocation and scheduling mechanisms.

4.3 Architecture Instantiation

It is known that effective mechanisms addressing stacked fog layers (cloudlets) constitutes still a challenge and open issue [4], since a universal standard of multi-layered fog computing architecture has not yet been adopted. We address this issue by considering data privacy aspects as inherent to a multi-layered fog architecture, in which resource allocation mechanisms are needed to efficiently process privacy-sensitive information.

Indeed, Fog computing technology enables the deployment of cloud-like computing services at the edge of the network. Fog-aware mechanisms and procedures to implement efficient resource management for applications with different levels of latency requirements remains a non trivial challenge [18]. Several efforts have addressed this body of work to provide a latency-driven fog computing environment for IoT applications using only one fog layer (i.e. *cloudlets*, *micro data centres*, or *fog nodes*) between the edge and the cloud (as illustrated in the proposed fog-cloud topology in Figure 1) [31,3,25].

Next, it is common knowledge that some architectures [26] for IoT assume that infrastructural elements in fog and edges are trusted entities. Thus, to preserve data privacy, those architectures may adopt data randomisation techniques such as differential privacy [8] based on a trust curator model [28]. But, in a complete untrustworthy component model, which is at the centre of our considerations, infrastructural components are not necessarily assumed to be trusted entities, so they should not access sensible data. We propose hence that our model should adopt techniques of secure multi-party computation to preserve data privacy.

Adoption of these techniques will have a direct impact on resource usage adequacy for IoT, in terms of processing power and exchanged messages among devices. In particular, they produce a huge increase in both the length and the number of messages, so they should be adapted to large-scale scenarios with IoT.

Finally, for the deployment of these protocols, end-to-end automated systems for deploying large-scale SMPC protocols between end users, called MPSaaS (acronym for system-as-a-service) have been designed and tested [2].

5 Preliminary Results and Discussion

In the field of smart homes with IoT, secure communications between smart devices endowed with low computational capabilities and a home gateway are frequently established via unsecured wireless communication channels. In [5] we explored secure data transfer generated by IoT using standard Diffie-Hellman secret sharing which can be applied via buffering to higher throughput communications. Our assumptions were that at setup both the user and the gateway have enough processing power to perform - say - one time secured RSA encrypted communication, hence relaxing the need for a trusted secure server outside the domain, and that the protocol should at least be secure for a range of known attacks, as replay or DoS attacks. In turn, the approach failed impersonation attacks via Man in the Middle (MiM) so that a possible way out may lie in the implementation of SMPC as proposed in the architecture proposed here, since authentication would require a consensus (for example defined using Shamir's secret schem) to be build upon trusted devices. Hence, as mentioned above, joint computation of trust among parties can be performed by matrix multiplication, as in [6], where we proposed an *ad hoc* Strassen-Winograd multiplication protocol that use a combination of partial homomorphic encryption schemes and additive masking techniques together with a (novel) schedule for the location and encryption layout of all intermediate computations, such that privacy is preserved. It turned out that the asymptotic communication volume and computational time was reduced from $O(n^3)$ to $O(n^{2.81})$.

Thus, the main challenge now is to incorporate the developed protocols to the proposed computing hierarchy. This involves the prior study of performance of devices from the edge to the cloud to evaluate the possible ways of allocating resources for protocol processing and also assess communication costs. As a second step, evaluating the performance of the protocols in a real-world scenario, considering requirements as the amount of data inputs and needed response time (e.g. in the pandemics scenario described in Section 4), is key for establishing limitations and pinpoint where research must be focused to enable performance-effective multiparty computation at the edge.

This involves developing optimization algorithms for dynamic grouping or clustering, as presented in Figure 2, which should consider multiparty computation aspects both in terms of privacy and response times: a too large set of IoT devices may impair performance for response time, while a too small set may result in poor output for the application requiring data collection and processing. To evaluate these aspects in the platform, we plan to assess metrics such as response time when compared to a centralized cloud deployment and accuracy of deployed algorithms for grouping optimization. Also, networking metrics will be collected to evaluate how/if the framework can reduce data traffic while

maintaining processing capacity and accuracy at the same level as a centralized system. Lastly, power consumption of devices using multi-party computation can be evaluated to learn if the framework can be used in practice without impairing users' experience with mobile devices.

6 Conclusions and future work

This work presented the main concepts and initial findings about the design and evaluation of privacy preserving protocols in an IoT-Fog-Cloud ecosystem computing hierarchy. We briefly discussed how this can be applied in a scenario as the present times of pandemics, but also covered general aspects that need attention for a generalized application of privacy-preserving mechanisms at the edge of the network.

We considered the paradigms of Fog and Edge computing, together with a multi-party computation mechanism to enable secure privacy-preserving data processing at the edge. The proposed architecture demands further developments in terms of understanding performance limitations and also in terms of resource allocation for data collection and processing of multiparty computation protocols, which can be the focus for further research. Also, the formalization of the proposal, with the modeling of inputs and outputs for multiparty computation at the edge, is a key future work to enable proper evaluation of efficiency and limitations of the proposal.

Acknowledgments

This work was partially funded by the São Paulo Research Foundation (FAPESP), grants #2018/23126-3 and #2015/24494-8, CAPES, and CNPq, Brazil, grants 432943/2018-8 and 309562/2019-8. Partial support from Universidad de La Frontera, Vicerrectoría de Investigación grant DI20-0060 is also acknowledged.

References

1. Ankele, R., Küçük, K.A., Martin, A., Simpson, A., Paverd, A.: Applying the trustworthy remote entity to privacy-preserving multiparty computation: Requirements and criteria for large-scale applications. In: 2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld). pp. 414–422 (2016)
2. Barak, A., Hirt, M., Koskas, L., Lindell, Y.: An end-to-end system for large scale p2p mpc-as-a-service and low-bandwidth mpc for weak participants. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. pp. 695–712 (2018)
3. Bittencourt, L., Diaz-Montes, J., Buyya, R., Rana, O., Parashar, M.: Mobility-Aware Application Scheduling in Fog Computing. IEEE Cloud Computing **4**(2), 26–35 (March 2017)

4. Bittencourt, L., Immich, R., Sakellariou, R., Fonseca, N., Madeira, E., Curado, M., Villas, L., DaSilva, L., Lee, C., Rana, O.: The internet of things, fog and cloud continuum: Integration and challenges. Internet of Things **3-4**, 134 – 155 (2018), http://www.sciencedirect.com/science/article/pii/S2542660518300635
5. Díaz Arancibia, J., Ferrari Smith, V., López Fenner, J.: On-The-Fly Diffie-Hellman for IoT. In: 2019 38th International Conference of the Chilean Computer Science Society (SCCC). pp. 1–5. IEEE (2019)
6. Dumas, J.G., Lafourcade, P., Lopez Fenner, J., Lucas, D., Orfila, J.B., Pernet, C., Puys, M.: Secure multiparty matrix multiplication based on strassen-winograd algorithm. In: International Workshop on Security. pp. 67–88. Springer (2019)
7. Dumas, J.G., Lafourcade, P., Orfila, J.B., Puys, M.: Dual protocols for private multi-party matrix multiplication and trust computations. Computers & security **71**, 51–70 (2017)
8. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) Theory of Cryptography. pp. 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
9. Elijah, O., Rahman, T.A., Orikumhi, I., Leow, C.Y., Hindia, M.N.: An overview of internet of things (iot) and data analytics in agriculture: Benefits and challenges. IEEE Internet of Things Journal **5**(5), 3758–3773 (2018)
10. Gentry, C.: Fully homomorphic encryption using ideal lattices. In: Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing. p. 169–178. STOC '09, Association for Computing Machinery, New York, NY, USA (2009)
11. Ghinita, G., Kalnis, P., Khoshgozaran, A., Shahabi, C., Tan, K.L.: Private queries in location based services: Anonymizers are not necessary. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. p. 121–132. SIGMOD '08, Association for Computing Machinery, New York, NY, USA (2008)
12. Goldwasser, S., Micali, S., Rackoff, C.: The knowledge complexity of interactive proof systems. SIAM Journal on computing **18**(1), 186–208 (1989)
13. Group, O.C.A.W., et al.: Openfog reference architecture for fog computing. OPFRA001 **20817**, 162 (2017)
14. Hossain, E., Khan, I., Un-Noor, F., Sikander, S.S., Sunny, M.S.H.: Application of big data and machine learning in smart grid, and associated security concerns: A review. IEEE Access **7**, 13960–13988 (2019)
15. Kerschbaum, F.: Privacy-preserving computation. In: Annual Privacy Forum. pp. 41–54. Springer (2012)
16. Lindell, Y.: Secure multiparty computation for privacy preserving data mining. In: Encyclopedia of Data Warehousing and Mining, pp. 1005–1009. IGI Global (2005)
17. Liu, W., Luo, S.s., Wang, Y.b., Jiang, Z.t.: A protocol of secure multi-party multi-data ranking and its application in privacy preserving sequential pattern mining. In: 2011 Fourth International Joint Conference on Computational Sciences and Optimization. pp. 272–275. IEEE (2011)
18. Mahmud, R., Ramamohanarao, K., Buyya, R.: Latency-Aware Application Module Management for Fog Computing Environments. ACM Trans. Internet Technol. **19**(1), 9:1–9:21 (Nov 2018)
19. von Maltitz, M., Bitzer, D., Carle, G.: Data Querying and Access Control for Secure Multiparty Computation . In: 2019 IFIP/IEEE International Symposium on Integrated Network Management. Washington, DC, USA (2019)
20. von Maltitz, M., Carle, G.: A performance and resource consumption assessment of secret sharing based secure multiparty computation. In: Data Privacy Management Workshop '18. Springer International Publishing (2018)

21. von Maltitz, M.L.: Secure and Privacy-preserving Services Based on Secure Multiparty Computation. Network Architectures and Services NET (2019)
22. Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M.: Deep learning for iot big data and streaming analytics: A survey. IEEE Communications Surveys & Tutorials **20**(4), 2923–2960 (2018)
23. Pallas, F., Raschke, P., Bermbach, D.: Fog computing as privacy enabler. IEEE Internet Computing (2020)
24. Papcun, P., Kajati, E., Cupkova, D., Mocnej, J., Miskuf, M., Zolotova, I.: Edge-enabled iot gateway criteria selection and evaluation. Concurrency and Computation: Practice and Experience p. e5219 (2020)
25. Peterson, L., Anderson, T., Katti, S., McKeown, N., Parulkar, G., Rexford, J., Satyanarayanan, M., Sunay, O., Vahdat, A.: Democratizing the Network Edge. SIGCOMM Comput. Commun. Rev. **49**(2), 31–36 (May 2019)
26. Pinto, S., Gomes, T., Pereira, J., Cabral, J., Tavares, A.: Iioteed: An enhanced, trusted execution environment for industrial iot edge devices. IEEE Internet Computing **21**(1), 40–47 (2017)
27. Raafat, H.M., Hossain, M.S., Essa, E., Elmougy, S., Tolba, A.S., Muhammad, G., Ghoneim, A.: Fog intelligence for real-time iot sensor data analytics. IEEE Access **5**, 24062–24069 (2017)
28. Rao, F.Y., Bertino, E.: Privacy techniques for edge computing systems. Proceedings of the IEEE **107**(8), 1632–1654 (2019)
29. Reddy, K.U.K., Shabbiha, S., Kumar, M.: Design of High Security Smart Health Care Monitoring System using IoT. International Journal of Emerging Trends in Engineering Research **8**(6), 2259–2265 (June 2020)
30. Saleem, A., Khan, A., Malik, S.U.R., Pervaiz, H., Malik, H., Alam, M., Jindal, A.: Fesda: Fog-enabled secure data aggregation in smart grid iot network. IEEE Internet of Things Journal pp. 1–1 (2019)
31. Shah-Mansouri, H., Wong, V.W.S.: Hierarchical Fog-Cloud Computing for IoT Systems: A Computation Offloading Game. IEEE Internet of Things Journal **5**(4), 3246–3257 (2018)
32. Sousa, P.R., Antunes, L., Martins, R.: The present and future of privacy-preserving computation in fog computing. In: Fog Computing in the Internet of Things, pp. 51–69. Springer (2018)
33. Tao, H., Bhuiyan, M.Z.A., Abdalla, A.N., Hassan, M.M., Zain, J.M., Hayajneh, T.: Secured data collection with hardware-based ciphers for iot-based healthcare. IEEE Internet of Things Journal **6**(1), 410–420 (2018)
34. Tonyali, S., Akkaya, K., Saputro, N., Uluagac, A.S., Nojoumian, M.: Privacy-preserving protocols for secure and reliable data aggregation in iot-enabled smart metering systems. Future Generation Computer Systems **78**, 547–557 (2018)
35. Van Dijk, M., Juels, A.: On the impossibility of cryptography alone for privacy-preserving cloud computing. HotSec **10**, 1–8 (2010)
36. Vorakulpipat, C., Rattanalerdnusorn, E., Thaenkaew, P., Hai, H.D.: Recent challenges, trends, and concerns related to iot security: An evolutionary study. In: 2018 20th International Conference on Advanced Communication Technology (ICACT). pp. 405–410. IEEE (2018)
37. Yan, Z., Zhang, P., Vasilakos, A.V.: A survey on trust management for internet of things. Journal of network and computer applications **42**, 120–134 (2014)
38. Zhang, Y., Ren, J., Liu, J., Xu, C., Guo, H., Liu, Y.: A survey on emerging computing paradigms for big data. Chinese Journal of Electronics **26**(1), 1–12 (2017)

APPADS. Propuesta de una Aplicación Open Source Interactiva, Reactiva y Dinámica para el Análisis Simbólico de los Datos

Scheffer María del Valle [1], Herrera Myriam Beatriz[1], Romagnano María[1]

[1] Instituto de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Argentina
{maruscheff, myriamhrrr}@gmail.com, maritaroma@iinfo.unsj.edu.ar

Resumen. Las organizaciones diariamente tienen que lidiar con millones de datos que deben analizar, procesar, visualizar para lograr una toma de decisión relevante. Datos de diferente naturaleza extraídos de variables categóricas, no categóricas, etc, que proceden de múltiples dispositivos y son parte de nuestro día a día. Estos proceden de entidades individuales para las cuales se graba un valor individual en cada variable. Pero cuando los elementos de interés son clases o grupos de algún tipo, como ciudadanos que viven en una determinada ciudad o alumnos o egresados que provienen de determinadas carreras, por ejemplo, hay variabilidad inherente en los datos. El análisis basado en este interés podría determinar patrones de comportamiento para apoyar a la toma de decisiones. Así mismo, resulta necesario y de interés asegurar que la toma de decisiones sea adecuada, y para ello se debe garantizar que los datos sean correctos y que la información sea la indicada. El análisis de datos simbólicos (ADS) proporciona un marco que permite representar datos con variabilidad, usando nuevos tipos de variables. El presente trabajo propone una herramienta open source, alternativa de SODAS (Symbolic Official Data Analysis System). Se trata de una aplicación web interactiva que utiliza el concepto de programación reactiva, debido a que el usuario puede interactuar con los datos sin tener que manipular el código. Por otra parte, el concepto de reactividad hace referencia al hecho de poder vincular los valores de entrada con los de salida permitiendo que la aplicación responda dinámicamente ante un evento.

Palabras claves: Dato Simbólico, Ciencia de los Datos, Aplicación Web.

1 Introducción

Permanentemente se cuenta con información socio-económica de poblaciones para lo cual el análisis de parámetros de tendencia, dispersión y distribuciones son importantes. Para ello se utilizan diferentes herramientas y técnicas descriptivas propias de la Ciencia de los Datos. Esta última es un campo interdisciplinario que

agrupa Estadística, Inteligencia Artificial (incluyendo el Aprendizaje de Máquina), Minería de Datos, etc. Además, involucra técnicas de procesamiento, análisis y visualización que permiten la extracción de información con un enfoque más holístico de los datos, para su posterior conversión a un mensaje manejable [1].

Los algoritmos desarrollados en el contexto del análisis multidimensional de datos se adaptan a diferentes niveles de complejidad de la información: datos numéricos, textuales, simbólicos. Es decir que el dato puede ser algo más que un único valor numérico resultado de la asignación de una medida o código a una unidad de análisis: puede ser una palabra, un conocimiento, una posibilidad, una conjunción de valores, en fin, un objeto simbólico.

Específicamente uno de los aportes fundamentales del Análisis Multivariado de Datos consiste en otorgar la mayor importancia a los individuos. La estadística clásica se interesa más en las variables, ya sea buscando regularidades o leyes de probabilidad y los individuos en sí mismos no son considerados más que en su totalidad.

En el caso de los métodos factoriales los individuos aparecen proyectados en el espacio generado por los ejes factoriales y es analizada su estructura al mismo tiempo que la de las variables. Complementariamente esos mismos individuos son sometidos a un proceso de clasificación.

En la perspectiva de Diday [2] acerca de la clasificación, el énfasis está puesto más que en los individuos, en los objetos mismos. De esta manera se sube un nivel en generalización.

El objetivo del presente trabajo es una herramienta open source, una alternativa al software SODAS. Se trata de una aplicación web interactiva que utiliza el concepto de Programación Reactiva. Facilita a los expertos en el Análisis de los Objetos Simbólicos, y que colaboran en la toma de decisiones, poder obtener de manera sencilla y amigable a partir de un archivo de texto plano, la generación de los objetos simbólicos, la tabla correspondiente a los mismos; como así también una representación gráfica a través de la cual puedan establecer a simple vista una relación entre estos de manera Interactiva y Dinámica haciendo uso del concepto de Programación Reactiva.

En adelante, el trabajo se estructura de la siguiente forma: Metodología, Desarrollo, Propuesta, Descripción de la aplicación web y Conclusiones.

2 Metodología

En la mayoría de las organizaciones actuales, con o sin fines de lucro, el cúmulo de datos a procesar es exponencial. El trabajar con bases de datos de grandes volúmenes es una tarea a diario.

Estas bases de datos contienen elementos o componentes los cuales son descritos por variables que generalmente pertenecen a una categoría, o son numéricas. Así éstas últimas pueden ser transformadas para incluirse en una categoría.

Para el procesamiento de los datos la Inteligencia Artificial ha provisto de subdisciplinas que se especializan en el procesamiento de los datos, tales como Big

Data, Data Mining, Descubrimiento del conocimiento a través de patrones, Adquisición del conocimiento, entre otras [3], [4].

Por lo general se trabaja con datos que proceden de observaciones únicas, de determinadas variables sobre individuos únicos. Pero cuando los elementos de interés son clases o grupos de algún tipo, como los ciudadanos que viven en una ciudad determinada o alumnos o egresados que provienen de determinadas carreras, por ejemplo, hay variabilidad inherente en los datos.

Las bases de datos pueden extenderse en profundidad como en ancho, esto quiere decir en cantidad de datos y cantidad de variables a analizar, lo cual presenta ciertas características en cuanto al procesamiento:

- *Gran cantidad de columnas.* Los analistas muchas veces deben limitar el número de variables a examinar cuando realizan análisis manual, debido a limitaciones de tiempo. Sin embargo, variables que son descartadas porque parecen no poseer importancia pueden proveer información acerca de modelos desconocidos. Si se cuenta con un proceso de Data Mining de alto rendimiento se puede explorar toda la base de datos, sin necesidad de preseleccionar un subconjunto de variables.
- *Gran cantidad de filas.* Muestras mayores producen menos errores de estimación y desvíos, y permite a los científicos de datos hacer inferencias acerca de pequeños pero importantes segmentos de población.

Resulta importante resumir estos datos en término de sus conceptos con el sentido de extraer nuevos conocimientos. Es decir, identificar un concepto que sea representativo de un conjunto de datos. Estos conceptos se pueden describir por tipos de datos más complejos, llamados Datos Simbólicos, que contienen variación interna y son estructurados de modo que crea un puente entre la Estadística y el Aprendizaje Automático.

Diday [1], [2] formaliza los conceptos de intención y extensión. La intención de un concepto constituye su descripción, mientras que la extensión es el conjunto de individuos cuya descripción es acorde a la del concepto.

La intención, que se representa por un objeto simbólico, se describe por los datos simbólicos y por un mecanismo de reconocimiento de los individuos de la extensión. Diday introduce los objetos simbólicos y presenta una formalización que permite tratar conocimientos más ricos que los datos habituales y establece una relación con el modelo clásico de Análisis de Datos [2], [6].

Un objeto simbólico representa una intención, un concepto y se define, en términos generales, como una conjunción de valores, o conjuntos de valores que pueden ser ponderados. Constituye una descripción en intención de una clase de individuos que constituyen la extensión.

Los objetos simbólicos representan conceptos, entendidos como la intención y extensión del mismo. La intención de un concepto representa las propiedades que lo definen y que lo hacen distinto de los demás conceptos. La extensión de un concepto se compone de los individuos que se definen por el concepto o que cumplen las propiedades que definen el concepto. Se describen por variables y datos simbólicos y proporcionan un mecanismo de vuelta a bases de datos o conjuntos de individuos en el sentido de conocer aquellos que se adecuan o relacionan con las descripciones

simbólicas representadas por los objetos (las intenciones), según determinadas relaciones que también forman parte de las intenciones.

El Análisis de Datos Simbólicos es una generalización de las técnicas de Análisis de Datos aplicadas a matrices de datos simbólicos. Las definiciones y notación de variables simbólicas y objetos simbólicos han estado en constante evolución desde sus inicios.

2.1 Datos Simbólicos

Sea $\Omega = \{\omega_1,..., \omega_k\}$ un conjunto de individuos, sea Y un conjunto o dominio de posibles valores observados y $E = \{e_1,..., e_n\}$ un conjunto de objetos. Como casos particulares más frecuentes se tiene que E es un subconjunto de Ω o un subconjunto de las clases de Ω, es decir, $E \subseteq \Omega$ o $E \subseteq P(\Omega)$. En el segundo caso, los datos simbólicos correspondientes, describen clases de individuos de Ω.

La descripción de un elemento $e \in E$ por una variable con dominio Y puede darse por: un elemento del conjunto Y (variable clásica), un subconjunto de elementos del conjunto Y, un intervalo de valores del conjunto Y donde se ha definido un orden, un subconjunto de elementos del conjunto Y donde cada uno de ellos es ponderado por un peso o modo.

La matriz de datos en el Análisis de Datos Simbólicos es la matriz [X], cuyas filas representan n unidades u objetos del conjunto E descritos por un vector de variables simbólicas $X = (X_1,..., X_p)$. Es decir, las celdas de una fila se corresponden con los datos simbólicos descritos por el vector X aplicado a un elemento de E.

El Análisis de Datos Simbólico (ADS) tiene como objetivo reemplazar los individuos del análisis de datos tradicional por individuos de más alto nivel, más complejos y más aptos para representar conocimientos, porque están definidos en intensión, utilizando el poder de la lógica: son los objetos simbólicos (OS) [7].

También se puede decir que las variables son de más alto nivel en el análisis de datos simbólicos, porque las variables no van a tomar un sólo valor por cada celda, sino que pueden tomar varios valores. Por ejemplo: cuando se describe una clase, los individuos de la clase pueden tomar distintos valores. Si se describe una clase de empresas, que tienen beneficios de distinto orden, se puede tomar el beneficio en un intervalo para esta clase de empresas. Si una empresa pertenece a una determinada clase, en la variable beneficio tendré un intervalo de valores (y no uno sólo). Además, es muy importante porque va a plantear todos los problemas teóricos también a un mayor nivel, se va a subir un nivel en toda la teoría del AMD.

En la estadística clásica, cuando se define una variable aleatoria, por ejemplo, la variable aleatoria x, toma valores en un determinado intervalo y en este caso cuando se escribe: x = [3,5] se dice que este suceso correspondiente a esta variable aleatoria se refiere a un conjunto de individuos en el espacio Ω (los individuos cuyos valores en la variable x varían entre 3 y 5). El objetivo es estudiar las leyes asociadas a estas variables aleatorias. Estas leyes rigen sobre los individuos de Ω, siendo Ω el espacio muestral de los valores obtenidos.

En el ADS, Ω no es un conjunto de valores tomados por los individuos sino es el conjunto de eventos o sucesos. A continuación, se muestra la manera de encontrar a partir de los datos el OS y el dato simbólico con un ejemplo en particular.

3 Desarrollo

Se partió de una matriz de datos correspondientes a egresados de la FCEFN de la UNSJ de orden nxp, cuyas variables corresponden a las opiniones de los egresados respecto a la parte administrativa de la facultad, el aula, la biblioteca y el laboratorio. Si trabajan en relación a la carrera de la que egresaron, la opinión de la UNSJ en la comunidad en la que se encuentran, y la carrera a qué departamento corresponde en la facultad. De acuerdo a las necesidades de los usuarios, los departamentos de la Facultad cuentan con la matriz de datos de orden nxp, siendo n los departamentos y p las variables.

Los datos simbólicos implican tablas de datos más complejas llamadas tablas de datos simbólicos. Una celda de tales tablas no necesariamente contiene valores categóricos o cuantitativos simples, sino muchos valores, que pueden tener pesos o estar unidos por reglas lógicas y taxonomías. Por ejemplo, una celda puede contener un intervalo o una distribución [5].

El análisis de este tipo de datos fue generado para analizar y tomar decisiones sobre grandes bases de datos, especialmente para datos de encuestas. Si bien resumen en gran proporción las bases de datos, éstas preservan los datos esenciales. Además, permiten visualizar, comparar y clasificar objetos.

En el ADS, el espacio muestral utilizado para determinar la base de datos no es un conjunto de valores tomados por un individuo, sino es el conjunto de eventos o sucesos, así es como se extiende la teoría. Por ello se pudieron demostrar axiomas básicos de la teoría probabilística [5].

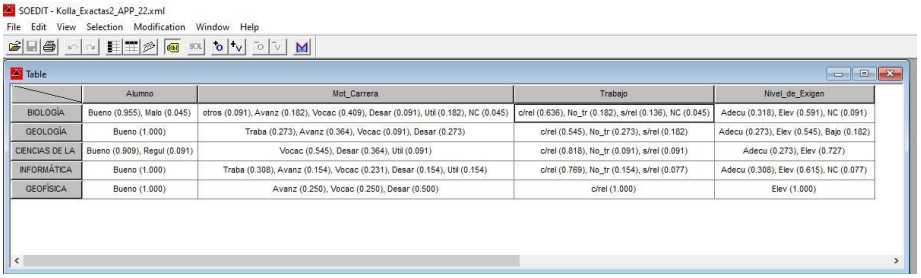
La herramienta informática para el manejo de este tipo de datos es SODAS (Symbolic Official Data Analysis System), un software realizado con la cooperación de 17 grupos de investigación europeos y tres Institutos Nacionales de Estadística: EUSTAS/ España – INE/ Portugal – ONS/ Londres.

Dado un conjunto de datos, obtenido a partir de la recolección y limpieza de los datos con los que se desea trabajar, el análisis de datos simbólicos se realiza utilizando dicho software. Pero para ello, es necesario realizar una serie de pasos diferentes y complejos:

1. Así, este software sólo permite importar archivos cuya extensión sea la del propio software; archivos con extensión .sds y/o archivos cuya extensión sea .xml. Por lo tanto, si en una primera instancia sólo se cuenta con un conjunto de datos que se encuentren en un formato de archivo de Excel o de texto plano, separado por comas (.csv), es necesario previamente procesarlo desde la herramienta Microsoft Access y a través de la misma seleccionar la variable candidata a ser el Objeto Simbólico, determinando que la primera columna, la identificación (ID) indica cuales son los objetos simbólicos, siendo en este caso “Departamento” la variable candidata.
2. Una vez realizado todo este proceso, se obtiene el archivo con el cual se procede a realizar el análisis de datos simbólicos haciendo uso de SODAS (ver Figura 1).

Las evidencias muestran que la manera de encontrar la tabla simbólica es muy laboriosa y además este trabajo se debe hacer cada vez que el experto indique el OS con el cual desea trabajar. Por lo tanto, se considera necesario contar con un software libre al cual pueda utilizar el experto de manera sencilla y que le permita definir nuevos OS de manera interactiva sin necesidad de realizar nuevamente el proceso desde cero como es en el caso de SODAS.

La posibilidad de contar con una aplicación web interactiva bajo el concepto de reactividad y de software libre la cual permite trabajar con cualquier conjunto de datos que posean un formato con extensión .csv y que permita seleccionar el objeto interactivamente de acuerdo a la necesidad del experto, parece ser muy beneficiosa debido a que no sería necesario llevar a cabo la serie de pasos previos que SODAS necesita para poder generar la tabla de OS.


SOEDIT - Kolla_Exactas2_APP_22.xml

File Edit View Selection Modification Window Help

Table

	Alumno	Mot_Carrera	Trabajo	Nivel_de_Exigen
BIOLOGÍA	Bueno (0.955), Malo (0.045)	otros (0.091), Avanz (0.182), Vocac (0.409), Desar (0.091), Util (0.182), NC (0.045)	c/rel (0.636), No_tr (0.182), s/rel (0.136), NC (0.045)	Adecu (0.318), Elev (0.591), NC (0.091)
GEOLOGÍA	Bueno (1.000)	Traba (0.273), Avanz (0.364), Vocac (0.091), Desar (0.273)	c/rel (0.545), No_tr (0.273), s/rel (0.182)	Adecu (0.273), Elev (0.545), Bajo (0.182)
CIENCIAS DE LA	Bueno (0.909), Regul (0.091)	Vocac (0.545), Desar (0.364), Util (0.091)	c/rel (0.818), No_tr (0.091), s/rel (0.091)	Adecu (0.273), Elev (0.727)
INFORMÁTICA	Bueno (1.000)	Traba (0.308), Avanz (0.154), Vocac (0.231), Desar (0.154), Util (0.154)	c/rel (0.769), No_tr (0.154), s/rel (0.077)	Adecu (0.308), Elev (0.615), NC (0.077)
GEOFÍSICA	Bueno (1.000)	Avanz (0.250), Vocac (0.250), Desar (0.500)	c/rel (1.000)	Elev (1.000)

Fig. 1. Tabla de Objetos Simbólicos obtenida con SODAS a partir del archivo generado en Microsoft Access.

4 Propuesta

El desarrollo de la Aplicación Web Interactiva y Reactiva, a la cual hemos denominado APPADS (APP de Análisis de Datos Simbólicos), fue realizado mediante el uso del lenguaje de programación R. Para ello se utilizó el entorno de RStudio. Por otra parte, también se hizo uso de los paquetes Shiny de RStudio y RSDA de R, siendo este último el que permitió la creación del Objeto Simbólico a partir de las distintas variables candidatas como así también la generación de la Tabla de Objetos Simbólicos, necesarios para llevar a cabo el análisis.

5 Descripición de la Aplicación Web

Si bien en un principio se trata de una aplicación web, la misma puede ser ejecutada de manera local en cualquier ordenador en el que se encuentre instalado el entorno de desarrollo RStudio. Esto se debe a que la aplicación está basada en una arquitectura del tipo cliente/servidor haciendo uso de tecnología websocket, entre otras. Utiliza parámetros y variables dinámicas que dotan a la aplicación de la capacidad de permitir al usuario interactuar con los datos y obtener diferentes resultados en función de los mismos.

En el contexto de la creación del objeto simbólico, esto permite que el usuario pueda seleccionar la variable candidata a ser convertida en el objeto simbólico entre un conjunto de variables previamente seleccionadas; generar en función de la misma la tabla de objetos simbólicos, visualizar su representación gráfica, y si lo desea seleccionar otra variable candidata y ver automáticamente reflejados todos los cambios de acuerdo a los nuevos datos generados a partir del nuevo objeto simbólico creado.

5.1 Detalles de las entradas

En la Figura 2 puede observarse que dado un conjunto de datos que se encuentra principalmente en formato de archivo .csv, la aplicación permite al usuario seleccionarlo desde la ubicación en la que se encuentre almacenada dentro del computador. El tamaño de los archivos puede llegar a ser de hasta 30 MB, y previo a la carga de los mismos, se debe seleccionar desde la aplicación si los mismos se encuentran en un formato separados por coma, punto y coma o tabulación. Luego en la Figura 3 se muestra el archivo seleccionado, con información detallada respecto al mismo y con el cual se comenzará a trabajar.

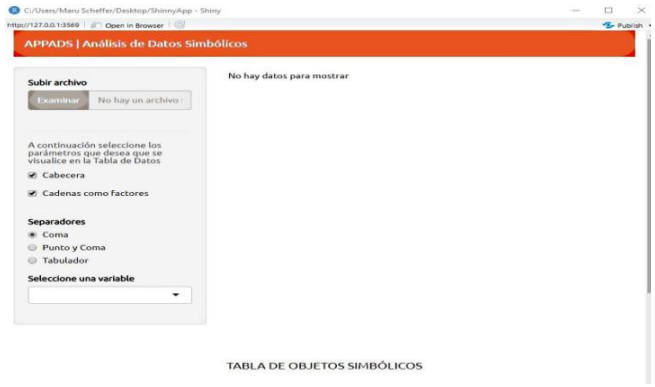


Fig. 2. Selección del tipo de formato del archivo fuente.

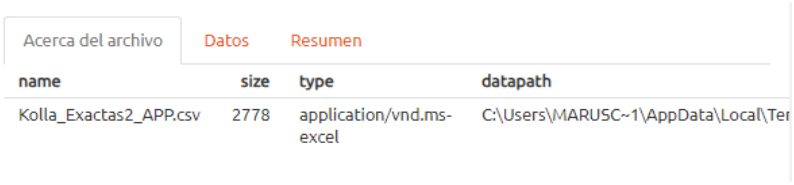


Fig. 3. Archivo fuente seleccionado

5.2 Detalles del Procesamiento

Una vez realizada la carga del archivo seleccionado, la aplicación brinda la posibilidad de seleccionar de manera interactiva la variable candidata a partir de la

cual se creará el Objeto Simbólico (ver Figura 4). Inmediatamente, la aplicación brinda información relacionada con el conjunto de datos del archivo, construye la tabla de Datos Simbólicos y muestra en pantalla una gráfica de radar que permite realizar el análisis entre las distintas clases que forman parte del objeto en cuestión.

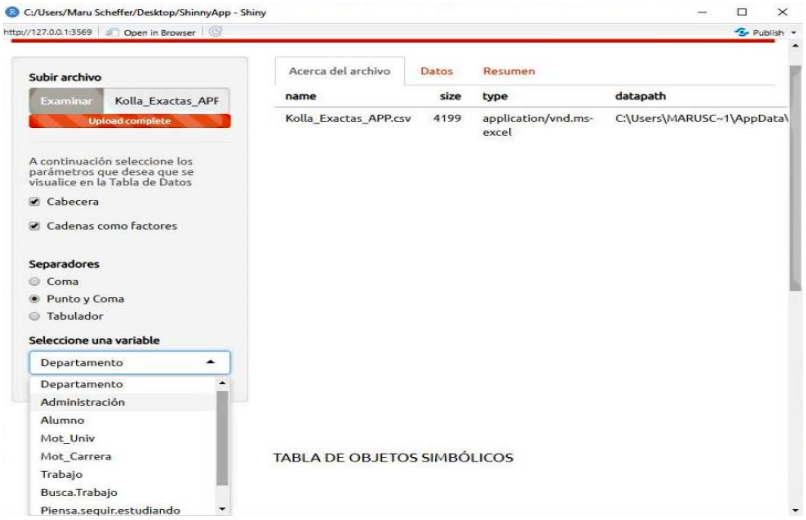


Fig. 4. Selección de la variable para realizer el análisis simbólico.

La tabla que genera en un principio APPADS, acorde al formato de la tabla creada con el paquete RSDA de R, muestra en pantalla a modo de delimitadores entre cada variable de la clase objeto, el tipo de variable bajo la etiqueta $M indicando con esto que la variable es del tipo "Modal". Seguidamente muestra el nombre de la variable, por ejemplo "alumno.modal", indicando en dicha columna la cantidad de categorías correspondientes a la variable en cuestión y en las próximas columnas los nombres de cada una de dichas categorías. Ese formato se repite por cada una de las variables del objeto simbólico presentadas en la tabla de objetos simbólicos.

Con el objetivo de brindar al usuario una visualización más amigable, se llevaron a cabo modificaciones en el formato estándar de la tabla generada, eliminando las columnas etiquetadas con $M para identificar el tipo de variable, como así también reemplazando el nombre de las variables seguidas por el tipo por el nombre de las mismas en mayúsculas con el objetivo de que sirvan a su vez como una forma de distinguir o delimitar el comienzo de las variables entre sí, por ejemplo "alumno.modal" por "ALUMNO". Esto puede observarse en la Tabla 1.

Tabla 1. Tabla de Objetos Simbólicos modificada respect al format estándar provisto por el paquete RSDA.

OBJETO	ALUMNO	Bueno	Malo	Regular	MOT_CARRERA	Avanzar	Desarrollo_Per	NC	otros	Trabalo	Util	Vocacion	TRABAJO	c/rel_prof	NC.1	No_trab	s/rel_prof	NIVEL DE EXIGEN	Adecuado	Bajo	Elev
BIOLOGÍA	3	0.95	0.05	0.00	7	0.18	0.09	0.05	0.09	0.00	0.18	0.41	4	0.64	0.05	0.18	0.14	4	0.32	0.00	0.59
CIENCIAS DE LA SALUD	3	0.91	0.00	0.09	7	0.00	0.36	0.00	0.00	0.00	0.09	0.55	4	0.82	0.00	0.09	0.09	4	0.27	0.00	0.73
GEOFÍSICA	3	1.00	0.00	0.00	7	0.25	0.50	0.00	0.00	0.00	0.00	0.25	4	1.00	0.00	0.00	0.00	4	0.00	0.00	1.00
GEOLOGÍA	3	1.00	0.00	0.00	7	0.36	0.27	0.00	0.00	0.27	0.00	0.09	4	0.55	0.00	0.27	0.18	4	0.27	0.18	0.55
INFORMÁTICA	3	1.00	0.00	0.00	7	0.15	0.15	0.00	0.00	0.31	0.15	0.23	4	0.77	0.00	0.15	0.08	4	0.31	0.00	0.62

5.3 Detalles de las salidas

En la Tabla 1, se muestra en primer lugar la tabla simbólica, en ella la 1° columna corresponde a los objetos en estudio, en este caso son los siete departamentos de la facultad. La 2° columna corresponde al número de categorías, en este caso 3, de la variable Alumnos la cual se refiere a la relación que tuvo el egresado con sus compañeros cuando fue alumno. La 3° columna indica el porcentaje de egresados que tuvieron una buena relación con sus compañeros en la facultad. La 4°columna cuando fue mala y la 5° a la categoría cuando fue regular.

La 6° columna indica el número de categorías que corresponden a la variable Motivos por los que eligieron la carrera en la que egresaron, en este caso son siete. Las siete columnas a continuación muestran los valores obtenidos en cada departamento (objeto).

La 14° columna indica el número de categorías correspondientes a la variable Trabajo para conocer cuál es la situación laboral luego de haber egresado, la cuales figuran en las cuatro columnas que siguen a continuación.

Finalmente, en la 19° columna figura la cantidad de categorías de la variable nivel de exigencia de la carrera, la cual tiene como objetivo conocer cuál es la opinión del egresado en relación a su experiencia con la carrera de la cual egresó. Las últimas cuatro columnas corresponden a las categorías de dicha variable.

De este modo la tabla representa los porcentajes obtenidos en las características de las variables Alumno (relación con sus pares), Mot_Carrera (motivo por el que eligió esa carrera), Trabajo (situación laboral actual) y Nivel de exigencia (respecto a su experiencia durante el cursado de la carrera). Las variables consideradas se escogieron de acuerdo a la importancia que tienen según los expertos y la asociación entre las mismas.

Se nota que quizás sea difícil determinar de una manera rápida el comportamiento de los egresados en cada departamento. Por ello se realiza un gráfico de radar que permite visualizar y comparar los resultados de los objetos simbólicos; el cual fue obtenido haciendo uso de paquetes provistos por R. En el gráfico se puede observar a simple vista el comportamiento de los egresados en cada departamento (Biología, Ciencias de la Salud, Geología, Geofísica e Informática). El mismo consiste en un pentágono en el cual se ve a simple vista que la relación entre alumnos es muy similar en todos los departamentos, mientras que existen diferencias en los departamentos respecto a si los egresados que egresan trabajan en relación a la carrera.

En la Figura 5 se puede observar la gráfica obtenida a partir del análisis de las

variables del objeto simbólico “Informática” y los valores correspondientes a sus categorías.

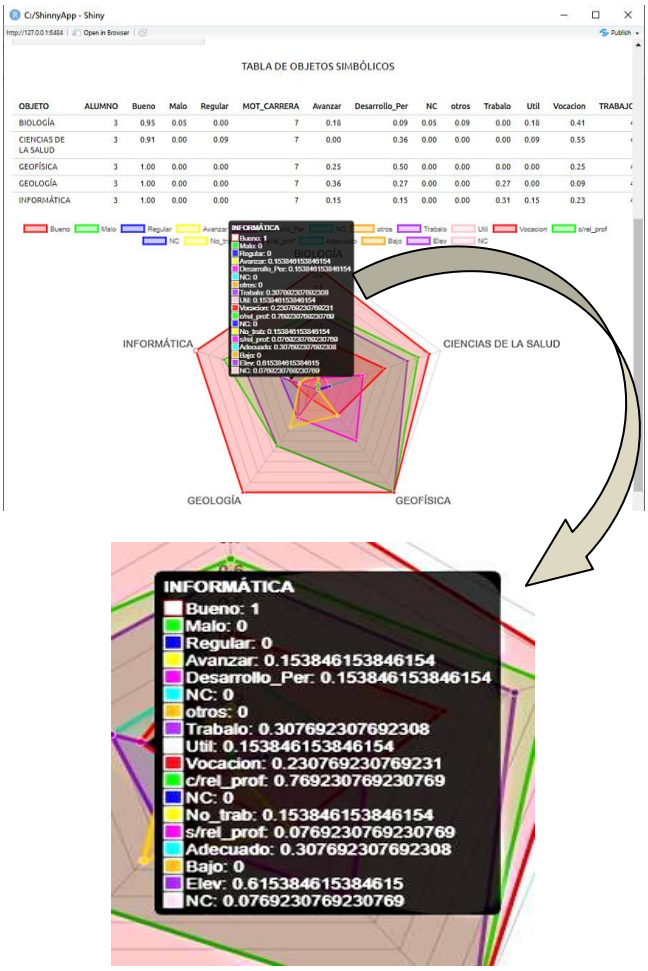


Fig.5. Valores de las categorías correspondientes a las variables del objeto simbólico “Informática”

6 Conclusiones

La ventaja de APPADS, es que la misma implica varias mejoras siendo una de ellas y como se mencionó, el hecho de poder ser implementada ya sea a través de la web o de un ordenador de manera local sin tener que pagar una licencia propietaria. Otro punto importante a destacar, es el hecho de poder trabajar directamente a partir de un archivo tan sencillo como lo es un archivo con extensión .csv, esto ahorra tiempo y simplifica la tarea al usuario.

Por otra parte, permite seleccionar a través de un menú desplegable la variable candidata para determinar el objeto simbólico de manera reactiva y dinámica. A partir de esto se encontrarán diferentes tablas de objetos simbólicos con sus respectivos gráficos con lo que a simple vista se podrá realizar una comparación entre ellas de manera aislada e individual. De este modo APPADS es una herramienta innovadora y amigable al usuario de tablas simbólicas.

Como conclusión, la herramienta propuesta en el presente trabajo, tiene como objetivo brindar una alternativa de código abierto que facilite a los expertos en el Análisis de los Objetos Simbólicos y que colaboran en la toma de decisiones, poder obtener de manera sencilla y amigable a partir de un archivo de texto plano, la generación de los objetos simbólicos, la tabla correspondiente a los mismos; como así también una representación gráfica a través de la cual puedan establecer a simple vista una relación entre estos de manera Interactiva y Dinámica haciendo uso del concepto de Programación Reactiva.

7 Agradecimientos

El presente trabajo se encuentra enmarcado en el proyecto “Aplicación de la Ciencia de los Datos para determinar la calidad de vida de la población del gran San Juan en relación con la Nación”, según resolución 0591-20-R., aprobado y financiado por la Universidad Nacional de San Juan.

8 Referencias

1. R. A.,Fisher. “The Use of Multiple Measurements in Taxonomic Problems”. Annals of Eugenics. Vol. 7, pp. 179-188 (1936). https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
2. L., Billard; E, Diday.“Symbolic Data Analysis: Conceptual Statistics and Data Mining”. Wiley. pp. 330 (2007). ISBN: 978-0-470-09016-9
3. L, Billard. “Special Issue of Statistical Analysis and Data Mining”. Wiley, on behalf of the American Statistical Association. Vol. 4. Issue 2, pp. 147-246 (2011).
4. J. Hernández-Orallo; Q. M.J., Ramírez Quintana; C., Ferri Ramírez. “Introducción a la Minería de datos”. Pearson Prentice Hall. Vol. 18, pp. 656 (2007). ISBN: 9788420540917 8420540919.
5. E. Diday. “Probabilist, possibilist and belief objects for knowledge analysis”. Annals of Operations Research. Vol. 55. Issue 2, pp. 225–276 (1995). https://doi.org/10.1007/BF02030862.
6. H.-H, Bock; E, Diday. “Analysis of Symbolic Data: Exploratory methods for extracting statistical information from complex data”. Springer-Verlag. Berlin-Heidelberg (2020) ISBN: 978-3-642-57155-8.
7. L, Billard & E, Diday. “From the Statistics of Data to the Statistics of Knowledge”. Journal of the American Statistical Association. Vol. 98. Issue 462, pp. 470-487 (2003). DOI: 10.1198/016214503000242

Development of Information Systems for Sexual and Reproductive Health: A systematic review

Leydi Erazo[1[0000-0001-5058-847X]], Erika Gallego[1[0000-0002-1019-1627]] and Helder Castrillón Cobo[1[0000-0003-1504-8544]]

[1] Corporación Universitaria Comfacauca, Popayán, Colombia.

{lerazo, erikagallego, hcastrillon}@unicomfacauca.edu.co

Abstract. In recent years, information systems have improved the quality of life of people, contributing significantly to generating changes that are reflected in their way of life. Perhaps one of the most important contributions is related to health, seeking to produce a standardization of user information. In order to identify problems and research gaps, many researchers use a variety of comprehensive methodologies, such as systematic reviews. The aim of this systematic review (SR) is to identify, analyze and evaluate research published in the context of Sexual and Reproductive Health (SRH), on guidelines, functional requirements and information necessary for the development of health information systems. A total of 36 articles were identified that report primary studies published in congresses and journals in the context of SRH between 2015 and 2019. In selecting the articles, a set of inclusion/exclusion criteria were used. One of the greatest difficulties that became evident in the SR was the lack of unification of terms associated with the SRH domain; to mitigate this situation, searches were made using terms related to SRH. New research is therefore urged that would make it easier to define and standardize the necessary information and requirements for the design of health information systems.

Keywords: Ontology, Framework, Sexual and reproductive health, Health information systems.

1 Introduction

Improving access to health services for most population groups has allowed significant changes that are reflected in the way in which people's health information is gathered, processed and stored. Despite this, in developing countries there are challenges to be met to make their health systems sustainable and ensure coverage and quality of access to health services for the entire population. Problems related to public finance, unequal distribution of living conditions, poor management of resources, among others, generate great difficulties in ensuring the sustainability of public health and the quality of health services [1]. Achieving standardization of information collected from users, ensuring it is complete and correct, represents one of the most im-

portant challenges to be faced to better understand the needs of the population and establish plans for promotion, prevention and action by the corresponding entities [2]. Involving information technologies in health issues is a task that has been carried out in recent years through the development of applications in different environments. These applications are widely used, since they improve the quality of health care and reduce the costs of care [3]. Although there are many applications related to health issues, not all are of high quality, a consequence usually of failing to employ established standards for developing and evaluating applications in this area [4].

A number of standards, methodologies, models, frameworks, etc. are found in the literature that can be used for the development of applications in different environments. These standards provide useful guidelines for generating good quality applications, but in order to develop applications for the health sector it is necessary to take into account that these standards involve the most relevant aspects of the context, to carry out an adequate capture, processing and storage of the information [5]. Hence the International Organization for Standardization (ISO) created standard ISO/IEC 62304: Medical device software - Software life cycle processes, which is a complement to the standard for medical devices ISO/IEC 13485: Medical devices - Quality management systems - Requirements for regulatory purposes; Organization HL7 [6] has further presented a set of standards adopted by different countries and organizations in order to standardize the development of systems and the standardization of information. Even so, countries such as Colombia lack their own models for standardizing information in this sector and because of local conditions, implementation of international models is difficult. Initiatives such as Decree 780 of 2016 [7] propose the definition of a common standard to achieve data interoperability. However, these solutions leave out aspects of system design and to a greater extent in a number of domains, such as Sexual and Reproductive Health (SRH).

The review aims to identify whether or not recommendations described in frameworks, methodologies, guides and standards for the design of Health Information Systems (HIS) can be used in the SRH domain and recommend both functional aspects and information that a HIS must support in this domain. It also seeks to identify whether or not these allow for collaborative work in HIS design. The rest of the article is organized as follows. Section 2 describes the process followed in carrying out the systematic review, research questions and the articles selection process. The results are summarized in Section 3. In Section 4 an analysis and discussion of the primary studies selected is carried out. Finally, conclusions drawn from the systematic review are presented in Section 5.

2 Systematic review process

This systematic review was carried out based on the guide proposed by Kitchenham in [8]. A systematic review (SR) follows a well-defined method and is a means to identify, analyze, evaluate and compare relevant research to answer a research question on a specific topic and allows to identify new research papers [9]. The main ob-

jective of this SR is to identify, analyze and compare the existence of proposed solutions for the development of HIS, especially in the SRH domain.

The SR was carried out in three phases, as shown in Figure 1. In the first phase the protocol to guide the search was defined, which includes the research questions, keywords and inclusion and exclusion criteria. In the second phase, articles were searched for each of the research questions and the most relevant selected according to the established criteria. Finally, in the third phase, categorization and processing of the studies that would be used as the primary source of information in the systematic review was done.

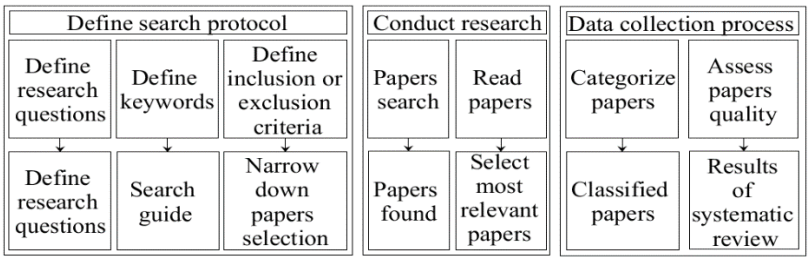


Fig. 1. Systematic review process

2.1 Define search protocol

This protocol is used to define the elements necessary to carry out the SR, establishing the criteria for the search, selection and analysis of the information. Each element of the process carried out in the SR is presented below:

Research questions
The research questions established that are expected to be resolved in carrying out the SR are:

- **RQ 1.** What are the main guidelines being used in developing Health Information Systems?
- **RQ 2.** What is the required and current information used in Sexual and Reproductive Health?
- **RQ 3.** What computer programs exist to support Sexual and Reproductive Health?

RQ1 seeks to identify the guidelines that have been used up till now in HIS development. Among these guidelines are methodologies, methods, frameworks, architectures, etc. that establish a set of steps or rules to follow in HIS development; RQ2 looks to define information necessary for operating SRH programs supported by HIS in processes of information capture and visualization; while RQ3 is focused on determining those computer programs or applications that have been designed to support the different lines of SRH.

Define keywords

Once the research questions were clearly defined, a set of keywords (Table 1) in English and Spanish associated with each of research question were identified. Similarly, bibliographic databases on which the articles were searched were identified. For each database the search string was established using the keywords.

Table 1. Keywords

RQ	English	Spanish
RQ1	Methodology, framework, guideline, model, architecture, method, ontology, standard, paradigm, pattern, structure, politic, health, healthcare, system information, software, application, app, information technology, development	Metodología, marco de trabajo, lineamiento, modelo, arquitectura, método, ontología, estándar, paradigma, patrón, estructura, política, salud, cuidado, sistemas de información, software, aplicación, app, tecnologías de la información, desarrollo
RQ2	Data, Information, record, registry, sexually transmitted infections, HIV, unintended pregnancy, safe abortion, violence related to gender and sexuality, sexual health, sexual orientation and gender identity; family planning, sexual health programs, cervical cancer, prostate cancer	Datos, información, registro, enfermedades de transmisión sexual, VIH, embarazo no deseado, aborto seguro, violencia relacionada con género y sexualidad, salud sexual, orientación sexual e identidad de género, planificación familiar, programas de educación sexual, Cáncer de cuello uterino, cáncer de próstata
RQ3	application, app web, app mobile, software, information system, web site, sexually transmitted infections, HIV, unintended pregnancy, safe abortion, violence related to gender and sexuality, sexual health, sexual orientation and gender identity; family planning, sexual health programs	aplicación, aplicación web, aplicación móvil, software, sistema de información, sitio web, infecciones de transmisión sexual, VIH, embarazo no deseado, aborto seguro, violencia relacionada con el género y la sexualidad, salud sexual, orientación sexual e identidad de género, planificación familiar, programas de salud sexual

Define inclusion or exclusion criteria

In order to select only those research papers related to the objective of this SR and that answer the research questions, a set of inclusion and exclusion criteria were defined as follows:

- The article is unrelated to the research aim: despite containing keywords in the title and abstract, the article reveals no information relevant to the research questions.
- For RQ2, only those articles that explicitly contained the keywords in the title were included, due to the ambiguity of the searched terms.
- Duplicate articles describing identical or similar results were excluded, but published elsewhere. Only the most recent or that with the most detailed explanation of the work carried out was included.
- Articles with insufficient information, that do not allow full access, that do not provide detailed information on the research carried out or that are considered incomplete were excluded.

- Published research older than 5 years was excluded, because topics related to ICT and health must be constantly updated.
- Those articles not written in English, Spanish or Portuguese were discarded.
- Articles not belonging to the ICT and Health sectors were excluded.
- A review was carried out in Google Scholar and due to the number of articles obtained, research papers lacking citations were excluded.

2.2 Conduct information search

Having established the search protocol, the literature search of the scientific research was begun.

Article search

Using the keywords, articles were searched in Scopus, IEEE Xplore, ACM Digital Library, SpringerLink, PubMed and Google Scholar. For each electronic database the search string was modified according to the required format. The articles were classified taking into account the year of publication and the research question with which they were related, the distinctions by year and research question are shown in Table 2.

Table 2. List of articles published between 2015 and 2019

RQ	2015	2016	2017	2018	2019	Total
RQ 1	792	1396	693	743	70	3694
RQ 2	321	481	257	351	109	1519
RQ 3	332	654	643	648	203	2480
				Total studies		**7693**

A total of 7,693 articles were found in all electronic databases: 3694 articles related to RQ1; 1519 related to RQ2; and 2480 to RQ3. As an exclusion criterion, a period of 5 years was established from 2015 to 2019.

Reading of articles found

The reading and selection of primary articles was carried out by applying the inclusion/exclusion criteria in three steps, as seen in Figure 2. In the first step, 7434 articles were discarded in which no relation to the objective of the investigation was revealed. Despite containing some keywords in the title and the abstract, the article does not show relevant information regarding the research questions. In the second step, 259 articles passed. Of these, the introduction, results and conclusions were read to determine if they offered significant contributions related to the SR objective. In this phase, 40 articles were selected and 219 discarded. Finally, in the third step, these 40 articles were read thoroughly and 36 selected as the primary information source. Each article was allocated to the research question to which it was linked.

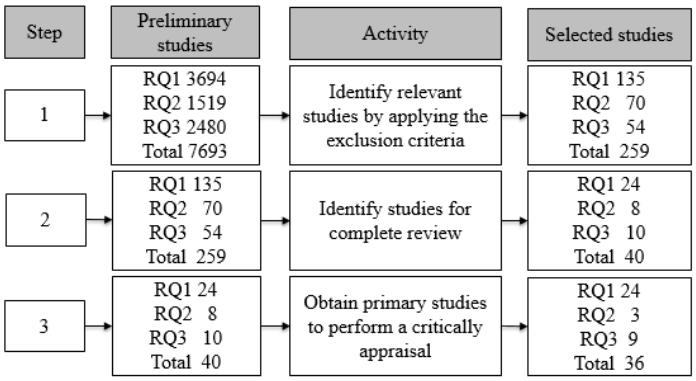


Fig. 2. Phases of the process of reading and selecting articles

3 Processing of collected data

3.1 Assessing the quality of the studies

The quality of the selected articles was established taking into account the type of article (case study, empirical studies, surveys, primary study, secondary study, etc.). Based on the reading of each article, analysis of the data allows identification of research gaps and offers suggestions for future research. All this information obtained was recorded in an Excel document containing the following fields:

- Title of article
- Authors
- University/institute,
- Abstract/important contributions
- Year of publication
- Language
- Keywords
- Bibliographic database in which it was published
- Research question with which it is associated

To assess the quality of the article, the tool developed by [10], was used, in which a set of 11 questions were considered to determine the quality of the studies. The questions seek that the studies meet three quality criteria necessary to be selected as relevant for this SR. These criteria are: rigor (completeness of the research method used in the study), credibility (the findings are well presented and significant) and relevance (the findings are useful for the industry and the research community). Data extraction and quality assessment of each study was carried out in parallel for each of the articles, by three reviewers.

4 Results

Overall, 36 research studies were identified as a source of primary information in the bibliographic databases, as shown in Figure 2. The first 24 studies were classified as

relevant to identify guidelines, frameworks, methodologies, methods, ontologies, and architectures, among others, used for the development of HIS (RQ1); 3 studies mentioned necessary and existing information for the context of SSR (RQ2); and 9 primary studies related to computer programs developed to support SRH such as HIV, safe motherhood, family planning, etc. (RQ3). The studies information is presented in Table 3.

Table 3. Primary studies selected

Ref	Title	Type solution	Health domain
[11]	Design and Development of a Sharable Clinical Decision Support System Based on a Semantic Web Service Framework	Framework	Other
[12]	OntoCR A CEN ISO-13606 clinical repository based on ontologies	Ontology	Other
[13]	Towards improving information quality requirements for online health information systems: A review on the recent frameworks	Framework	Other
[14]	Semantic interoperability and pattern classification for a service-oriented architecture in pregnancy care	Ontology	Safe motherhood
[15]	Trust Requirements in E-Health System: A Conceptual Framework	Framework	Other
[16]	The eClinical Care Pathway Framework: a novel structure for creation of online complex clinical care pathways and its application in the management of sexually transmitted infections	Framework	Other
[17]	Development of automated methods for the critical condition risk prevention, based on the analysis of the knowledge obtained from patient medical records.	Ontology	Other
[18]	Behavioral Reference Model for Pervasive Healthcare Systems	Architecture	Other
[19]	Youth Technology and HIV: Recent advances and future directions.	Guide	HIV
[20]	Prevention of sexually transmitted infections using mobile devices and ubiquitous computing.	Methodology	STIs
[21]	Design of a framework for the deployment of collaborative independent rare disease-centric registries: Gaucher disease registry model	Model	Other
[22]	Self-Management Education Through mHealth Review of Strategies and Structures	Guide	Other
[23]	A Population Health Measurement Framework: Evidence-Based Metrics for Assessing Community-Level Population Health in the Global Budget Context.	Framework	Other
[24]	A Framework for Developing Prediabetes Self-care Application	Framework	Other
[25]	An Ontology for a Patient-Centric Healthcare Interoperability Framework	Ontology - Framework	Other
[26]	Conceptual framework for the security of mobile health applications on Android platform	Framework	Other
[27]	A Security Framework for MHealth Apps on Android Platform	Framework	Other
[28]	The Evolution of a Healthcare Software Framework: Reuse, Evaluation and Lessons Learned	Framework	Other
[29]	Measuring health systems strength and its impact: Experiences from the African Health Initiative	Framework	Other
[30]	Integrative system of virtual electronic health record with online community-based health determinant data for home care service: MHealth development and usability test	Guide	Other
[31]	Ontologies for the Representation of Electronic Medical Records: The Obstetric and Neonatal Ontology	Ontology	Safe motherhood
[32]	Assessment of evaluation frameworks for design of a sexual risk prevention game for black adolescent girls	Framework	Other
[33]	A Granular Ontology Model for Maternal and Child Health Information System	Ontology	Other

[34]	FAMAP: A Framework for Developing m-Health Apps	Framework	Other
[35]	Knowledge, sources and use of family planning methods among women aged 15-49 years in Uganda: a cross-sectional study.	Guide	Family planning
[36]	Improving Linkage to HIV Care Through Mobile Phone Apps: Randomized Controlled Trial.	Guide	ITS y VIH
[37]	Sexual health in your hands How the smartphone apps can improve your sexua lwellbeing	Guide	SRH
[38]	Development of a Mobile App on Contraceptive Options for Young African American and Latina Women.	Model	SRH
[39]	Introducing the World Health Organization Postpartum Family Planning Compendium	Guide	Family planning
[40]	Collecting family planning intentions and providing reproductive health information using a tablet-based video game in India.	System of information	Family planning
[41]	Development, use, and integration of a nationally-distributed HIV/AIDS electronic health information system.	System of information	ITS y VIH
[42]	A Mobile Gaming Intervention to Increase Adherence to Antiretro-viral Treatment for Youth Living With HIV: Development Guided by the Information, Motivation, and Behavioral Skills Model	System of information	ITS y VIH
[43]	Girl Talk: A Smartphone Application to Teach Sexual Health Education to Adolescent Girls	Application mobile	SRH
[44]	Social Support in a Virtual Community: Analysis of a Clinic-Affiliated Online Support Group for Persons Living with HIV/AIDS	Application web	ITS y VIH
[45]	Development of an mHealth platform for HIV care: Gathering user perspectives through co-design workshops and interviews	Application web	ITS y VIH
[46]	Desarrollo de software educativo que contribuya a la promoción de la salud sexual y reproductiva en la institución técnica Jorge Eliécer Gaitán de González	System of information	SRH

As shown in Figure 3 for RQ1 it was to identify the guidelines used for the design of information systems in SRH. With this in mind, after performing a rigorous classification of the research studies identified, 24 solutions were obtained among architectures, frameworks, guides, methodologies, models, ontologies and other solutions.

Fig. 3. List of articles according to the type of solution and health domain

It was seen that 52% of the solutions classified correspond to framework and that 100% do not support the SRH domain. The remaining 48% corresponds to other solutions that were not developed for HIS design, but are used in this domain. Therefore, these studies are an important basis for proposing a suitable solution for the design of HIS in this domain. Regarding RQ2, it was proposed to identify the variables that a HIS in SRH should support, but no sets of variables defined for this domain were identified. However, in Table 4 some identified works that propose certain variables for some special cases are described.

Table 4. Studies that propose variables for specific topics

Paper	Contribution
[35]	Database to analyze family planning methods in women between 5 and 49 years of age. The variables included are ● Information source of family planning record ● Reliability of information source ● Current, preferred and intended use of family planning methods ● Level of knowledge about family planning methods ● Perception of what the method involves ● Which family planning method is safer
[36]	Solution for the integration of laboratory systems for HIV-associated tests • ID data of the person • Attention center • Access credentials Clinical laboratory document (does not specify model or standard)
[37]	Describes the information that the teenagers prefer to have in a SRH mobile app. ● Reminder of fertile days in women ● Ensure timely use of contraceptives ● Identify STDs ● Check pregnancy and symptoms ● Information about SRH in a creative and fun context

With question RQ3, it was proposed to identify applications used in the SRH domain, mainly those that are based on guidelines for the development of systems. Initially, 68 applications were identified. Only 9, however, met the conditions established in the SR. To make an adequate analysis of the results, classification was carried out on four aspects: the associated domain or subdomain, the type of intervention carried out from the solution, the platform on which the solution operates and the type of solution developed.

According to the results, only 22% comprehensively address SRH, the rest focus on family planning and sexually transmitted diseases. Those relating to intervention type can be divided in two: educational interventions, with 67%, and support in health aspects, with 33%. Of solutions developed for mobile devices, 78 % could be grouped into five types: expert systems, interactive web, video games, social networks and integration of information, with the most common being the development of video games, with 22%. Table 5 shows the labeling of the studies selected in the SR according to the four above-mentioned aspects. Furthermore, Figure 4 comprises a summary of the results, in percentages.

In general, it can be observed that important studies that have been carried out in this domain. Nevertheless, there are no clear criteria on how to design and develop

solutions. Neither are the functional and non-functional requirements that must be taken into account defined.

Table 5. Studies that answer and contribute to RQ3

Ref.	Domains/SubDomain	Type of intervention	Platform	Type of Solution
[38]	Family planning	Education	Mobile	Expert system
[38]	Family planning	Education	Web	Interactive web
[40]	Family planning	Education	Mobile	Videogame
[38]	Sexually Transmitted Infections	Health Support	Mobile	Information integration
[38]	Sexually Transmitted Infections	Education	Mobile	Videogame
[38]	Sexual and Reproductive Health	Education	Mobile	Videogame
[44]	Sexually Transmitted Infections	Health Support	Mobile	Social network
[45]*	Sexually Transmitted Infections	Health Support	Mobile	Expert system
[46]	Sexual and Reproductive Health	Education	Web	Social network

(*) Development carried out collaboratively

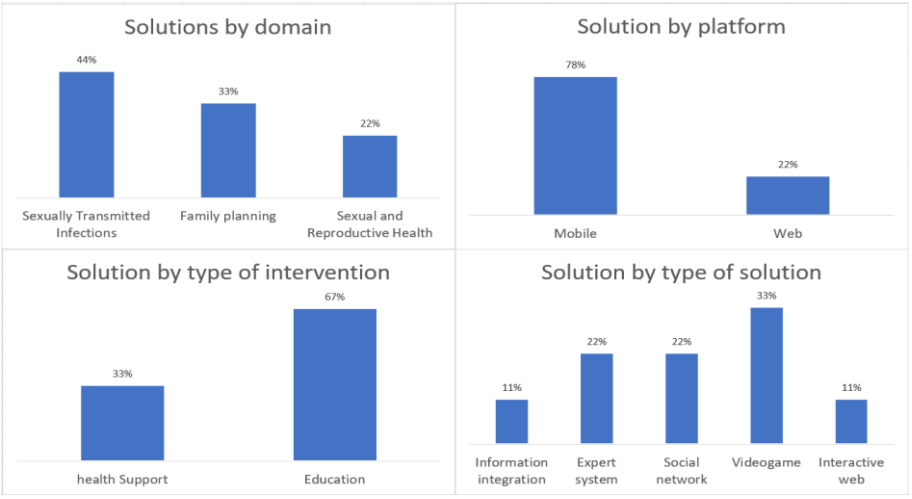


Fig. 4. List of studies according to the type of solution

5 Discussion

One of the greatest difficulties that became evident in carrying out the SR was the lack of unification of terms associated with the SRH domain. To mitigate this situation, searches were made with related terms. However, the queries on the bibliographic databases yielded many results unrelated to the aim of the SR. It can be seen that guidelines exist for the design and development of HIS according to RQ1, but that there is a sizeable gap regarding the adaptation of these guidelines (models, framework, guide, etc.) to the SRH domain.

Meanwhile, taking RQ2 into account, there is no evidence of variable sets of information required in this domain. While some have been defined in the studies reviewed, carrying out a solution based on these findings can cause problems of integra-

tion, interoperability and insufficiency of the solutions. Practically, the research works reviewed were developed for a specific problem. However, it is worth highlighting [38] that it was the study carried out collaboratively that made it possible to define a solution that meets the needs of all the participants in the design and the compatibility with other solutions, mitigating some of the problems previously described.

Finally, the solutions identified with RQ3 show a wide variety of applications that support different domains in different ways. The nine solutions prioritized describe a formal process in its development. Some features in the design and development of a system can be considered static as its quality, reliability etc. Some depend on the organizations or on the client of the solution, but there are other characteristics that are specific to the domain and cannot be easily specified. Characteristics such as the sufficiency of information and the adaptation of functionality are specific to the domain and cannot be taken from other domains.

6 Conclusions

This study was carried out to obtain an overview of the existing research for the development of HIS in the area of sexual and reproductive health; 36 articles published between 2015 and 2019 were identified. The panorama of work carried out, contributions and challenges were described and reveal needs on which to focus research efforts. The greatest amount of work carried out is related to the treatment, education and clinical support for HIV cases. However, this problem does not represent the entire domain of SRH. According to the interpretation of RQ1 there is no adequate guideline for the design and development of information systems in this domain; in RQ2 there is no formal set of variables required for systems in this same domain; and according to RQ3 there are important solutions to support specific problems that are in the SRH domain, but they are not based on standard models that ensure an adequate level of support for this domain.

In conclusion, the solutions identified are not sufficient for the development of information systems in the SRH domain. However, they do form the basis for supporting an appropriate solution. The main contributions of this document relate to identification of research studies conducted in the SRH context; application or adaptation of relevant information, to be applied in research work in Colombia; and research in which information systems were developed to support issues related to SRH. Among the challenges identified, it was found that: (i) the standardization of information on health-related issues is important; (ii) cover of other components of SRH, including such as gender violence, family planning, and voluntary termination of pregnancy; and (iii) definition of functional and non-functional requirements necessary for the development of HIS. This work was used as the basis for the creation of a domain ontology [47] and a framework for the development of Information Systems as a support tool for SRH programs[48].

12

References

1. Gutiérrez, C.: El sistema de salud colombiano en las próximas décadas: cómo avanzar hacia la sostenibilidad y la calidad en la atención. (2018)
2. Mukasa, E., Kimaro, H., Kiwanuka, A., Igira, F.: Challenges and strategies for standardizing information systems for integrated tb/hiv services in Tanzania: A case study of Kinondoni municipality. Electron. J. Inf. Syst. Dev. Ctries. 79, 1–11 (2017). https://doi.org/10.1002/j.1681-4835.2017.tb00581.x
3. Lee Ventola, C.: Social media and health care professionals: Benefits, risks, and best practices. P T. 39, 491–500 (2014)
4. Van Velthoven, M.H., Smith, J., Wells, G., Brindley, D.: Digital health app development standards: A systematic review protocol. BMJ Open. 8, 1–5 (2018). https://doi.org/10.1136/bmjopen-2018-022969
5. Yesid, H., Cobo, C., Giovanni, N., Ippolito, A.: Characterization of information systems for primary health care. Rev. Ing. 18, 141–164 (2019)
6. HL7 International: Introduction to HL7 Standards, http://www.hl7.org/implement/standards/index.cfm?ref=nav
7. Ministerio de Salud y Protección Social: Decreto número 1990 de 2016. 5 (2016)
8. Kitchenham, B.: Procedures for Performing Systematic Review. Ann. Saudi Med. 37, 79–83 (2017). https://doi.org/10.5144/0256-4947.2017.79
9. Kitchenham, B., Charters, S.: Guidelines for performing Systematic Literature Reviews in SE. 1–44 (2007). https://doi.org/10.1145/1134285.1134500
10. Dybå, T., Dingsøyr, T.: Empirical studies of agile software development: A systematic review. Inf. Softw. Technol. 50, 833–859 (2008). https://doi.org/10.1016/j.infsof.2008.01.006
11. Zhang, Y.F., Gou, L., Tian, Y., Li, T.C., Zhang, M., Li, J.S.: Design and Development of a Sharable Clinical Decision Support System Based on a Semantic Web Service Framework. J. Med. Syst. 40, (2016). https://doi.org/10.1007/s10916-016-0472-y
12. Lozano-Rubí, R., Muñoz Carrero, A., Serrano Balazote, P., Pastor, X.: OntoCR: A CEN/ISO-13606 clinical repository based on ontologies. J. Biomed. Inform. 60, 224–233 (2016). https://doi.org/10.1016/j.jbi.2016.02.007
13. Ighe, M.A., Mohammed, S.A.B.: Towards improving information quality requirements for online health information systems: A review on the recent frameworks. Proc. - Int. Conf. Inf. Commun. Technol. Muslim World 2018, ICT4M 2018. 138–143 (2018). https://doi.org/10.1109/ICT4M.2018.00034
14. Moreira, M.W.L., Rodrigues, J.J.P.C., Sangaiah, A.K., Al-Muhtadi, J., Korotaev, V.: Semantic interoperability and pattern classification for a service-oriented architecture in pregnancy care. Futur. Gener. Comput. Syst. 89, 137–147 (2018). https://doi.org/10.1016/j.future.2018.04.031
15. Kamalrudin, M., Winarsih, N.A.S., Sidek, S.: Trust requirements in e-health system: A conceptual framework. J. Telecommun. Electron. Comput. Eng. 10, 29–33 (2018)
16. Gibbs, J., Sutcliffe, L.J., Gkatzidou, V., Hone, K., Ashcroft, R.E., Harding-Esch, E.M., Lowndes, C.M., Sadiq, S.T., Sonnenberg, P., Estcourt, C.S.: The eClinical Care Pathway Framework: A novel structure for creation of online complex clinical care pathways and its application in the management of sexually transmitted infections.

BMC Med. Inform. Decis. Mak. 16, 1–9 (2016). https://doi.org/10.1186/s12911-016-0338-8

17. Naydanov, C., Palchunov, D., Sazonova, P.: Development of automated methods for the critical condition risk prevention, based on the analysis of the knowledge obtained from patient medical records. Proc. - 2015 Int. Conf. Biomed. Eng. Comput. Technol. Sib. 2015. 33–38 (2015). https://doi.org/10.1109/SIBIRCON.2015.7361845
18. Tahmasbi, A., Adabi, S., Rezaee, A.: Behavioral Reference Model for Pervasive Healthcare Systems. J. Med. Syst. 40, 1–23 (2016). https://doi.org/10.1007/s10916-016-0632-0
19. Hightow-weidman, L.B., Building, B., Hill, C., Muessig, K.E., Hill, C., Carolina, N., Bauermeister, J., Arbor, A., Zhang, C., Hill, C., Carolina, N., Legrand, S.: Youth, Technology and HIV: Recent Advances and Future Directions. 12, 500–515 (2016). https://doi.org/10.1007/s11904-015-0280-x.Youth
20. Besoain, F., Perez-Navarro, A., Caylà, J.A., Aviñó, C.J., de Olalla, P.G.: Prevention of sexually transmitted infections using mobile devices and ubiquitous computing. Int. J. Health Geogr. 14, (2015). https://doi.org/10.1186/s12942-015-0010-z
21. Bellgard, M.I., Napier, K.R., Bittles, A.H., Szer, J., Fletcher, S., Zeps, N., Hunter, A.A., Goldblatt, J.: Design of a framework for the deployment of collaborative independent rare disease-centric registries: Gaucher disease registry model. Blood Cells, Mol. Dis. 68, 232–238 (2018). https://doi.org/10.1016/j.bcmd.2017.01.013
22. Bashi, N., Fatehi, F., Fallah, M., Walters, D., Karunanithi, M.: Self-management education through mhealth: Review of strategies and structures. JMIR mHealth uHealth. 6, (2018). https://doi.org/10.2196/10771
23. Hatef, E., Lasser, E.C., Kharrazi, H.H.K., Perman, C., Montgomery, R., Weiner, J.P.: A population health measurement framework: Evidence-based metrics for assessing community-level population health in the global budget context. Popul. Health Manag. 21, 261–270 (2018). https://doi.org/10.1089/pop.2017.0112
24. Subramaniam, S., Dhillon, J.S., Ahmad, M.S., Teoh, C., Leong, J.W.S.: A framework for developing prediabetes self-care application. Int. Conf. Res. Innov. Inf. Syst. ICRIIS. (2017). https://doi.org/10.1109/ICRIIS.2017.8002463
25. Azarm, M., Ave, K.E., Peyton, L., Ave, K.E.: An Ontology for a Patient-Centric Healthcare Interoperability Framework. 2018 IEEE/ACM Int. Work. Softw. Eng. Healthc. Syst. 34–41 (2018)
26. Hussain, M., Zaidan, A.A., Zidan, B.B., Iqbal, S., Ahmed, M.M., Albahri, O.S., Albahri, A.S.: Conceptual framework for the security of mobile health applications on Android platform. Telemat. Informatics. 35, 1335–1354 (2018). https://doi.org/10.1016/j.tele.2018.03.005
27. Hussain, M., Al-Haiqi, A., Zaidan, A.A., Zaidan, B.B., Kiah, M., Iqbal, S., Iqbal, S., Abdulnabi, M.: A security framework for mHealth apps on Android platform. Comput. Secur. 75, 191–217 (2018). https://doi.org/10.1016/j.cose.2018.02.003
28. Macedo, A.A., Baranauskas, J.A., Bulcão-Neto, R.F.: The evolution of a healthcare software framework: Reuse, evaluation and lessons learned. Proc. 2018 Fed. Conf. Comput. Sci. Inf. Syst. FedCSIS 2018. 15, 1043–1051 (2018). https://doi.org/10.15439/2018F173
29. Sherr, K., Fernandes, Q., Kanté, A.M., Bawah, A., Condo, J., Mutale, W.: Measuring

health systems strength and its impact: Experiences from the African Health Initiative. BMC Health Serv. Res. 17, (2017). https://doi.org/10.1186/s12913-017-2658-5

30. Boonchieng, W., Boonchieng, E., Tuanrat, W., Khuntichot, C., Duangchaemkarn, K.: Integrative system of virtual electronic health record with online community-based health determinant data for home care service: MHealth development and usability test. 2017 IEEE Healthc. Innov. Point Care Technol. HI-POCT 2017. 2017-Decem, 5–8 (2017). https://doi.org/10.1109/HIC.2017.8227571
31. Mauricio Barcellos Almeida, F.F.: Ontologies for Representation of Electronic Medical Records: the Obstetric and Neonatal Ontology. J. Am. Ceram. Soc. (2008)
32. Sockolow, P., Joppa, M., Zhu, J.: Assessment of evaluation frameworks for design of a sexual risk prevention game for black adolescent girls. Stud. Health Technol. Inform. 250, 101–105 (2018). https://doi.org/10.3233/978-1-61499-872-3-101
33. Ismail, S., Alshmari, M., Latif, K., Ahmad, H.F.: A Granular Ontology Model for Maternal and Child Health Information System. J. Healthc. Eng. 2017, (2017). https://doi.org/10.1155/2017/9519321
34. Iván Gracia, Manuel Gozalez, G.P.: FAMAP: A Framework for Developing m-Health Apps. 10th Eur. Conf. Inf. Syst. Manag. Acad. Conf. Publ. Ltd. 1, 20–28 (2016). https://doi.org/10.1007/978-3-319-77703-0
35. Alege, S.G., Matovu, J.K.B., Ssensalire, S., Nabiwemba, E.: Knowledge, sources and use of family planning methods among women aged 15-49 years in Uganda: A cross-sectional study. Pan Afr. Med. J. 24, 1–12 (2016). https://doi.org/10.11604/pamj.2016.24.39.5836
36. Venter, W., Coleman, J., Chan, V.L., Shubber, Z., Phatsoane, M., Gorgens, M., Stewart-Isherwood, L., Carmona, S., Fraser-Hurt, N.: Improving linkage to HIV care through mobile phone apps: Randomized controlled trial. JMIR mHealth uHealth. 6, (2018). https://doi.org/10.2196/mhealth.8376
37. Eleuteri, S., Rossi, R., Tripodi, F., Fabrizi, A., Simonelli, C.: Sexual health in your hands: How the smartphone apps can improve your sexual wellbeing? Sexologies. 27, e57–e60 (2018). https://doi.org/10.1016/j.sexol.2018.04.004
38. Akinola, M., Hebert, L.E., Hill, B.J., Quinn, M., Holl, J.L., Whitaker, A.K., Gilliam, M.L.: Development of a Mobile App on Contraceptive Options for Young African American and Latina Women. Heal. Educ. Behav. 46, 89–96 (2019). https://doi.org/10.1177/1090198118775476
39. Sonalkar, S., Gaffield, M.E.: Introducing the world health organization postpartum family planning compendium. Int. J. Gynecol. Obstet. 136, 2–5 (2016). https://doi.org/10.1002/ijgo.12003
40. Bertozzi, E., Bertozzi-Villa, A., Kulkarni, P., Sridhar, A.: Collecting family planning intentions and providing reproductive health information using a tablet-based video game in India. Gates Open Res. 2, 20 (2018). https://doi.org/10.12688/gatesopenres.12818.2
41. Milberg, J.A.: Development, use, and integration of a nationally-distributed HIV/AIDS electronic health information system. J. Am. Med. Informatics Assoc. 23, 1190–1194 (2016). https://doi.org/10.1093/jamia/ocv212
42. Whiteley, L., Brown, L., Lally, M., Heck, N., van den Berg, J.J.: A mobile gaming intervention to increase adherence to antiretroviral treatment for youth living with hiv:

Development guided by the information, motivation, and behavioral skills model. JMIR mHealth uHealth. 6, (2018). https://doi.org/10.2196/mhealth.8155

43. Brayboy, L.M., Schultz, L., Mills, B.S.L., Sepolen, A., Mezoian, T., Wheeler, C., Clark, M.A.: Girl Talk: A Smartphone Application to Teach Sexual Health Education to Adolescent Girls. 30, 23–28 (2017). https://doi.org/10.1016/j.jpag.2016.06.011.Girl
44. Flickinger, T.E., DeBolt, C., Waldman, A.L., Reynolds, G., Cohn, W.F., Beach, M.C., Ingersoll, K., Dillingham, R.: Social Support in a Virtual Community: Analysis of a Clinic-Affiliated Online Support Group for Persons Living with HIV/AIDS. AIDS Behav. 21, 3087–3099 (2017). https://doi.org/10.1007/s10461-016-1587-3
45. Marent, B., Henwood, F., Darking, M., Consortium, E.: Development of an mHealth platform for HIV care: Gathering user perspectives through co-design workshops and interviews. JMIR mHealth uHealth. 6, (2018). https://doi.org/10.2196/mhealth.9856
46. Ocaña, U.F. de P.S.-: Desarrollo de un Software Educativo que Contribuya a la Promoción de la Salud Sexual y Reproductiva. Tesis. 1–49 (2017)
47. Gallego, E., Castrillón, H., Erazo, L., Obando, F.: OntoSSR : Una solución para diseño estándar de sistemas de información en el dominio de salud sexual y reproductiva. Rev. Ibérica Sist. e Tecnol. Informação. (2019)
48. Castrillón-Cobo, H.Y., Erazo-Paruma, L.R., Gallego-Gómez, E.L.: Conceptual Framework for the Collaborative Design of Health Information Systems. Rev. Fac. Ing. 29, e11757 (2020). https://doi.org/10.19053/01211129.v29.n54.2020.11757

SISTEMA DE INFORMACIÓN PARA LA GESTIÓN DE DATOS DEL LABORATORIO DE ACUAPONÍA LESTOMA

Andrés Camilo Tinoco Barreto [0000-0002-3141-9854],
Cristian Camilo Medina Albarracín [0000-0002-4625-9013],
Alexander Espinosa García [0000-0002-6571-7294],
Jaime Eduardo Andrade Ramirez [0000-0001-6077-2449]

Universidad de Cundinamarca, Facatativá 253052, Colombia
atinoco@ucundinamarca.edu.co
ccamilomedina@ucundinamarca.edu.co
aespinosa1@ucundinamarca.edu.co
jeandrade@ucundinamarca.edu.co

***Resumen*—** Los sistemas de producción acuapónicos son sistemas que juntan la acuicultura y la producción hidropónica de las plantas. En Colombia, los recursos hídricos y la amplia variedad de especies acuáticas y vegetales permiten la implementación de los sistemas acuapónicos, por ende, resulta indispensable y necesario fomentar la realización de proyectos que vinculen la acuaponía. Los sistemas de información permiten administrar, procesar y distribuir datos relevantes para los procesos fundamentales que se producen dentro de los sistemas acuapónicos con el fin de obtener resultados y poder visualizar la información generada. Este proyecto se realiza con el propósito de implementar un sistema de información para administrar los datos generados por el laboratorio experimental de sistemas tecnológicos orientados a modelos acuapónicos (LESTOMA) ubicado en la Universidad de Cundinamarca, extensión Facatativá. El proyecto se realizará bajo los parámetros de la metodología de investigación cuantitativa, la cual permite reunir y administrar información de un número relativamente grande de las plantas implementadas en el laboratorio.

Palabras claves - Transmisión de datos, monitoreo de sistemas, acuaponía, sistemas de información, gestión de datos.

1. Introducción

La acuaponía es una combinación de peces y plantas en un sistema de producción. Hay diferentes técnicas de cultivos, como las aguas profundas, la recirculación del agua y sus componentes, pero su objetivo está en preparar el nutriente en el agua, para que la planta crezca teniendo en cuenta las variables medioambientales como la salinidad, la temperatura, la intensidad de luz y la recirculación de elementos, que hagan de esta combinación un aspecto vital, para el ahorro y la recuperación de recursos.[1]

La acuaponía como los distintos procesos organizacionales, requiere de un sistema capaz de gestionar los datos resultantes de las actividades. El propósito de un sistema de información está basado en el conjunto de datos que interactúan entre sí. Con el objetivo de gestionar información que permite ayudar a administrar, recolectar, recuperar, procesar, visualizar, almacenar y distribuir información relevante para los procesos y actividades.[2]

La importancia del sistema de información radica en la eficiencia en la correlación de los datos capturados a través de los procesos ya diseñados para cada área con el objetivo de producir información. los procesos ya mencionados anteriormente se basan en un contexto plasmados en la Fig. 1.

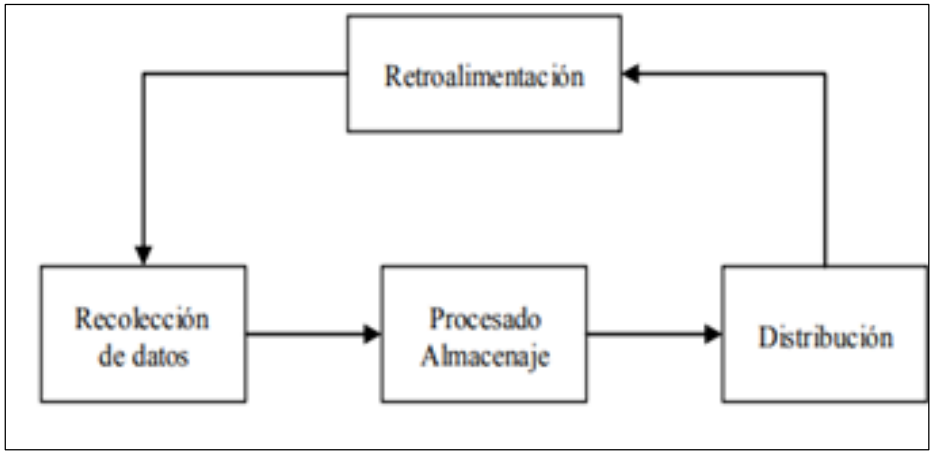


Fig 1. Modelo general de un sistema de información. Fuentes: [2].

Un sistema de información, controla sus variables por medio de los actuadores y sensores conectados por red inalámbrica, a partir de una plataforma web o móvil; el internet de las cosas unifica los datos enviados para que el usuario acceda a ellos y monitoree su sistema; allí nace la unión de una infraestructura de una red en sistemas informáticos con virtualización de datos, disponibles en la internet generando valor sobre los usuarios.[3]

El laboratorio de acuaponía LESTOMA (laboratorio experimental de sistemas tecnológicos orientados a modelos acuapónicos) está orientado a el proceso de diseño, construcción e implementación de una estrategia de incremento de la productividad de las unidades de producción agrícolas (UPAs) de la región sabana de occidente a partir de una alternativa tecnológica sostenible basada en un prototipo de sistema acuapónico automatizado. Pero este laboratorio no dispone de un mecanismo de recolección y centralización de información, que permita a los investigadores y usuarios aprovechar y almacenar los datos que llegan. Este proyecto propone un sistema de información que centraliza, almacena y gestión toda la información para su presentación visual y análisis posterior de los datos a través de módulos, generando así una estructura capaz de administrar la información.

2. Estructura de un sistema acuapónico

Los sistemas acuapónicos dan la opción de producir su propio alimento orgánico de peces y plantas en área urbana o rural. Este proyecto se basó en el diseño en implementación de un sistema de información acuapónico para el control de variables.[4]

Los sistemas acuapónicos tradicionales constan principalmente de cinco componentes, los cuales son tanque de cultivo de plantas, tanque de cultivo de peces, filtro mecánico, filtro biológico y sumidero.[5] Este sistema acuapónico se basa en gran parte en el flujo de agua que recircula por medio de la gravedad como se ilustra en la Figura 2.

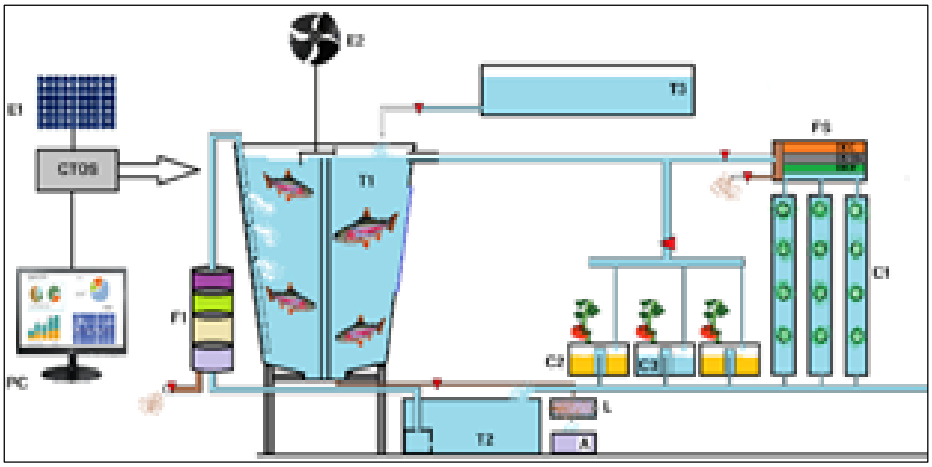


Fig 2. Diseño acuapónico implementado. Fuente: Autores.

Estos componentes se encuentran conectados por medio de un sistema de circulación de agua en el cual se benefician mutuamente peces, plantas y bacterias. Los compuestos generados por la descomposición del alimento no consumido por los peces y las heces de los mismos, se encuentran disueltos en el agua que circula inicialmente del tanque de peces hacia el filtro mecánico para la eliminación de las partículas de mayor tamaño, seguidamente el agua circula al filtro biológico que se encarga de la oxidación del amonio a nitritos y posteriormente a nitratos por medio de bacterias del tipo nitrosomonas y nitrobacterias que se encuentran adheridas al filtro biológico.

Se contempla que, en la estructura de un sistema acuapónico tradicional, la perdida de información generada se puede dar en grandes cantidades, porque cuya gestión se basa solo en la producción y no en el análisis productivo a gran escala. La información es un eje fundamental en el seguimiento y análisis de los comportamientos alrededor del ciclo de producción, además, la gestión de información generada por los dispositivos, recursos y demás procesos puede ser de utilidad para la toma de decisiones, cuando de esta se requiera.[3]

A continuación, se dará a conocer los procesos estructurados con los que se están llevando a cabo el desarrollo del sistema de información, los parámetros que se están tomando en cuenta para el proyecto están constituidos dentro del centro de innovación y tecnología (CIT), se utilizan estos como objetivo de diseñar un sistema capaz de gestionar los datos resultantes del laboratorio.

3. Procedimiento de diseño

Los procesos estructurados con los que se están llevando a cabo el desarrollo del sistema de información están basados en una serie de etapas constituidas para el correcto funcionamiento y ciclo de vida de un sistema de información, iniciando desde un proceso de planeación estratégica, pasando a un análisis, hasta su posterior control y mantenimiento requerido.[7]

El proceso de desarrollo del sistema de información constaría de siete etapas fundamentales:

a) *Definición*: En esta etapa se determinaría si se presenta problemas y como esto pueden solucionarse mediante la implantación de un sistema de información. En ella se identificarán cuáles son los objetivos del uso del sistema de información y como estos se ubican dentro de la estrategia del proyecto.

b) *Análisis de sistemas*: Tras haber identificados los diferentes problemas de la organización estos serán analizados más detenidamente, identificando las causas que lo originan y planteando diversas soluciones. En esta fase se producirá un estudio de factibilidad, para ver si las soluciones son posibles dados los recursos que se posee.

c) *Diseño de Sistemas*: Una vez elegida aquella solución que resuelva los problemas, se detallará cómo el sistema de información satisface los requisitos planteados por la organización. A la hora de diseñar los sistemas, hemos de indicar que componentes de los sistemas de información utilizaremos (nivel hardware, software y tecnología de las telecomunicaciones) y como se relacionarán dichos componentes entre sí.

d) *Programación*: Se traducirán las especificaciones del sistema desarrolladas en la etapa anterior, llevándose a cabo la programación y el desarrollo del software.

e) *Fase de pruebas:* Para evaluar el correcto funcionamiento del sistema de información será necesario lleva a cabo un proceso exhaustivo y profundo para determinar si el sistema de información funciona en diversas

condiciones y si los resultados se corresponden con lo que se esperaba.

f) *Conversión:* Una vez comprobado que el sistema de información funciona correctamente se llevará a cabo la implantación de este.

g) *Producción y mantenimiento:* Una vez instalado el nuevo sistema de información se dice que el sistema está en producción. A partir de aquí existir un proceso constante de evaluación del sistema de información por parte de los usuarios y personal especializado.

Las etapas mencionadas son las más utilizadas por su estructura definida y concreta, capaz de generar una organización clara a la planeación del sistema de información, delimitando obstáculos o problemas que puedan surgir.[8] Es claro que el sistema de información no es la única forma de realizar la gestión de datos, pero si hay que aclarar que es el más eficiente, en la figura 3 se observa la comparación de un sistema de información con otros tipos de estructuras de gestión.

Enfoque	***Características***	***Ventajas***
Ciclo de vida de Sistemas	• Secuencial • Realización de un proceso formal • Especificaciones y aprobaciones por escrito • Los usuarios tienen un papel limitado	• Necesario para Sistemas y proyectos muy complejos y grandes
Elaboración de prototipos	• Requerimientos especificados dinámicamente con Sistema experimental • Proceso rápido, informal e iterativo • Los usuarios interactuan rápido con el proceso	• Rápido y barato • Util cuando existe incertidumbre en los requisitos de información o los usuarios finales son importantes
Paquete de software para la aplicación	• El software comercial evita necesidad de programas desarrollados internamente	• Se reducen el diseño, programación, instalación y mantenimiento • Ahorro en tiempo y coste • Disminuye la necesidad de poseer recursos internos
Desarrollo de usuarios finales	• Sistemas creado por y para usuarios finales • Rápido e informal • Poca influencia especialistas de la Información	• Usuarios controlan la construcción de los Sistemas • Ahorra el coste y tiempo de desarrollo

Fig 3. Comparativa de sistemas. Fuentes: [7].

La totalidad de las fases de desarrollo de un sistema de información constituirían el denominado ciclo de vida del sistema implementado.

4. Variables importantes del sistema

El laboratorio requiere de las variables resultantes del proceso productivo de las plantas, peces, circulación del agua y comida, esto implica factores como el monóxido de carbono, temperatura del ambiente, tamaño, cantidad, entre otras. Además el crecimiento, altura, tiempo, riego y demás componentes necesarios para la administración de los datos con respecto al crecimiento de las plantas, sin embargo se contemplan también los posteriores requerimientos que pueda solicitar el sistema, debido a que la toma de datos siempre está en expansión, en donde se podrían incluir nuevos aspectos que causen una generación nueva de información, que también deberá ser almacenada por el sistema, sin importar que nuevos recursos se implementen. El sistema de hardware instalado en el laboratorio hará la transmisión de dos tipos de variables, las analógicas y las digitales, a partir de esto el proyecto gestionará las variables permitiendo el ingreso de nuevos datos sin realizar cambios en el código fuente del sistema.

Tabla 1. Posibles variables del sistema de información. Fuente: Autores.

Categoría	**Posibles variables**
Peces	Tamaño Peso Cantidad Especie Crecimiento
Alimentos	Cantidad de alimento Hora
Circulación del agua	Perdida de agua Reutilización de agua Temperatura Cantidad
Plantas	Altura (Crecimiento) Tiempo (vida de la planta) Riego Crecimiento

5. REQUERIMIENTOS DEL SISTEMA

- El número de variables del sistema.
- El sistema debe almacenar toda la información resultante del proceso de producción, según especificaciones.
- El sistema deberá incorporar nuevas variables y dispositivos a través del tiempo.
- El sistema debe mostrar la visualización de la información.
- El sistema debe brindar un control específico a un ente, para que este

vigile y administre la gestión de los demás participantes o usuarios.

- El sistema debe ser capaz de incorporar cambios a posteriores, ya que el sistema puede generar expansión y estos deben ser incluidos en el proceso de administración y gestión del sistema.

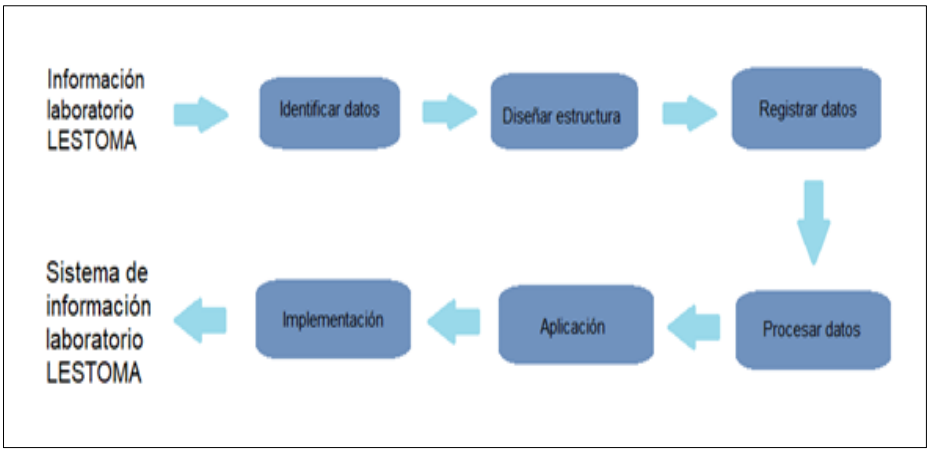


Fig 4. Diagrama de bloques para administrar la información. Fuente: Autores.

6. Conclusión

Existen pocos estudios sobre automatización y controles de almacenamiento de datos en sistemas acuapónicos, generalmente es un sistema desarrollado de forma tradicional donde la mayoría de actividades de control y seguimiento de la producción se realizan de forma manual. Los parámetros son medidos y analizados en laboratorios repercutiendo en la imposibilidad de tomar decisiones en tiempo real del comportamiento del sistema acuapónico, sin olvidar que no se realiza un almacenamiento correcto de toda la información resultante de este.

Es claro que el sistema de información es la estructura más completa en la que se puede desarrollar una verdadera gestión y administración de la información, dando una organización de datos, un control y una facilidad en la toma de decisiones.

El sistema permitirá la gestión de toda la información que genere el laboratorio acuapónico para mejorar su utilidad y facilitar aspectos como el análisis y optimización de los recursos implementados en el laboratorio.

Referencias

1. M. A. C. Gomez, “Monitoreo acuapónico y tecnología IOT en el CBA,” *Rev. Siembra CBA*, no. 1, pp. 83–88, 2019.
2. R. Hernández, C. Fernández, and P. Bapista, “Cómo se originan las investigaciones cuantitativas, cualitativas o mixtas,” *Metodol. la Investig.*, pp. 24–29, 2014.

8

3. J. M. D. D. E. P. ORTIZ, “DESARROLLO DE UN SISTEMA DE CONTROL PARA EL RIEGO Y LA ILUMINACIÓN EN UNA GRANJA VERTICAL,” *l*, vol. ى1, no. ى, p. ى, 2017.
4. W. P. and K. M. B. V. J. Caiozzo, F. Haddad, S. Lee, M. Baker, “DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA PARA EL CONTROL Y MONITOREO DE VARIABLES APLICADO A UN CULTIVO ACUAPÓNICO A PEQUEÑA ESCALA BASADO EN IOT,” pp. 1–19, 2019.
5. K. Krishnasamy, J. Nair, and B. Bäuml, “Hydroponic system for the treatment of anaerobic liquid,” *Water Sci. Technol.*, vol. 65, no. 7, pp. 1164–1171, 2012, doi: 10.2166/wst.2012.031.
6. Luis Felipe Hernández Zambrano, “Diseño, construcción y evaluación de un sistema acuapónico automatizado de tipo tradicional y doble recirculación en el cultivo de Tilapia Roja (Oreochromis Mossambicus) y Lechuga Crespa (Lactuca Sativa),” p. 127, 2017.
7. A. Hernández Trasobares, “Los sistemas de información: evolución y desarrollo,” *Proy. Soc. Rev. Relac. laborales*, no. 10, pp. 149–165, 2013.
8. E. Upc, *Vicenç Fernández Alarcón Desarrollo de sistemas de información*.

Computational strategies for implementation of 2D elastic wave modeling in GPU

Anderson Paez Chanaga[1], Ivan Sanchez Galvis[2], and Ana Ramirez Silva[1]

[1] CPS Group, Universidad Industrial de Santander, Santander, Colombia
[2] CEMOS Group, Universidad Industrial de Santander, Santander, Colombia

Abstract. Elastic wave modeling presents a challenge to implement since it is a computational costly procedure. Nowadays, due to GPU increased power jointly with development in HPC computation, it is possible to execute elastic modeling with better execution times and memory use. This study evaluates the performance of 2 strategies for implementing elastic modeling using different kernel launching layouts, CPML memory allocation strategies and wavefield storage management. The performance measures show that the algorithm which includes 2D kernel launching layout, CPML reduced memory strategy and GPU global memory storage to save wavefield cube peaks up to 88.4% better execution time and uses 13.3 times less memory to obtain same elastic modeling results. There is also an increasing trend of enhancement in execution times and memory savings when working with models of bigger sizes with this strategy.

Keywords: GPU · HPC · CUDA · Elastic Wave Modeling · CPML

1 Introduction

Seismic imaging as part of exploration seismology is focused on building physical properties images to gain better insights of earth subsurface; it is founded on how seismic waves can collect information of those properties while traveling into determined medium and this feature is greatly exploited in several imaging reconstruction methods [1, 13] and its applications [11]. Due to the aforementioned, simulation of wave propagation has become a necessity, however, implementation of numerical modeling is widely known to be a costly computational method mainly because the amount of information to process; furthermore, this cost is increased if complex subsurface models are considered as it is the case of elastic model.

Owing to high computing resources consumption, implementation of numerical elastic modeling is still a challenge to deal in terms of computational performance, nevertheless, development in graphical processing units (GPU) [5] and emergence of new programming paradigms as heterogeneous computing and High Performance Computing (HPC) [2, 4] have let to tackle this problem by reducing time elapse, memory consumption and other metrics; even though the necessity of improving performance is present.

Based on the previously mentioned, finding new strategies to implement numerical elastic modeling which explore new programming techniques, technological developments and hardware equipment is highly appreciated. This study presents a comparison of two elastic wave modeling algorithms running in GPU which exploit three different features of hardware programming such as kernel launching layout, CPML memory allocation management and CPU-GPU memory transference; at the end an evaluation between both algorithms is performed using metrics as execution time and memory consumption to determine which one has better performance.

2 Methods

2.1 Elastic wave modeling

Numerical wave modeling is considered a powerful tool for simulating how seismic waves travel over different subsurface layers in earth. Elastic wave modeling considers that the earth could be modeled by three parameters named λ, μ and ρ; those parameters are related each other by elastic wave propagation set of equations presented by Virieux in the P-Sv form [12] in equations 1, 2, 3, 4, 5.

$$\frac{\partial v_x}{\partial t} = \frac{1}{\rho}\left(\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xz}}{\partial z}\right), \tag{1}$$

$$\frac{\partial v_z}{\partial t} = \frac{1}{\rho}\left(\frac{\partial \sigma_{xz}}{\partial x} + \frac{\partial \sigma_{zz}}{\partial z}\right), \tag{2}$$

$$\frac{\partial \sigma_{xx}}{\partial t} = (\lambda + 2\mu)\frac{\partial v_x}{\partial x} + \lambda\frac{\partial v_z}{\partial z} + \varphi_{xx}, \tag{3}$$

$$\frac{\partial \sigma_{zz}}{\partial t} = (\lambda + 2\mu)\frac{\partial v_z}{\partial z} + \lambda\frac{\partial v_x}{\partial x} + \varphi_{zz}, \tag{4}$$

$$\frac{\partial \sigma_{xz}}{\partial t} = \mu\left(\frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x}\right), \tag{5}$$

Where x and z are coordinates on the plane, v_x and v_z are velocity components, σ_{xx}, σ_{xz} and σ_{zz} are the stresses and finally φ correspond to the source which generates the wave to be propagated. Simulation of source is performed commonly using a Ricker wavelet given by expression 6

$$\varphi = (1 - 2\pi^2 f^2 t^2)e^{-\pi^2 f^2 t^2} \tag{6}$$

2.2 CPML implementation

When implementing numerical modeling, an undesired behavior of waves reflection is found at boundaries of the model, in order to avoid this issue is necessary to establish some energy absorbing artificial zones near the boundaries and prevent reflections to appear. There are several methods to perform this task, among

them Convolutional Perfectly Matched Layers (CPML) [6] is widely used because it has proved to be very effective to reduce energy from seismic waves compared with other classic options. Basically, CPML method add auxiliary variables to the modeling expressions presented, as shown in equations 7, 8, 9, 10 and 11.

$$\frac{\partial v_x}{\partial t} = \frac{1}{\rho}\left(\frac{\partial \sigma_{xx}}{\partial x} + \frac{\partial \sigma_{xz}}{\partial z}\right) + \Omega_{xx} + \Omega_{xz}, \tag{7}$$

$$\frac{\partial v_z}{\partial t} = \frac{1}{\rho}\left(\frac{\partial \sigma_{xz}}{\partial x} + \frac{\partial \sigma_{zz}}{\partial z}\right) + \Omega_{xz} + \Omega_{zz}, \tag{8}$$

$$\frac{\partial \sigma_{xx}}{\partial t} = (\lambda + 2\mu)\frac{\partial v_x}{\partial x} + \lambda\frac{\partial v_z}{\partial z} + \varphi_{xx} + (\lambda + 2\mu)\Psi_{xx} + \lambda\Psi_{zz}, \tag{9}$$

$$\frac{\partial \sigma_{zz}}{\partial t} = (\lambda + 2\mu)\frac{\partial v_z}{\partial z} + \lambda\frac{\partial v_x}{\partial x} + \varphi_{zz} + \lambda\Psi_{xx} + (\lambda + 2\mu)\Psi_{zz}, \tag{10}$$

$$\frac{\partial \sigma_{xz}}{\partial t} = \mu\left(\frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x}\right) + \mu\Psi_{zx} + \mu\Psi_{xz}, \tag{11}$$

those auxiliary variables are updated following expressions in 12

$$\Omega_{ik}^{n} = b_i\Omega_{ik}^{n-1} + a_i\partial_i^{n+\frac{1}{2}} \qquad \Psi_{ik}^{n} = b_i\Psi_{ik}^{n-1} + a_i\partial_i^{n+\frac{1}{2}} \tag{12}$$

where i and k are x or z coordinates depending which variable is working, a and b represent attenuation coefficients given by the method.

2.3 Discretization

There are several numerical methods for solving the propagation equations, among them, finite differences (FD) is widely used because its easiness to implement. A decision to use second order in time and forth order in space option was made since it allows to achieve an adequate balance for numerical efficiency and small truncation error. Equations 13, 14 and 15 represent discretization expressions for differential operators which are applied over propagation equations.

$$\frac{\partial f_{x,z}^{t}}{\partial t} = \frac{1}{\Delta t}\left[f_{i,k}^{n+\frac{1}{2}} - f_{i,k}^{n-\frac{1}{2}}\right] \tag{13}$$

$$\frac{\partial f_{x,z}^{t}}{\partial x} = \frac{1}{\Delta x}\left[\frac{9}{8}\left(f_{(i+\frac{1}{2}),k}^{n} - f_{(i-\frac{1}{2}),k}^{n}\right) - \frac{1}{24}\left(f_{(i+\frac{3}{2}),k}^{n} - f_{(i-\frac{3}{2}),k}^{n}\right)\right] \tag{14}$$

$$\frac{\partial f_{x,z}^{t}}{\partial z} = \frac{1}{\Delta z}\left[\frac{9}{8}\left(f_{i,(k+\frac{1}{2})}^{n} - f_{i,(k-\frac{1}{2})}^{n}\right) - \frac{1}{24}\left(f_{i,(k+\frac{3}{2})}^{n} - f_{i,(k-\frac{3}{2})}^{n}\right)\right] \tag{15}$$

The discretization procedure generates a grid distribution of the model where all physical dimensions are adapted into memory spaces that will contain information related to elastic parameters, in this grid the length and depth dimensions are turned into memory locations N_x-N_z depending on step sizes Δx and Δz; it is common for avoiding anisotropic effects in modeling to choose $\Delta x = \Delta z = \Delta h$; discretization result is shown in Fig. 1a. In order to properly implement elastic modeling is necessary to use a staggered-grid scheme [8] since velocity fields update depend on stress fields located in other position as shown in Fig. 1b.

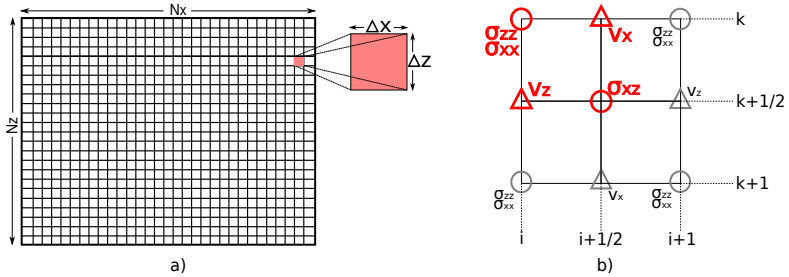


Fig. 1. Discretization result a) grid schematization of model, b) staggered-grid scheme for elastic wave modeling implemented.

2.4 Thread/Block launching layout on GPU for elastic wave modeling

Nvidia Graphic Processing Units (GPU) contains Streaming Multiprocessors (SM) as the main component for launching threads and groups of threads known as *Blocks*, SM schedules necessary resources to run *Kernels* containing the code to be parallelized through the application of Threads/Blocks layouts [4, 10], in the case of elastic wave modeling the parallel code comprise the propagation equations already presented. Once the grid from discretization is created is possible to allocate GPU memory (that could be seen as a matrix) and design launching layouts to properly run kernels and implement modeling.

When implementing a specific layout over a matrix usually the first approach consider using one thread to perform calculations per every memory space allocated since is the simplest and easiest option, this distribution leads to a 1D layout where there is one big block containing all threads processing completely the memory space giving a total number of threads per block of $N_x \times N_z$. Even though the mentioned 1D layout is the first option, is not a practical implementation since hardware architecture limits block sizes to contain a maximum of 1024 threads per block which directly limits the model size to be processed with this layout.

Another design considers allocation of threads to correspond with N_x memory spaces and the number of blocks will be bound to N_z value which represent the rows in matrix configuration, this configuration still keeps the 1D fashion introduced in previous example but in this case the number of blocks utilized is incremented, this design is presented in Fig. 2a. This layout improves over the previous allowing to work with bigger size models but still is limited for block size limit since in this case a row with more than 1024 components could no be processed properly.

In addition to previously mentioned of resource allocation in Nvidia hardware when launching layouts, it is important to mention that GPU architecture considers internally groups of 32 threads as a structure called *warp*, when launching

a Kernel all threads allocated in selected layout are divided to adjust them into the warp size and later some of those warps run concurrently.

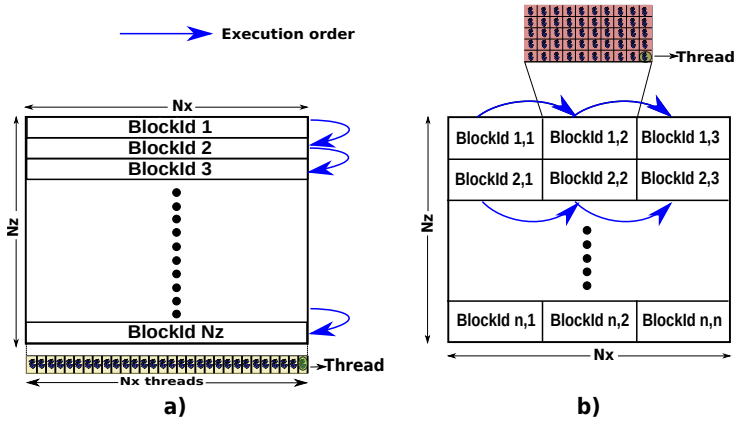


Fig. 2. Threads/Blocks layout for Kernel execution:a) 1D layout, b) 2D layout

The amount of warps available to run simultaneously depends on distribution of registers and shared memory made when one layout is established however this resource scheduling is not directly controlled by the programmer because it is performed automatically by GPU. For this reason, appropriate layout design is important for better use of GPU resources, however there are not specific instructions to follow when designing launching layouts and thus it becomes an empiric procedure highly depending on the application and hardware availability.

Despite the aforementioned, Nvidia has given some suggestions [9] that could help in layout design; among them they recommend to keep the number of threads per block a multiple of warp size, avoidance of small block sizes and carry the number of blocks to a much greater number than SMs. By taking those recommendations it is possible to propose a new layout as depicted in Fig. 2b.

This distribution is achieved by dividing the discretization matrix in smaller 2D matrixes of specific size to cover the space of the bigger, those smaller matrixes will correspond to the blocks in the layout and each of these blocks will have a 2D thread distribution internally. In this way is possible to work with a greater quantity of blocks while it keeps threads per block as multiple of warp size, this layout also let dealing with bigger model sizes easily with no major changes. There are many possibilities for selecting a 2D layout suitable for determined model, a useful suggestion would be execute kernels based on block size of 128 or 256 threads.

2.5 Common and reduced memory CPML memory allocation strategies

When implementing numerical modeling is crucial the definition of absorbing zones for CPML processing, those areas of fixed thickness and length are depicted in Fig. 3a. In the image is clearly differentiated two big areas named Regular and CPML zones, in the first, seismic waves travel unaltered while in the latter wave energy is reduced as you go deeper in that area.

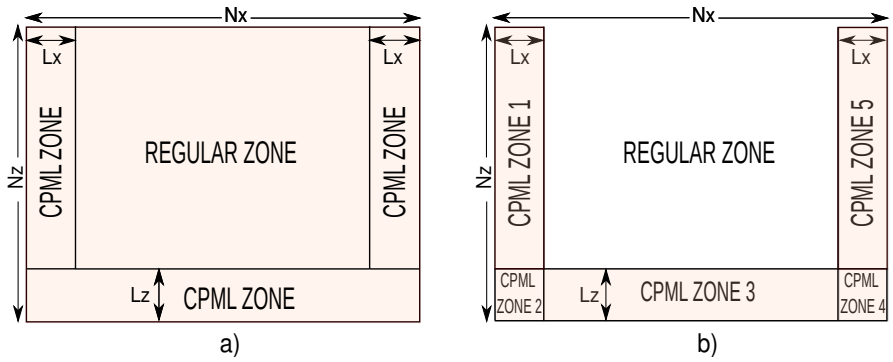


Fig. 3. CPML zones distribution and memory allocation a) common CPML allocation strategy, b) CPML reduced memory strategy.

Each CPML zone has a specific thickness L_x or L_z that extend to cover fully the corresponding dimension N_z or N_x, when implementing over GPU those zones must have their own memory space apart from that used for updating velocity and stress fields where specific calculations are performed. It is common to find CPML implementations in elastic wave modeling that allocate memory spaces of $N_x \times N_z$ bytes looking to keep the easiness in code programming but sacrificing valuable GPU memory resources.

When using common scheme the amount of GPU resources invested in operations is greater since on every iteration a CPML kernel is executed it covers completely (coloured in Fig. 3a) the model memory space allocated including regular zone where is supposed not to be performed any instruction. Despite this, indexing memory locations, jumping over indexes and internal instructions come about which lead to GPU memory and processing threads to be misused, in addition, this inefficient behavior could be worsen if bigger size models are considered.

Another alternative to deal with CPML processing considers 5 CPML zones with reduced memory allocation to keep better control on memory resource and code execution, furthermore Kernels for CPML operations avoid to access the regular zone as shown in Fig. 3b. Unlike the common implementation, the new proposed distribution considers 2 independent memory allocation zones of size

$L_x \times (N_z - L_z)$ bytes, another 2 of $L_x \times L_z$ bytes, finally 1 area of $(N_x - 2L_x) \times L_z$ bytes; using this scheme assures that total memory used in GPU decreases even in cases where model sizes are bigger since allocations formulas only rely on 1 dimension from the model while maintaining constant the other which in most cases is a low value L_x or L_z.

It is important to mention that the amount of GPU resources saved in the new scheme decrease when model size dimensions are diminished leading at some point to a virtual *match* in performance for both implementations of CPML processing, however in a practical application of elastic modeling where is common finding survey areas covering several kilometers of length and depth this strategy could be useful for more efficient resources management in the GPU.

2.6 CPU-GPU memory management for wavefield cube storage

According to Nvida Best practices guidelines [9], GPU memory optimization is the key area to develop when algorithm performance is look for, this is particularly applied in elastic wave modeling where high performance levels are expected specially in execution time and memory consumption because this method is usually used as the first stage in more complex procedures as inversion [3, 13] and migration [7].

Programming over GPU involves the implementation of heterogeneous computing [2] paradigm where basically a CPU act as the controller for execution flow of the program and GPU perform necessary heavyweight operations; each hardware device has its own memory space and one algorithm running under this paradigm necessarily will share information between those spaces for proper execution.

NVIDIA GPU architecture based its kernel execution procedure over a memory hierarchy which includes several memory types; among them the registers, shared and global spaces are worth to be considered since they are the most commonly used in general applications. Registers and shared memory are the fastest memory in GPU but those resources are limited to some Kylobytes, whereas global memory (VRAM) is slower but there are much more available for use which make it ideal for dealing with big size wavefield cubes from modeling.

When executing elastic wave modeling is important to save velocity and stress fields updated and later perform the creation of wavefield cube with those fields, to achieve that, Fig. 4 presents one approach where GPU updates one wavefield and transfers to CPU system memory on every iteration. In this approach, elastic modeling is performed by executing Kernels that update wavefield iteratively, once the field is updated a data transference carry that field from GPU global memory to CPU over PCI-e bus; when the wavefield reaches the system memory it is saved and execution of next iteration for wavefield update in GPU is allowed.

The main drawback in this strategy is associated with the reduced bandwidth available in the PCI-e x16 channel giving a maximum of 16GB/s whereas the VRAM-GPU bandwidth could peak a maximum of 128GB/s (in a Nvidia Turing Gtx 1650 card). In addition to bandwidth, *overhead* due to GPU-CPU transfers could impact execution time of wave modeling.

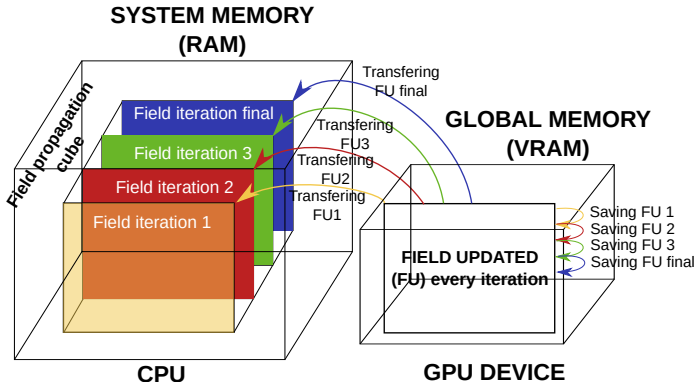


Fig. 4. Wavefield cube saved in CPU system memory (RAM) every iteration GPU updates the wavefield

Another strategy for wavefield cube storage is depicted in Fig. 5, in this case the GPU updates the wavefield but this time is kept in global memory and saved on it creating the cube, once the propagation is finished, a full cube transference is performed to CPU. This strategy overcomes most disadvantages of previous scheme by exploiting the higher bandwidth on GPU transfers to speed up execution of kernels and construction of propagation cube, likewise it performs only one big transfer between GPU-CPU memory spaces which could reduces impact in execution time due to overhead.

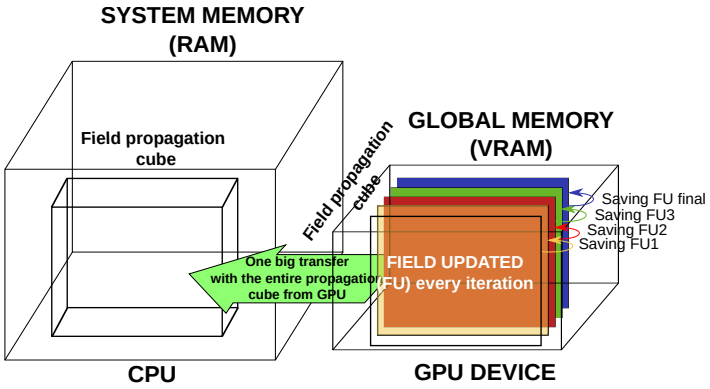


Fig. 5. Wavefield cube saved in GPU global memory (VRAM) every iteration GPU updates the wavefield

3 Results

Proposed tests intent to quantify the impact of GPU launching layout schemes, GPU memory management and data transference when an elastic modeling algorithm is executed. Two different algorithms named *Algorithm 1* and *Algorithm 2* were developed to execute elastic modeling, flowchart in Fig. 6 depicts the general steps included in programming, differences in execution come about when choosing CPML memory scheme (number 1), thread/block layout (number 2) and storage of wavefield cube (number 3).

Algorithm 1 implements a 2D thread/block layout for launching GPU Kernels, it also uses the CPML reduced memory allocation strategy and it manages wavefield creation of v_x field in VRAM in the GPU, on the other side, *Algorithm 2* considers initially a 2D thread/block layout and later is changed to 1D layout, it includes common CPML memory allocation and wavefield construction for v_x field is performed using RAM memory of the CPU.

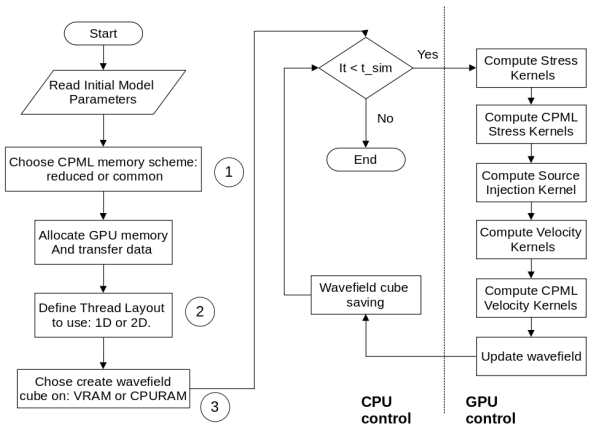


Fig. 6. Flowchart for implementation of elastic modeling in a heterogeneous system highlighting specific locations where both strategies differ

Wave propagation implemented uses 2^{nd} order in time and 4^{th} in space finite differences option, space steps of $\Delta x = \Delta z = 5$ m and time step of $\Delta t = 1$ ms, an isotropic and homogeneous medium with two planar layers with velocities $V_p = 3000$ m/s and $V_p = 1500$ m/s respectively, the two layers shared other parameters as $V_s = 1730$ m/s, $\rho = 2500$ Kg/m^3; the source is simulated by Ricker wavelet with central frequency of 15 Hz. Hardware specifications include a standalone station with a CPU Ryzen 5 3550H, 8GB of RAM; it contains a GPU GTX 1650 from Nvidia with 4GB of Video RAM; the system run under Debian OS and the language used for programming is CUDA C. Fig. 7 shows four

snapshots of elastic wave modeling implemented over an area of 2560 m×1280 m which corresponds to a wavefield of 512×256 points with selected discretization.

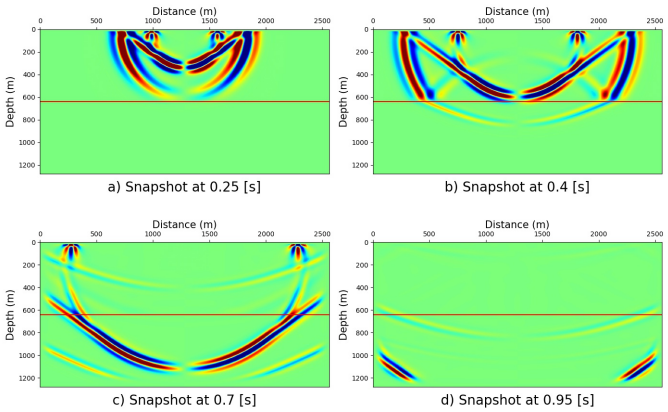


Fig. 7. Elastic wave modeling snapshots in isotropic medium with two layers. a)t = 0.25 s,b) t = 0.4 s, c) t = 0.7 s, d) t = 0.95 s

Table 1 summarizes the results for execution times of each algorithm implemented (Alg 1 and Alg 2) in 3 tests where elastic modeling was performed at 3 simulations times using 3 different models grids. In this case, both codes use 2D layout in order to isolate the impact of memory management and data transference from Kernel launching scheme. Measurements show that Algorithm 1 improves execution time over algorithm 2 ranging from a minimum of 8.9% to 18.1%, there is also a continuous enhancement in execution time when rising both simulation time or model size; however a specially pronounced improvement trend is detected when model size is incremented.

Table 1. Execution times for algorithms 1 and 2 considering 3 different model sizes and 3 simulation times, both algorithms use 2D Layout

Model size	$t_{sim} = 1.0\ s$			$t_{sim} = 1.5\ s$			$t_{sim} = 2.0\ s$		
(points)	Alg1 (s)	Alg2 (s)	Diff (%)	Alg1 (s)	Alg2 (s)	Diff (%)	Alg1 (s)	Alg2 (s)	Diff (%)
256×256	1.01	1.10	8.9%	1.37	1.50	9.4%	1.69	1.86	10.0%
512×256	1.48	1.64	10.8%	2.00	2.30	15.0%	2.57	2.98	15.9%
512×512	2.25	2.56	13.7%	3.24	3.72	14.81%	4.14	4.89	18.1%

Table 2. Execution times for algorithms 1 and 2 considering 3 different model sizes and 3 simulation times, algorithm 1 implements 2D Layout whereas algorithm 2 implements 1D Layout

Model size (points)	$t_{sim} = 1.0\ s$			$t_{sim} = 1.5\ s$			$t_{sim} = 2.0\ s$		
	Alg1 (s)	Alg2 (s)	Diff (%)	Alg1 (s)	Alg2 (s)	Diff (%)	Alg1 (s)	Alg2 (s)	Diff (%)
256×256	1.01	1.44	42.5%	1.37	2.01	46.7%	1.69	2.52	49.1%
512×256	1.48	2.22	50.0%	2.00	3.17	58.5%	2.57	4.29	66.9%
512×512	2.25	4.04	79.5%	3.24	5.83	79.93%	4.14	7.83	88.4%

Table 3. VRAM use for Algorithms 1 and 2 considering 5 different model sizes and a simulation time of 1.5 s

Model size (points)	$t_{sim} = 1.5\ s$				
	Alg1 (MB)	Alg2 (MB)	Diff (MB)	Model size (MB)	Proportion / model size
256×256	435	437	2	0.250	8.0
512×256	815	819	4	0.500	8.0
512×512	1577	1583	6	1.000	6.0
768×512	2345	2361	16	1.500	10.6
768×768	3515	3545	30	2.250	13.3

Table 2 collects same information as in table 1 and similar tests were performed, however in this case algorithm 2 implements 1D layout. The idea at this point is to measure execution times in codes which use not only different CPML memory management strategy but also different launching schemes. Results show vastly gains in time of algorithm 1 over 2 varying from 42.5% climbing to 88.4%, likewise there is an improvement trend going upwards with increments in simulation time and model size. It is important to mention that major gains in execution time were obtained with bigger model sizes tested according to data from both tables 1 and 2.

Table 3 includes data related to GPU memory used for each algorithm over different model sizes, the *Diff* column shows how much more memory algorithm 2 used compared with algorithm 1, this value ranges from 2 to 30MB. Column *Proportion/model size* presents how many times memory used in algorithm 2 is above the memory necessary for saving a model in GPU. Figures on those columns allow us to identify a rising trend in memory consumption of algorithm 2 over 1 due to the strong relation between memory used and model dimensions N_x-N_z for CPML calculation in algorithm 1, this relation is reduced when implementing CPML reduced memory strategy and thus the impact on memory used.

4 Conclusions

The results show that one algorithm using only CPML reduced memory strategy together with wavefield cube creation in GPU improves execution time by a maximum of 18.1% in tests, there also exists a continuous enhancement trend with model size increments and higher times of simulation. If 2D launching layout is added to the strategy, the figures are increased ranging from a minimum of 42.5% to a peak of 88.4%. In all cases major gains were obtained when working with the bigger model sizes for all times of simulation. From data it is noticeable the tremendous influence that Kernel launching layout has over execution time compared with CPML memory management, however the combined effect is outstanding.

Bigger memory gains for algorithm 1 over 2 were found when working with bigger model sizes reaching a maximum of 30MB in tests which make something close to 1% of saving compared to GPU total memory used, however if compared with memory consumption of the model is possible to say that common memory management strategy took 13.3 times more memory than CPML reduced memory strategy when executing wave modeling, also it can be identified a continuous increasing trend in memory consumption when model sizes rise up.

References

1. Alaei, B.: Seismic Modeling of Complex Geological Structures. InTech (2012)
2. Arora, M.: The architecture and evolution of cpu-gpu systems for general purpose computing (2012)
3. Brittan, J., Bai, J., Delome, H., Wang, C., Yingst, D.: Full waveform inversion – the state of the art. First Break **31**(10), 75–81 (2013)
4. Cheng, J., Grossman, M., McKercher, T.: Professional CUDA C Programming. John Wiley & Sons. (2014)
5. Keckler, S.W., Dally, W.J., Khailany, B., Garland, M., Glasco, D.: Gpus and the future of parallel computing. IEEE Micro **31**(5), 7–17 (2011)
6. Komatitsch, D., Martin, R.: An unsplit convolutional perfectly matched layer improved at grazing incidence for the seismic wave equation. Geophysics **72**(5) (Sep 2007)
7. Menke, W.: Geophysical Data Analysis. Elsevier Inc, 4th edition edn. (2018)
8. Moczo, P., Robertsson, J.O., Eisner, L.: The finite-difference time-domain method for modeling of seismic wave propagation. Advances in geophysics **48**, 421–516 (2007)
9. NVIDIA: CUDA C++ Best Practices Guide. Nvidia (Jun 2020)
10. NVIDIA: CUDA C++ Programming Guide. Nvidia (Jun 2020)
11. Permana, T., Sudarmaji, M.: The 2d finite difference numerical modelling of p-sv wave propagation in elastic heterogeneous medium using graphic processing unit: Case study of mount merapi topography, yogyakarta. In: International Conference on Physics 2014 (ICP-14) (2014)
12. Virieux, J.: P-sv wave propagation in heterogeneous media: Velocity-stress finite-difference method. Geophysics **51**(4), 889–901 (1986)
13. Virieux, J., Asnaashari, A., Brossier, R., Métivier, L., Ribodetti, A., Zhou, W.: An introduction to full waveform inversion. In: Encyclopedia of exploration geophysics, pp. R1–1. Society of Exploration Geophysicists (2017)

Traducción de imagen con redes generativas antagónicas condicionales aplicada a la resolución automática de captchas

Tomás Ortiz, Pablo Sández, Rosana Gimenez y Miguel Méndez Garabetti

Laboratorio de Investigación en Ciencia y Tecnología (LabICyT-UDA),
Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua,
Catamarca 147, Mendoza, Argentina
devtomas99@gmail.com

Resumen Las redes generativas antagónicas son redes neuronales que surgieron en 2014 y son una de las ideas más interesantes e innovadoras en informática actualmente debido a sus novedosas y variadas aplicaciones en diferentes campos y problemas. El artículo está orientado específicamente a la traducción de imagen a imagen, y a una variante de estas redes, las redes generativas antagónicas condicionales. Inicialmente, se explica esta potente tecnología, posteriormente se realiza un experimento realizado a modo de demostración, que consiste en el desarrollo de un algoritmo aplicado a un problema en concreto: la resolución automática de captchas. Si bien este problema ya ha sido resuelto con otras tecnologías, se dispone a resolverlo con esta, concretamente. Finalmente se obtienen y exponen los resultados obtenidos, estos son, los captchas legibles, y se verifica su calidad de imagen y eficiencia con la que fue solucionado.

Keywords: Inteligencia artificial · Redes generativas antagónicas · Traducción de imagen a imagen · Resolución automática de captchas · Procesamiento de imágenes

1. Introducción

Muchos problemas en el procesamiento de imágenes, gráficos y visión implican traducir una imagen de entrada en una imagen de salida correspondiente. Estos problemas a menudo se tratan con algoritmos específicos de cada problema, a pesar de que el proceso es siempre el mismo: asignar píxeles a píxeles. Las redes generativas antagónicas (GANs) son una solución de propósito general para una amplia variedad de estos problemas [1]. Estas redes son una de las ideas más interesantes en informática hoy en día, donde dos modelos son entrenados simultáneamente por un proceso de confrontación [3].

Las GANs han logrado resultados impresionantes para muchas aplicaciones del mundo real. Estas redes han logrado producir imágenes fotorrealistas, indistinguibles de las imágenes reales. Esta notable capacidad ha impulsado muchas

aplicaciones de reconocimiento visual, manipulación de imágenes, hasta la predicción de video. Desde su invención en 2014, se han propuesto muchas variantes de GANs, que producen muestras más realistas y con una mejor estabilidad de entrenamiento [8].

Las GANs son un enfoque prometedor para muchas tareas de traducción de imagen a imagen. Estas redes aprenden y se adaptan a la tarea y los datos disponibles, lo que las hace aplicables en una amplia variedad de entornos [1], siendo utilizadas en campos como la medicina, ingeniería, videojuegos, arte, edición, entre otros.

Uno de los principales objetivos de este trabajo es explicar y demostrar la potencialidad de las redes generativas antagónicas, más específicamente la variante condicional, aplicado a un problema concreto, en este caso, a la resolución automática de captchas.

El artículo está organizado de la siguiente manera: la sección 2 explica la traducción de imagen a imagen. La sección 3 habla sobre el tema principal, las redes generativas antagónicas condicionales. La sección 4 describe la arquitectura del generador y discriminador. La sección 5 expone el problema a resolver. La sección 6 consiste en la experimentación realizada. La sección 7 muestra los resultados obtenidos. Finalmente, la sección 8 realiza las conclusiones.

2. Traducción de imagen a imagen

Las GANs se han utilizado ampliamente en el dominio de la visión por computadora: superresolución de una imagen, traducción de texto a imagen, traducción de imagen a imagen, imagen en pintura, etc. También se han utilizado para otras tareas, como la generación de música, generación de texto y la mejora del habla [9]. Este artículo se centrará en la traducción de imagen a imagen.

Se define la traducción automática de imagen a imagen como la tarea de traducir una posible representación de una escena en otra, dados los datos de entrenamiento suficientes. Tradicionalmente, cada una de estas tareas se ha abordado con algoritmos diferentes para fines especiales, a pesar de que el entorno es siempre lo mismo: predecir píxeles a partir de píxeles [1]. Uno de los aspectos más importantes de estas redes es la generalidad, para una variedad de tareas y conjuntos de datos, como se observa en la Fig. 1.

Una característica principal de los problemas de traducción de imagen a imagen es que asignan una cuadrícula de entrada de alta resolución a una cuadrícula de salida de alta resolución. Tanto la entrada como la salida son representaciones de la misma estructura [1]. Algunas aplicaciones de las GANs en la traducción de imagen a imagen:

- Mapa $\longleftrightarrow$ Foto aérea.
- Fotos blanco y negro $\rightarrow$ Fotos en color
- Bordes $\rightarrow$ Foto
- Boceto $\rightarrow$ Foto
- Foto de día $\rightarrow$ Foto de noche
- Fotos térmicas $\rightarrow$ Fotos en color

• Foto con píxeles faltantes → foto pintada

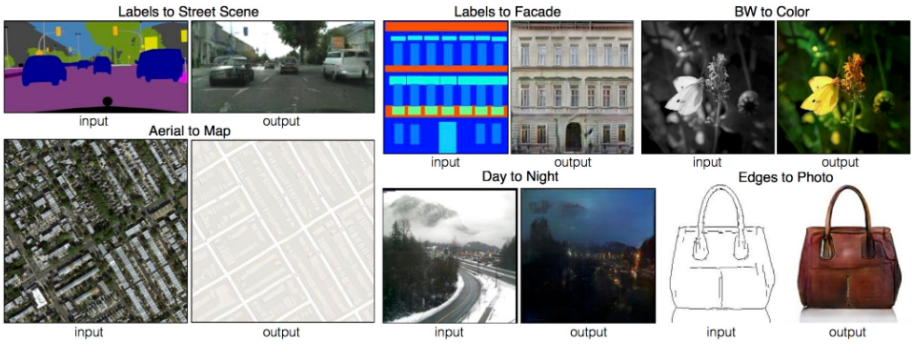


Figura 1. Ejemplos de resultados de varios problemas de traducción de imagen a imagen. En cada caso se usa la misma arquitectura y objetivo, simplemente entrenando con diferentes datos. (imagen extraída de [2])

3. Redes generativas antagónicas condicionales

Las redes generativas antagónicas (GANs) aprenden una pérdida que intenta clasificar si la imagen de salida es real o falsa, mientras que simultáneamente entrena un modelo generativo para minimizar esta pérdida. Estas redes se pueden aplicar a una multitud de tareas que tradicionalmente requerirían tipos muy diferentes de funciones de pérdida [1].

El generador G está entrenado para producir salidas que no pueden distinguirse de las imágenes 'reales' por un discriminador entrenado de manera adversa, D, que está entrenado para detectar las 'falsificaciones' del generador [1]. Por lo tanto, D devuelve la probabilidad de que la imagen de entrada (x) sea una imagen real de la distribución de datos. En este sentido, G está entrenado para engañar a D al generar imágenes que se parecen a las imágenes de entrenamiento reales [8]. Por el contrario, D está entrenado para maximizar la probabilidad de clasificar correctamente tanto a las muestras reales como a las muestras falsas, de G [7].

De manera crucial, G no tiene acceso directo a imágenes reales del conjunto de entrenamiento, ya que aprende únicamente a través de su interacción con D. Si G es capaz de igualar la distribución de datos reales, entonces D estará confundido al máximo, prediciendo 0.5 para todas las entradas [8].

En otras palabras, un generador sería el 'artista' que aprende a crear imágenes que parecen reales, mientras que un discriminador es el 'crítico de arte' que aprende a distinguir las imágenes reales de las falsas [3].

Para mejorar el rendimiento del modelo GAN original, se han propuesto muchas variantes en los últimos años. Se logró una mejora adicional al introducir

la supervisión en el entorno de entrenamiento, como en las GAN condicionales. Estas variantes condicionales mejoran la calidad de la imagen generada, al tiempo que mejoran la estabilidad del proceso de entrenamiento [9].

Las GANs pueden extenderse a un modelo condicional (CGANs) si tanto el generador como el discriminador están condicionados por alguna información adicional y. El valor y podría ser cualquier tipo de información auxiliar, como etiquetas de clase, y se proporciona en el conjunto de datos [11]. En este caso, la etiqueta y corresponde a la imagen objetivo a la que se quiere llegar.

Las CGAN son efectivas en problemas donde la salida es altamente detallada o fotográfica, como es común en el procesamiento de imágenes y tareas gráficas [1]. También se obtienen buenos resultados cuando la salida es menos compleja que la entrada, es decir, cuando no requiere tantos detalles.

Función objetivo GAN El objetivo de GAN, implica encontrar un equilibrio de Nash en un juego de minimax de dos jugadores [10]. Es decir, el discriminador debe:

- **Maximizar** la probabilidad de que una muestra real sea clasificada como real.
- **Minimizar** la probabilidad de que una muestra falsa sea clasificada como real.

Por el contrario, el generador debe:

- **Maximizar** la probabilidad de que una muestra falsa sea clasificada como real.

La función objetivo de la CGAN es la misma que una GAN excepto que al generador y discriminador se le agrega una nueva entrada y, que representa la información adicional que explicamos anteriormente. En la Fig. 2 se puede observar el flujo de datos para una CGAN.

Durante el entrenamiento, el generador progresivamente mejora en la creación de imágenes que parecen reales, mientras que el discriminador mejora en distinguirlas. Las imágenes generadas comienzan como ruido aleatorio, y a medida que el generador se entrena éste se acerca más a la imagen objetivo, es decir, aquella imagen a la que se quiere llegar para engañar al discriminador, como se observa en la Fig. 3. El proceso alcanza el equilibrio cuando el discriminador ya no puede distinguir las imágenes reales de las falsas [3].

Las GAN generalmente se entrenan utilizando técnicas de descenso de gradiente, que están diseñadas para encontrar un valor mínimo de una función de error. Este entrenamiento consiste en encontrar un equilibrio de Nash para un juego no cooperativo de dos jugadores. Cada jugador (generador y discriminador) busca minimizar su propia función de error [6].

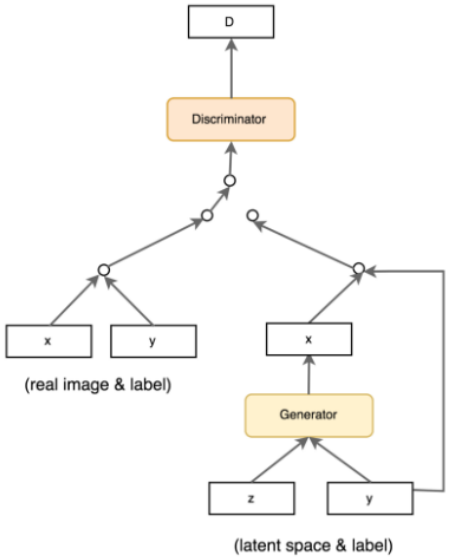


Figura 2. Flujo de datos para una CGAN. Se introduce como entrada la etiqueta y (imagen objetivo) y un ruido aleatorio z al generador para crear una imagen x (salida del generador) que se parezca al valor de la etiqueta y. El discriminador recibe como entrada dos partes: la imagen generada x y la etiqueta y, y la imagen real y la etiqueta y. (imagen extraída de [4])

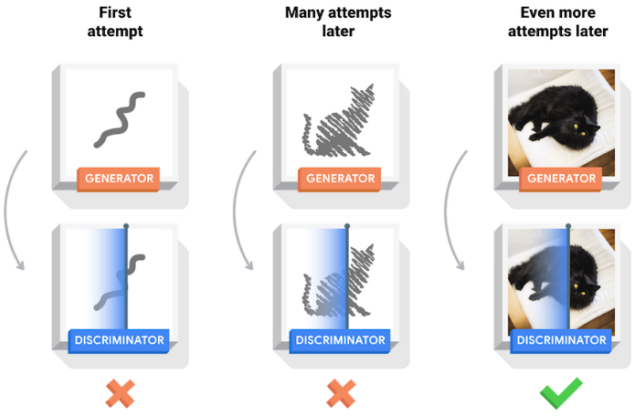


Figura 3. El discriminador, D, aprende a clasificar entre imágenes reales y falsas. Después de muchos intentos, el generador logra crear una imagen lo suficientemente realista como para engañar al discriminador. (imagen extraída de [3])

4. Arquitectura

A continuación, se describen las arquitecturas del generador y discriminador. Para el generador se va a tener en cuenta la arquitectura codificador-decodificador y U-Net, y para el discriminador PatchGAN.

4.1. Arquitectura del generador

Hay dos opciones para la arquitectura del generador: codificador-decodificador y U-Net. Esta última es un codificador-decodificador con conexiones de salto entre capas del codificador y decodificador [1]. Esto se observa en la Fig. 4.

Codificador-decodificador En esta arquitectura, el codificador es una red convolucional, en donde la imagen de entrada se pasa a través de una serie de capas secuenciales que disminuye progresivamente la muestra, hasta llegar a una capa de cuello de botella (punto de conexión entre el codificador y decodificador), en donde el codificador ha extraído la información relevante para la tarea que se quiere resolver. En la última capa el proceso se invierte. Esta inversión viene dada por el decodificador, que es una red deconvolucional, en la que la imagen se pasa por una serie de capas secuenciales que aumenta progresivamente la muestra, hasta alcanzar la última capa, que genera la imagen final, con el mismo tamaño con la que se introdujo al codificador.

El decodificador tiene que predecir pixeles en cada capa a partir de la entrada, es decir, se está generando una nueva imagen. Por lo tanto, cada capa realiza un procesamiento a una matriz de valores numéricos de diferente tamaño, que representa los pixeles de la imagen.

Esta arquitectura requiere que todo el flujo de información pase obligatoriamente a través de todas las capas, incluido el cuello de botella. Para muchos problemas de traducción de imágenes, existe una gran cantidad de información de bajo nivel compartida entre la entrada y la salida, y sería deseable transferir esta información directamente a través de la red [1].

U-Net En la arquitectura anterior, el cuello de botella podría no dejar pasar información suficiente ya que a este nivel dicha información estaría muy procesada o sintetizada, por lo cual no se podrían resolver algunas tareas con la calidad necesaria, como por ej. la segmentación. La solución a este problema son las conexiones de salto.

Para darle al generador un medio para evitar el cuello de botella, agregamos conexiones de omisión o salto que servirán como atajo para que se pueda saltar una serie de capas. Específicamente, agregamos estas conexiones entre capas del mismo nivel. Cada conexión de salto concatena todos los canales en la capa i con esos en la capa n - i [1].

Por consiguiente, una arquitectura U-Net permite que la información de bajo nivel tenga acceso directo a través de la red, y, por lo tanto, que no se pierda. Esto conduce a mejores resultados para ciertos problemas. Por lo tanto, el codificador-decodificador no puede aprender a generar imágenes realistas [1].

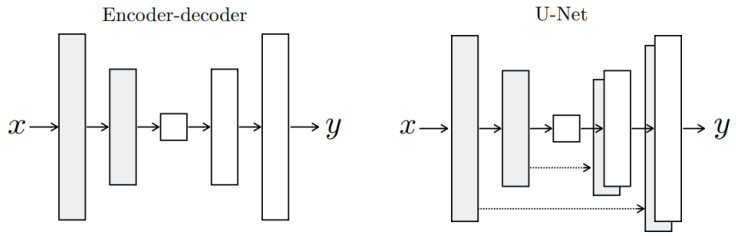


Figura 4. La arquitectura U-Net utiliza la arquitectura codificador-decodificador y agrega conexiones de salto entre capas del mismo nivel. (imagen extraída de [1])

4.2. Arquitectura del discriminador

PatchGAN La arquitectura del discriminador, PatchGAN, intenta clasificar si cada área de NxN en una imagen es real o falso. Se ejecuta este discriminador de forma convolucional, promediando todas las respuestas para proporcionar el resultado final de D, indicando si la muestra es real o falsa. N puede ser mucho más pequeño que el tamaño completo de la imagen y aun así producir resultados de alta calidad [1]. Por ejemplo, la imagen de entrada al discriminador puede ser de 256x256x3 y generar una salida de 16x16x1. Tal salida se modela como un mapa de colores, prediciendo qué áreas parecen reales y cuáles falsas, como se observa en la Fig. 5.

La utilización de esta arquitectura es ventajosa porque tiene menos parámetros, se ejecuta más rápido y se puede aplicar a imágenes grandes [1].

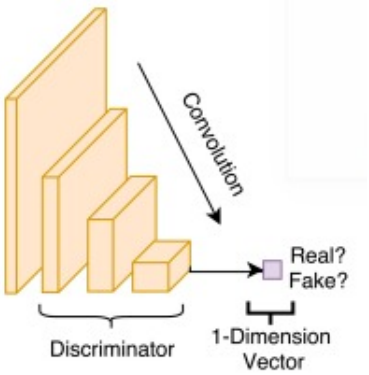


Figura 5. Arquitectura PatchGAN. La salida es una matriz de NxN de valores numéricos, que representa una imagen que clasifica como real o falsa una muestra recibida por el discriminador. (imagen extraída de [12])

5. Problema: resolución de captchas

El problema a resolver será la resolución de captchas, debido a que es un problema típico donde muchas veces se pierde tiempo intentando descifrar captchas complejos. Para esto, se introducirá una imagen de entrada que será el captcha a resolver, y el modelo generará una imagen de salida del captcha resuelto, en una forma más legible y prolija. A continuación, se construirá el modelo y se observará qué tan eficiente son las CGANs para esta tarea en concreto.

6. Experimentación

El experimento consiste en un desarrollo propio de un algoritmo con CGAN, y se describe el conjunto de tecnologías utilizadas, el banco de datos donde se extrajeron los datos, las arquitecturas seleccionadas, y por último el entrenamiento y validación del modelo.

6.1. Tecnologías utilizadas

Para la construcción del modelo, se utilizó el lenguaje de programación Python, con las bibliotecas TensorFlow y Keras, en Google Colab. Para visualizar gráficamente el error de entrenamiento y validación se empleó TensorBoard.

6.2. Banco de datos

Se utilizó un banco de datos de 100000 imágenes de 200x50, en formato PNG, de captchas sin resolver [5], como se observa en la Fig. 6. Sin embargo, para esta tarea no se empleará la totalidad de las imágenes provistas, sino una pequeña porción. Del banco de datos solo se utilizarán 5000 imágenes en total. El conjunto de datos se dividirá en dos: conjunto de entrenamiento y validación. Para el entrenamiento se utilizarán 4000 imágenes (80 %), y para la validación 1000 (20 %). Para entrenar al modelo es necesario no solo las imágenes de entrada sino también las imágenes objetivo correspondientes, en este caso, los captchas resueltos. En el banco de datos no existen tales imágenes, sin embargo, se pueden generar mediante código.

Tanto las imágenes de entrada y objetivo se normalizarán desde el rango [0, 255] a [-1, 1], para trabajar con valores más pequeños a fin de que sea más sencillo su manejo.

Figura 6. [5] Ejemplar del tipo de imagen utilizado para el entrenamiento y validación del modelo.

6.3. Arquitecturas seleccionadas

Arquitectura del generador La arquitectura del generador es una U-Net. El codificador es una red convolucional compuesta por bloques. Cada bloque tiene las siguientes capas: Convolution-BatchNorm-LeakyReLU, con k filtros. El decodificador es una red deconvolucional compuesta por bloques. Cada bloque tiene las siguientes capas: Deconvolution-BatchNorm-Dropout-ReLU, con k filtros y una tasa de abandono del 50 %.

En el generador el factor de reducción de resolución es de 2. Por lo tanto, en cada bloque del codificador, la imagen se reduce a la mitad. Por el contrario, en el decodificador aumenta a la mitad. El codificador y decodificador están compuesto por 8 y 7 bloques, respectivamente. A continuación, se especifican los filtros aplicados en cada bloque:

Codificador: 64 - 128 - 256 - 512 - 512 - 512 - 512 - 512

Decodificador: 512 - 512 - 512 - 512 - 256 - 128 - 64

Después del último bloque en el decodificador, se aplica una deconvolución para especificar el número de canales de salida (3, RGB), seguido de una función tangente hiperbólica.

Arquitectura del discriminador La arquitectura del discriminador es una PatchGAN, compuesta por una red convolucional, por lo tanto, de bloques del codificador, donde se comprime la imagen bloque por bloque hasta llegar a un tamaño NxN, que determine, mediante un mapa de colores, si la imagen es real o falsa. El discriminador está compuesto por 4 bloques del codificador. A continuación, se especifican los filtros aplicados en cada bloque:

Codificador: 64 - 128 - 256 - 512

Después del último bloque, se aplica una deconvolución para especificar una salida unidimensional (1 canal), seguida de una función sigmoidea.

Las arquitecturas utilizadas se basan en el artículo [1].

Todas las redes fueron entrenadas desde cero. Los pesos son inicializados a partir de una distribución gaussiana con media 0 y desviación estándar 0.02.

6.4. Entrenamiento y validación

El entrenamiento se basa en un esquema de aprendizaje supervisado, ya que en el conjunto de datos se conoce la salida deseada para cada imagen de entrada, y los parámetros se modifican en base al error cometido.

El conjunto de validación está compuesto por aquellas imágenes que el algoritmo nunca ha visto antes, y se utiliza para verificar si éste está aprendiendo adecuadamente, es decir, si está generalizando o memorizando los datos de entrenamiento.

El modelo solo fue entrenado por 60 épocas a causa de las limitaciones en el uso de los recursos que proporciona Google Colab. En cada época se introduce todo el conjunto de entrenamiento, y al finalizar cada una de ellas se realiza una

validación con 12 imágenes aleatorias extraídas del conjunto de validación. Durante todo el entrenamiento y validación se registran los errores de entrenamiento y validación: error del discriminador y generador.

7. Resultados obtenidos

En la Fig. 7 se observan algunos resultados aceptables, mientras que en la Fig. 8 resultados que no se asemejan fielmente a las imágenes objetivo. Estos resultados son obtenidos a partir de imágenes aleatorias de entrada del conjunto de validación, con su respectiva imagen objetivo. Además, se muestran gráficamente los errores tanto de entrenamiento -línea naranja- como de validación -línea azul-, desde el inicio hasta la finalización del entrenamiento, como se observa en la Fig. 9 para el error del discriminador y Fig. 10 para el error del generador. Estos errores se pueden visualizar numéricamente en la Tabla 1.

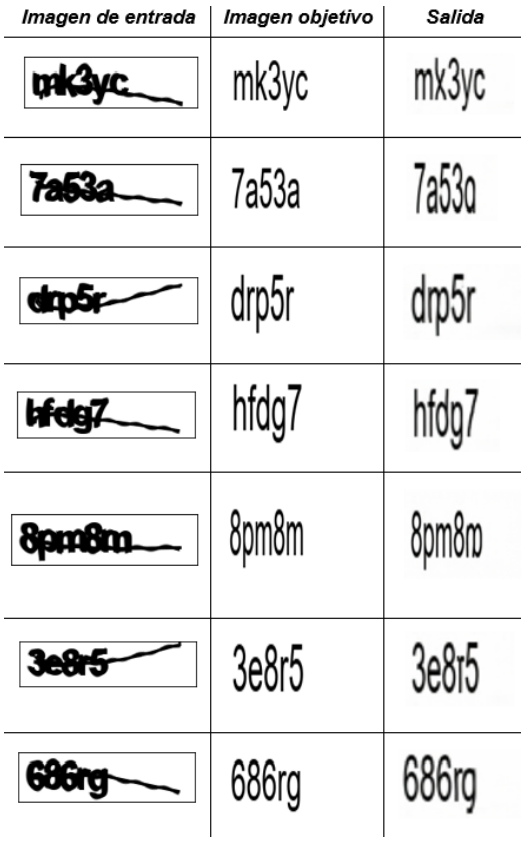


Figura 7. Resultados aceptables, obtenidos después de 60 épocas de entrenamiento, utilizando imágenes de validación como entrada.

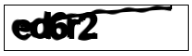
Imagen de entrada	*Imagen objetivo*	*Salida*
	2yexy	2yoxy
	wnrw2	wnnv2
	ed6r2	od6i2
	36aw8	36av8

Figura 8. Otros resultados que no se asemejan fielmente a las imágenes objetivo.

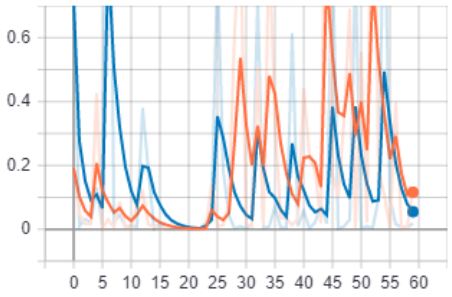


Figura 9. Error del discriminador registrado durante el entrenamiento y validación, hasta la época 60.

En la tabla y gráficos presentados se observa que, tanto para el entrenamiento y validación, el error del discriminador es mínimo, por lo que discrimina eficientemente las imágenes reales y falsas. En comparación, el error del generador es alto, por lo que genera imágenes que muchas veces no logran engañar al discriminador. Sin embargo, como se observa en la Fig. 7, el generador ha logrado crear imágenes que se asemejan a las imágenes objetivo.

8. Conclusiones

Con solo 60 épocas de entrenamiento del modelo, se lograron obtener resultados aceptables en la resolución automática de captchas. Sin embargo, el error del generador no fue mínimo, por lo que las imágenes generadas no eran perfectas e introducían artefactos, pero con más tiempo y muestras de entrenamiento se podrían obtener aún mejores resultados.

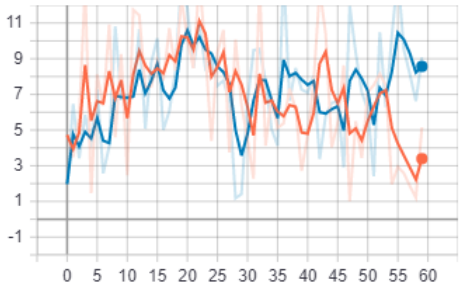


Figura 10. Error del generador registrado durante el entrenamiento y validación, hasta la época 60.

	Error discriminador		Error generador	
	Valor inicial	Valor final	Valor inicial	Valor final
Entrenamiento	0.1918	0.128	4.726	5.159
Validación	0.7211	0.0183	1.973	9.086

Tabla 1. Errores de entrenamiento y validación

Una de las grandes desventajas de las GANs es que requieren grandes cantidades de datos, diversos y muy concretos, para producir resultados admisibles. En consecuencia, esto demanda un tiempo considerable para el entrenamiento del modelo.

Referencias

1. P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, and B. A. Research, "Image-to-Image Translation with Conditional Adversarial Networks."
2. https://phillipi.github.io/pix2pix/
3. https://www.tensorflow.org/tutorials/generative/dcgan
4. https://jonathan-hui.medium.com/gan-cgan-infogan-using-labels-to-improve-gan-8ba4de5f9c3d
5. https://www.kaggle.com/digdeepbro/100000-labeled-captchas
6. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs."
7. I. J. Goodfellow et al., "Generative Adversarial Nets."
8. D. Bau et al., "GaN dissection: Visualizing and understanding generative adversarial networks."
9. M. Ben-Yosef and D. Weinshall, "Gaussian Mixture Generative Adversarial Networks for Diverse Datasets, and the Unsupervised Clustering of Images."
10. A. Brock, J. Donahue, and K. Simonyan, "Google DeepMind 2019 Large scale GaN training for high fidelity natural image synthesis."
11. M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets."
12. F. Yamashita, R. Orihara, Y. Sei, Y. Tahara, and A. Ohsuga, "Factors Affecting Accuracy in Image Translation based on Generative Adversarial Network."

REFLEXIONES SOBRE APRENDIZAJE PRESENCIAL VS. REMOTO

Un caso de estudio

Laura Nidia Aballay[1, 2][0000-0001-5462-7683], María Inés Lund[2, 1][0000-0002-1687-7267]

[1] Instituto de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Argentina
[2] Departamento de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Argentina
{laballay, mlund}@iinfo.unsj.edu.ar

Abstract. Este trabajo presenta un análisis y comparación de resultados de aprendizaje y de medición de satisfacción de alumnos de 4to año, en la asignatura Ingeniería de Software II de las carreras Licenciaturas en Sistemas de Información y Ciencias de la Computación, de la Universidad Nacional de San Juan, correspondientes al cursado 2019 (con modalidad presencial) y 2020 (con modalidad virtual). Las variables consideradas fueron las notas finales obtenidas por los alumnos y las valoraciones dadas por ellos en la encuesta final de cursado, agrupadas en dimensiones de calidad. Se exponen, en base a los resultados obtenidos, las ventajas y desventajas de cada una de las modalidades, y posibles mejoras para el contexto virtual.

Keywords: Educación presencial, Educación virtual, Aprendizaje.

1 Introducción

A principios del año 2020, se declara a COVID-19, como pandemia [1]. En Argentina, el 15/03 se anuncia el Aislamiento Social, Preventivo y Obligatorio (ASPO) lo que provocó la suspensión de clases [2]. La crisis por el Covid-19 ha obligado a los profesores a adaptar sus clases, según la UNESCO [3], afrontando la transición hacia una educación de emergencia, encarando planes de contingencia que permitan continuar el dictado de clases a través de entornos virtuales en poco tiempo, tratando de no afectar la calidad, promoviendo la inclusión social y minimizando la deserción por falta de conectividad.

Este formato pedagógico caracterizado por la migración abrupta al aprendizaje en línea, se conoce como educación on-line de emergencia o enseñanza remota de emergencia (ERE) [4] y se la está identificando como *"Coronateaching"*, que es la expresión de los esfuerzos docentes por usar los escasos recursos tecnológicos disponibles para dictar sus cursos, como si siguieran en situación de aula y sin la correspondiente capacitación para ello [3], [5].

Para afrontar este cambio de paradigma en la enseñanza de la asignatura Ingeniería de Software II, del 1er cuatrimestre de 4to año de las carreras de Licenciatura en Sistemas de Información (LSI) y Licenciatura en Ciencias de la Computación (LCC), carreras del Departamento de Informática de la Facultad de Ciencias Exactas, Físicas y Naturales (FCEFN) de la Universidad Nacional de San Juan (UNSJ), en la cual las autoras son docentes, se propuso realizar las actividades de aprendizaje aplicando un modelo probado de trabajo colaborativo en entorno distribuido, basado en la técnica Jigsaw [6], utilizando plataforma Zoom para las clases síncronas (de teoría y de práctica) y Google Drive, correo electrónico y grupo de WhatsApp para las comunicaciones y entregas. Una vez finalizada la cursada, se propuso analizar las diferencias entre la educación presencial y la virtual, de los años 2019 y 2020, respectivamente, donde la modalidad de trabajo es igual, pero cambió el ambiente de presencial a virtual. Se propone comparar la satisfacción del aprendizaje y el nivel de aprendizaje alcanzado. Los resultados de satisfacción serán analizados por medio de encuestas de satisfacción que se realizan todos los años y el nivel de aprendizaje en base a la calificación final obtenida (que es un valor ponderado de los controles teóricos y el proyecto desarrollado).

El artículo está ordenado como sigue: En la sección 2, se presentan los antecedentes en las actividades y prácticas que se aplican en el desarrollo del trabajo. A continuación el marco teórico, donde se aclaran los distintos tipos de educación y en la sección 4 se explica el caso de estudio. En las secciones 5 y 6 se presentan los datos y el análisis correspondiente a los años 2019 y 2020. La comparación de las variables analizadas se realiza en la sección 7, y finalmente en la sección 8, las conclusiones y trabajo a futuro o mejoras.

2 Antecedentes

La asignatura Ingeniería de Software II desde el año 2017 se dicta siguiendo un modelo de enseñanza-aprendizaje colaborativo, utilizando la técnica Jigsaw [6] para los contenidos teóricos, permitiendo que sean los propios alumnos quienes gestionen la adquisición de su conocimiento, y una metodología de Aprendizaje Basado en Problemas [7], [8], para la elaboración del diseño de un sistema real (usando UML como lenguaje de modelado) para la parte práctica. En el año 2014 el trabajo [9] presenta un modelo de proceso de enseñanza aprendizaje. Y en [10] se expone un trabajo siguiendo ese modelo.

También, desde hace varios años se aplica en la asignatura una encuesta de satisfacción, de fin de curso para los alumnos, donde los enunciados están correlacionados a una dimensión de calidad, que permite medir distintos aspectos que incumben a la satisfacción de la asignatura, como se presentó en los trabajos [11], [12], [13].

En el presente año, en el marco de la pandemia, adicionalmente a la encuesta de satisfacción se realizó una encuesta de co-evaluación entre los compañeros de trabajo colaborativo, para medir las habilidades blandas. Estos resultados se presentarán en [14], aún no está publicado.

3 Marco Teórico

La educación presencial o tradicional, ha ido cambiando en las últimas décadas con el advenimiento de internet, plataformas educativas, nuevas tecnologías de comunicación, etc. En esta sección se presentará brevemente las modalidades educativas generales y la actual que nos introdujo la pandemia de Covid-19. La formación online y la presencial persiguen la calidad de la enseñanza, cada una utiliza su propio método y establece sus propios canales y pautas de aprendizaje [15].

3.1 Educación presencial.

En la educación presencial o convencional, se requiere y exige la presencia física del alumno y del profesor en el aula. Los alumnos adquieren un papel más pasivo y se adaptan al ritmo y al método del profesor, quien se encarga de explicar, aclarar, evacuar dudas y comunicar ideas y experiencias [15], [16]. O, en su forma más dinámica, donde los alumnos participan en la adquisición de su conocimiento y el profesor pasa a ser un facilitador o guía de dicho aprendizaje. Ventajas: El trato con el profesor es más cercano. Posibilidad de socializar. La disciplina, al tener que cumplir con horarios prefijados. Desventajas: Los costos son más altos. El mantenimiento de las infraestructuras y los desplazamientos la encarecen. Los horarios son más rígidos. Los contenidos tardan en actualizarse [15].

3.2 Educación virtual (no presencial)

La educación virtual, se refiere al desarrollo de actividades de aprendizaje mediante entornos no presenciales, donde el educando puede instruirse sin la necesidad de asistir físicamente a un centro educativo [16]. La educación virtual recurre a Internet y a las tecnologías de la información y la comunicación (TIC), como chats, blogs, videoconferencias o documentos compartidos. La educación virtual también tiene sus ventajas y desventajas [15]. Ventajas: No hay barreras geográficas. Se puede acceder desde cualquier lugar. Los horarios son flexibles. Siempre se puede acceder a los contenidos. Menores costos. Contenidos en formato digital que facilita su renovación constante. Desventajas: Exige organización y fuerza de voluntad. Dificulta el contacto alumno-docente por ser virtual. Reduce socialización. Dentro de lo que es Educación virtual, se pueden distinguir variantes:

3.2.1 Educación a distancia (EaD)

La educación a distancia es una forma de enseñanza virtual. Requiere de la implementación de didácticas y metodologías adecuadas para lograr el aprendizaje. Requiere cierto grado de autodisciplina. Está concebida como enseñanza asincrónica, donde los alumnos aprenden a su propio ritmo e incluye características como evaluaciones continuas [17]. La organización de un programa de EaD demanda planificación, tiempo y recursos técnicos, humanos y financieros. Es un proceso que puede durar un año o algo

más, con el apoyo de diseñador instruccional, ilustrador, diseñador gráfico, corrector de estilo, productor (para video o audio), etc.[18].

3.2.2 Educación en línea efectiva (educación online o e-learning)

Es una variante de la EaD, la diferencia es que únicamente se basa en la www para sus actividades sincrónicas y asincrónicas. Es la forma de educación en la que los docentes y estudiantes participan en un entorno digital haciendo uso intensivo de las facilidades que proporciona Internet y las tecnologías digitales. Por ello, planificar, diseñar y desarrollar una unidad curricular/curso o programa en línea también requiere de mucha preparación, puede tomar entre seis y nueve meses diseñar un curso on-line [18]. Una de las grandes desventajas de la educación on-line es su alta dependencia de la tecnología [18], [19], [20],.

3.2.3 Educación Remota de Emergencia (ERE)

Este tipo de enseñanza también es conocida como "Docencia no presencial de Emergencia o Aprendizaje Remoto". El estado de emergencia decretado globalmente debido a la pandemia por COVID-19 requirió una transición forzada, para no demorar el dictado de clases, hacia un proceso de aprendizaje en línea en muy poco tiempo, sin demasiada planificación ni elaboración de material instruccional, y con docentes poco o nada preparados. Esta implementación de la ERE es la que se ha denominado, en el contexto actual, "*coronateaching"*, que de acuerdo a la definición de Luz Montero, directora UC Online de la Pontificia Universidad Católica de Chile, no es más que el proceso de *"transformar las clases presenciales a modo virtual, pero sin cambiar el currículum ni la metodología"* [3], conduciendo esto al riesgo de que podrían concluir en resultados poco óptimos, frustración y agobio, sin la correspondiente capacitación para ello [5]. La ERE se diferencia del aprendizaje en línea porque representa un cambio temporal y abrupto de la pedagogía debido a circunstancias de crisis; mientras que, la educación en línea, es una experiencia planificada [19], [20].

4 Caso de Estudio

Se analizaron y contrastaron los resultados de satisfacción del cursado y de nivel de aprendizaje de los alumnos de la asignatura Ingeniería de Software II, de 4to año de las carreras LSI y LCC del Departamento de Informática de la FCEFN de la UNSJ. Esta asignatura contempla entre sus objetivos, no sólo la formación en diseño de software, sino también promover en los alumnos habilidades de trabajo en grupo y habilidades blandas. Se implementa el aprendizaje colaborativo, siguiendo el método *Jigsaw* [6], brindándoles conceptos generales, guiándolos con las lecturas recomendadas y monitoreando la dinámica y cooperación de los grupos. Luego se realizan evaluaciones del conocimiento adquirido, mediante distintas técnicas (controles, exposiciones, *checklist*, evaluación entre pares, etc.). No se brindan apuntes, sino las presentaciones áulicas, lecturas recomendadas y una amplia lista de bibliografía para consultar. Para la práctica, ellos desarrollan el diseño de un sistema propuesto por ellos, un caso real. Una vez que

finaliza el cursado, se mide siempre la satisfacción de los alumnos por medio de una encuesta, respecto a los contenidos, a la forma de trabajar y a sus logros obtenidos. Esta encuesta tiene el formato de respuesta tipo *likert* desde "muy insatisfecho" hasta "muy satisfecho", con 5 niveles de valoración, además de contar con espacios de comentarios positivos y negativos. Los resultados de satisfacción fueron analizados por medio de las encuestas y el nivel de aprendizaje en base a la calificación final obtenida (que es un valor ponderado de los controles teóricos y el proyecto desarrollado).

4.1 Encuesta y dimensiones

La encuesta consta de varios ítems, que se encuentran agrupados en dimensiones de calidad para su medición y mejor interpretación. En este trabajo se detallan sólo las dimensiones referidas a la satisfacción de la asignatura y del aprendizaje. A continuación se exponen los enunciados por dimensión:

- *Dimensión Satisfacción General de la asignatura:* Los contenidos de la materia son interesantes. Me gustó mucho la materia. Estoy conforme con el dictado de la materia
- *Dimensión Satisfacción del aprendizaje con metodología colaborativa Jigsaw:* El trabajo colaborativo, como metodología de enseñanza aprendizaje de aspectos más teóricos, me pareció mejor que las clases tradicionales. Esta metodología ayuda a asimilar mejor los conocimientos que la forma tradicional. La metodología de enseñanza aprendizaje me permitió aprender en forma completa todos los temas del grupo. Me parece bien aplicar este tipo de estrategias para la enseñanza de conceptos teóricos. La metodología colaborativa me incentivó a gestionar la búsqueda de mi propio conocimiento.
- *Dimensión Satisfacción con los contenidos y planificación de las clases:* Los tiempos asignados para la asignatura son adecuados. El nivel (contenidos y exigencias) de la materia es apropiado para alumnos de 4º año. Los contenidos de la materia son dictados en forma organizada, coordinando teoría y práctica. Los contenidos de la materia son dictados en tiempo y forma. Los contenidos de la asignatura fueron abordados en forma clara
- *Dimensión Metodología de Práctica:* La práctica realizada permite comprender el objetivo y uso de los Diagramas UML. La metodología desarrollada en la práctica permitió asimilar los contenidos. La práctica es útil para aprender a aplicar los Diagramas UML. La aplicación de la práctica de modelamiento me guio en el reconocimiento de los requisitos del sistema. La práctica de modelamiento me ayudó a completar el diseño del sistema a desarrollar. Los ejemplos explicados me sirvieron para comprender las actividades a realizar
- *Dimensión Metodologías de Evaluación de la teoría (Jigsaw):* La exposición, como evaluación de los trabajos de grupos de especialistas, me pareció adecuada. El control, como evaluación individual del trabajo en grupo, me permitió reforzar los conocimientos del tema. El informe, como evaluación grupal de los trabajos colaborativos, me permitió integrar el conocimiento del grupo. Las distintas estrategias de

evaluación del aprendizaje me parecieron adecuadas. La forma de evaluar los contenidos teóricos me sirvió como método de aprendizaje.

5 Educación Presencial: Año 2019

En el año 2019, y como todos los años previos, la educación fue presencial. Se cumplió con la planificación de cátedra en tiempo y forma. Se dictaron clases teóricas y prácticas con metodología de trabajo en grupos colaborativos, los estudiantes debieron cumplir con trabajos prácticos grupales y rendir controles individuales. Al finalizar se les entregó una encuesta de satisfacción de la asignatura. Cursaron y aprobaron 23 alumnos, 18 de la LCC y 5 de la LSI.

5.1 Análisis de las encuestas

Del procesamiento de los resultados obtenidos de las respuestas de los 23 alumnos, se muestra en la figura 1 el porcentaje de satisfacción, teniendo en cuenta las respuestas puntuadas con 4 y 5 (satisfechos y muy satisfechos) respecto de los enunciados correspondientes a cada dimensión:

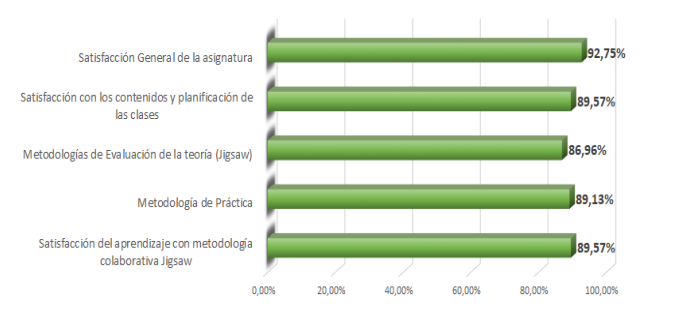


Fig. 1. Porcentaje de "satisfechos" y "muy satisfechos" de las dimensiones seleccionadas.

Se puede observar que casi el 93% de los alumnos se encuentra satisfecho y muy satisfecho con la asignatura, casi el 90% puntúa como satisfecho y muy satisfecho respecto a los contenidos de la asignatura y la planificación, con la metodología de práctica y con el aprendizaje colaborativo.
En la figura 2 se observa lo mismo que en la 1, pero discriminado por grado de satisfacción (4 y 5). Se visualiza el porcentaje de alumnos que puntuaron como "satisfechos" y los que puntuaron como "muy satisfechos. En general, salvo "metodología de evaluación de la teoría", la mayoría valora como "muy satisfechos" esas dimensiones. En la dimensión "Satisfacción general de la asignatura", el promedio general de los "satisfechos" junto con los "muy satisfechos" (figura 1), es del 92.75%, donde el 42.03% están "satisfechos" y el 50.72% "muy satisfechos" (figura 2).

Se puede concluir que el nivel de satisfacción de esta dimensión es alto, quedando apenas el 7.25% del total con valores menores a 4, donde 3 respuestas "neutrales" (valor 3 de la escala Likert) en el enunciado "Me gustó mucho la materia.", y 2 "neutrales" en

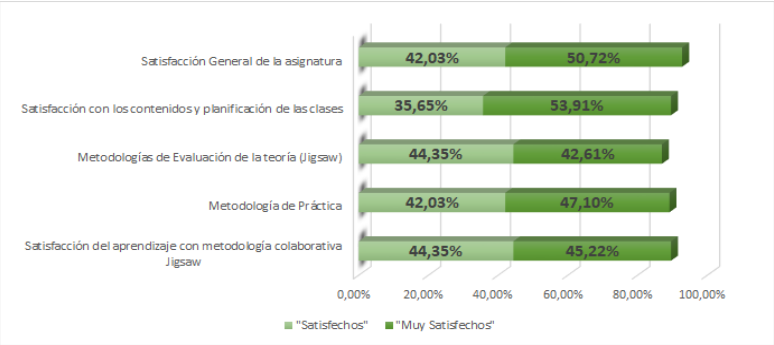


Fig. 2. Porcentaje de "satisfechos" y "muy satisfechos" de las dimensiones seleccionadas.

el ítem "Estoy conforme con el dictado de la materia", sin embargo no dejaron comentarios negativos al respecto que evidencien el motivo. No se registraron insatisfechos (valores menores que 3 de la escala). La dimensión "Metodologías de Evaluación de la teoría", es la que tiene menor porcentaje de satisfacción (86.96%), el 44.35% está "satisfecho" y el 42.61% "muy satisfecho". El motivo de que esta dimensión no alcance valores más altos se evidencia por la influencia de las puntuaciones del enunciado "El informe, como evaluación grupal de los trabajos colaborativos, me permitió integrar el conocimiento del grupo", con 3 respuestas "neutrales" y 2 "insatisfechos". Tal vez se debe a que a los estudiantes no les gusta redactar informes, o se les hizo difícil hacerlo en forma grupal. Sin embargo, en los comentarios de los casos en que los valores fueron "neutrales" o "insatisfecho" no se evidencia disconformidad. Este es un análisis elemental de los resultados de la encuesta, pero permite observar que hay un alto nivel de satisfacción con la asignatura y la forma de trabajar en ella.

5.2 Análisis de las notas

La materia es Promocional, por lo que se aprueba con nota 7. Sin embargo, en el 2019, la totalidad de los alumnos que cursaron lograron aprobar. La nota final fue resultado de un promedio ponderado de notas de controles teóricos individuales, exposiciones y trabajo final de diseño con UML. En la tabla 1 se muestran las notas de aprobación y la cantidad de alumnos que alcanzaron la misma. Podría interpretarse como el nivel de aprendizaje con respecto a la asignatura.

Tabla 1. Notas finales de los alumnos

Notas	7	8	9	10
Cantidad de alumnos	4	7	11	1

Se observa que la mitad más uno de los alumnos lograron un nivel de aprendizaje de 9 y 10, (muy bueno y excelente), lo que puede interpretarse que este tipo de Educación y modalidad de trabajo permite lograr exitosamente los objetivos propuestos por la asignatura.

6 Educación Virtual: Año 2020.

En este año se alcanzó a dar una semana presencial. La primera semana del ASPO las docentes se dedicaron a la tarea de buscar aplicaciones para dar las clases, en organizar la cursada por medios virtuales. Se consultó a los alumnos acerca de los medios de conexión y computadoras con que contaban y en poco tiempo se optó por la plataforma Zoom para las clases presenciales sincrónicas y las exposiciones de los alumnos. Los controles de evaluación se realizaron por Google Forms, y las consultas por Zoom o Meet de Google. Se siguió usando Google Drive, haciendo el seguimiento de la parte práctica en forma individual. La planificación original tuvo que ser readaptada a los nuevos tiempos y se pudo cumplir con ella. Esta modalidad educación es la ERE denominada en esta oportunidad como "Coronateaching", tal como se explica en el Marco Teórico. Cursaron y aprobaron 18 alumnos, 13 de la LCC y 5 de la LSI.

6.1 Análisis de las encuestas

Al finalizar el cursado, como todos los años, se les distribuyó la encuesta de satisfacción en la que los alumnos podían dejar sus opiniones acerca de la satisfacción o no respecto a diferentes aspectos de la materia y en general. Los resultados obtenidos de las 18 encuestas respondidas se muestran en la figura 3, indicando el porcentaje de satisfacción, teniendo en cuenta las respuestas puntuadas con 4 y 5 (satisfechos y muy satisfechos) respecto de los enunciados correspondientes a cada dimensión:

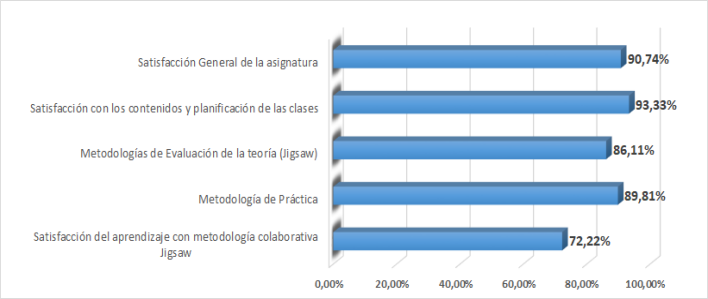


Fig. 3. Porcentaje de "satisfechos" y "muy satisfechos" (acumulado) de las dimensiones seleccionadas

Se puede observar que el 93,33% de los alumnos se encuentra "satisfecho" y "muy satisfecho" con los "contenidos y planificación de las clases", casi el 91% puntúa como satisfecho y muy satisfecho la "satisfacción general de la asignatura".
En la figura 4 se observa lo mismo que en la figura 3, pero discriminado por grado de satisfacción (4 y 5). Se visualiza el porcentaje de alumnos que puntuaron como "satisfechos" y los que puntuaron como "muy satisfechos. La dimensión "Satisfacción general de la asignatura" tiene casi un 91% de "satisfechos" y "muy satisfechos" (figura 3), donde el 35.19% están "satisfechos" y el 55.56% "muy satisfechos" (figura 4). Se puede concluir que el nivel de satisfacción de esta dimensión es alto, quedando apenas el 9.26% del total con valores insatisfechos o neutrales. Es notoria la valoración para la

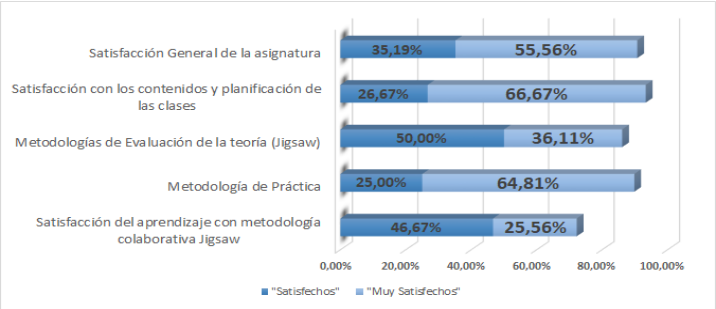


Fig. 4. Porcentaje de "satisfechos" y "muy satisfechos" de las dimensiones seleccionadas.

dimensión "Satisfacción del aprendizaje con metodologías colaborativas", es la que menor puntaje obtuvo (72.22%), donde el 46.67% está "satisfecho" y el 25.56% "muy satisfecho". El bajo puntaje de esta dimensión se evidencia debido a que el 27.78% de las respuestas fueron menores a 4; "neutrales" (valor 3 de la escala) es decir "ni satisfechos ni insatisfechos", ó "insatisfechos" (valor 2 de la escala). El resto de las dimensiones presentan valores cercanos o mayores al 90%, lo que evidencia conformidad por parte de los estudiantes.

6.2 Análisis de las notas

En la tabla 2 se muestran las notas de aprobación y la cantidad de alumnos que alcanzaron la misma. En esta oportunidad también se la considerará como el nivel de aprendizaje con respecto a la asignatura.

Tabla 2. Notas finales de los alumnos.

Cantidad de alumnos	8	5	5
Notas	8	9	10

Se observa que de 18 alumnos 10 de ellos lograron un nivel de aprendizaje de 9 y 10, (muy bueno y excelente), los restantes alcanzaron un 8 y no se presenta ninguna nota menor. Se podría interpretar que la ERE y la modalidad de trabajo permiten lograr exitosamente los objetivos propuestos por la asignatura.

7 Comparativa Presencial vs. Virtual

7.1 Comparativa de Satisfacción

En las secciones 5.1 y 6.1 se analizaron los resultados de las encuestas de satisfacción del año 2019, y del año 2020.

En la figura 5 se muestra la comparación de los años y diferencias entre cada una de las dimensiones. Comparando los valores para cada una de las dimensiones, se puede decir que en ambos años la "Satisfacción general de la asignatura" fue muy buena, con valores mayores al 90% (92.75% y 90.74% respectivamente).

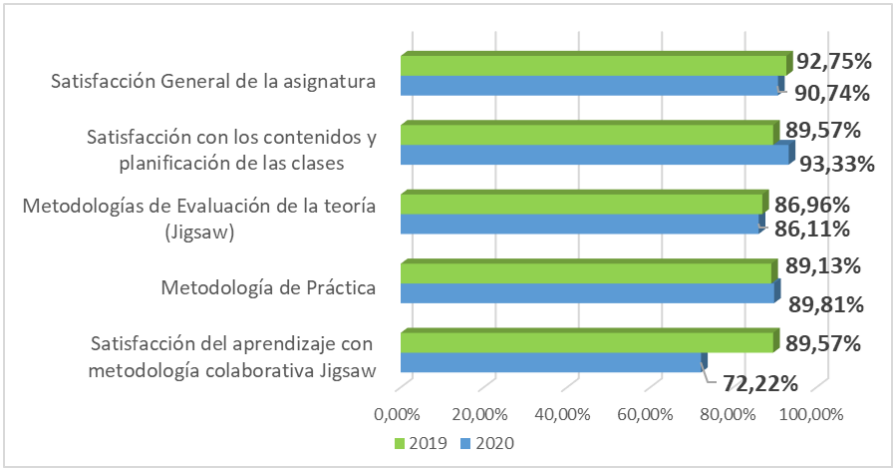


Fig. 5. Comparación de la satisfacción por dimensión correspondiente a los años 2019 y 2020.

En cuanto a la "satisfacción con los contenidos y planificación de clases" el año 2020 supera casi en un 4% al año anterior. Varios alumnos consideraron que la asignatura se adaptó muy bien a la virtualidad, algunos comentarios: "La forma de dictado de esta materia respecto a otras me gustó más, se notaba que las clases estaban bien preparadas, las diapositivas se podían ver bien, el audio se escuchaba bien, además los videos de las clases fueron de mucha utilidad", "Mucho más organizado, con tiempos mejor medidos y con buena predisposición por parte de las profesoras.", "El tener un cronograma de actividades en lo personal me ayudó a ser más organizada".

El resto de las dimensiones no han tenido grandes diferencias, excepto para la "Satisfacción del aprendizaje con metodología *Jigsaw*". Esta dimensión en el año 2020 fue mucho menor, con 72.22% de "satisfechos" y "muy satisfechos" respecto al 89.57 del año anterior, por los detalles ya explicados en la sección 6.1.

7.2 Comparativa de Nivel de Aprendizaje

Para poder comparar el nivel de aprendizaje, debido a que la cantidad de alumnos en cada año fue diferente, se calculó el porcentaje de alumnos que alcanzaron cada una de las notas, como se muestra en la figura 6.

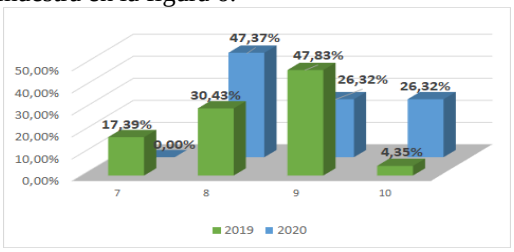


Fig. 6. Comparación de Notas obtenidas en los años 2019 y 2020

Se pueden observar las diferencias en el porcentaje obtenido en cada nota. En el año 2019 el 17,39% de alumnos lograron un nivel de aprendizaje de 7. En cuanto a la nota 8, se puede ver que el 30,44% de los alumnos del año 2019 alcanzaron esa nota respecto al 47,37% del año 2020, superando al año anterior y presencial. En el año 2019, el 47,83% de los alumnos alcanzaron un 9, en diferencia al 2020, que sólo un 26,32% lo hicieron. Y, la calificación excelente (diez) fue lograda por el 4,35% de los alumnos del año 2019 (corresponde a un sólo alumno, como se pudo observar en la tabla 2), frente a un 26,32% de los del año 2020 (5 alumnos, tabla 4). Los alumnos del año 2019 alcanzaron el 78,27% la calificación de muy bueno (8 y 9), casi 5% más que los del año 2020, en el que el 73,69% tuvieron esa calificación.

8 Conclusiones

En ambos años se mantuvo constante las variables de modalidad de trabajo colaborativa para los contenidos teóricos, evaluaciones, desarrollo de la práctica y encuesta de satisfacción. Lo que varió es la educación presencial versus virtual y, por supuesto, los alumnos y la cantidad de alumnos. Por ello fue posible realizar esta comparación.

La satisfacción general de los alumnos no se ve alterada. Los alumnos se encuentran satisfechos a pesar de la virtualidad forzada, inclusive hubo varios comentarios de que fue la materia mejor organizada y que cumplió con los tiempos y lo programado. No se observa un impacto negativo en la ERE. Se transcriben algunos comentarios positivos de la virtualidad: "*muy bueno, esta es la que mejor se adaptó a la modalidad virtual", "El contar con las lecturas recomendadas me permitió entender las bases, me permitió tener una mayor compresión de los temas", "No se nos exigió que estudiáramos de memoria sino que investigáramos por nuestra cuenta y nos explayáramos en los que nos parecía más importante."*. Otros comentarios del 2020, respecto a los problemas que se presentaron: *"Lo malo fue que no fue fluido, al ser sesiones de 40 minutos, pero por lo demás excelente", "En lo personal se me dificultó mucho el cursado al principio por no tener acceso a internet. Pero estuvo genial la dinámica y trabajo que hubo en la materia", "Me resultó complicada por no tener acceso a internet. Pero más allá de eso me resultó interesante".*

La dimensión que más se debería enfatizar o considerar mejorar, suponiendo un próximo año virtual, es la de "satisfacción con el aprendizaje colaborativo siguiendo la metodología *Jigsaw*", donde no están seguros sobre la intención del proceso de enseñanza aprendizaje siguiendo una metodología colaborativa guiada, o si aprenden más o mejor que de manera tradicional y en cómo identificar las situaciones de grupo (dificultades y desencuentros), a través de la virtualidad, como la del año 2020.

Respecto del nivel de aprendizaje, en la educación presencial las notas se distribuyeron entre 7, 8 y 9 y un único 10 (tabla 1); en la educación virtual las notas se distribuyeron entre 8, 9 y 10 (tabla 2). Se podría inducir que la ERE obtuvo mejores resultados en cuanto a nivel de aprendizaje.

Más allá de que los alumnos manifiestan satisfacción con la virtualidad, el trabajo ha sido intenso, con muchas horas de dedicación, y con disponibilidad casi 5x12hs o

más. La cátedra considera muy fructífera la experiencia y quizá para otros años pueda considerarse una modalidad mixta, para potenciar lo positivo de cada una.

Para terminar las conclusiones se cita de InGenio Learning lo siguiente: "**la educación presencial siempre será la mejor experiencia de aprendizaje"**, ya que a nivel evolutivo los seres humanos somos criaturas sociables, y si existe un contacto más directo con otras personas interesadas en los mismos temas, la experiencia se vuelve más enriquecedora" [16]. Muchos expertos concuerdan que la educación presencial, sigue siendo por excelencia la mejor experiencia educativa.

Referencias

1. Comunidad wikipedia, "Pandemia de enfermedad por coronavirus de 2019-2020 - Wikipedia, la enciclopedia libre," 2020. [Online]. Available: https://es.wikipedia.org/wiki/Pandemia_de_enfermedad_por_coronavirus_de_2019-2020.
2. "Pandemia de COVID-19 en Argentina ," *Wikipedia, la enciclopedia libre*, 2020. [Online]. Available: https://es.wikipedia.org/wiki/Pandemia_de_COVID-19_en_Argentina.
3. C. Pérez Escobar, "El gran test de las clases 'online' - La Tercera," *La tercera*, 2020. [Online]. Available: https://www.latercera.com/tendencias/noticia/el-gran-test-de-las-clases-online/JOJOMO7S2BAB3FNRJYPPHGUZ3I/.
4. "Coronateaching ¿síndrome o nueva oportunidad para la reflexión? I/II – UNESCO-IESALC." [Online]. Available: https://www.iesalc.unesco.org/2020/07/02/coronateaching-sindrome-o-nueva-oportunidad-para-la-reflexion-i-ii/.
5. F. Henriquez Ordenes, "El rol clave del profesor en la educación on line durante la crisis sanitaria," *El Universal*, 16-Apr-2020. Available: https://eluniversal.cl/contenido/11348/el-rol-clave-del-profesor-en-la-educacion-on-line-durante-la-crisis-sanitaria.
6. E. Aronson and S. Patnoe, *The Jigsaw Classroom: Building Cooperation in the Classroom*, 2nd ed. Longman, 1997.
7. C. Y. Shim, M. Choi, and J. Y. Kim, "Promoting Collaborative Learning in Software Engineering by Adapting the PBL Strategy," *World Acad. Sci. Eng. Technol.*, vol. 53, pp. 1167–1170, 2009.
8. L. P. Uribe Rico, "Aprendizaje basado en problemas: un método que contribuye en la calidad de la educación superior en Colombia," Universidad Militar NUEVA GRANDADA, 2018.
9. M. I. Lund, L. N. Aballay, M. C. Gomez, and E. G. Ormeño, "Propuesta de un Proceso de Software de Enseñanza-Aprendizaje para la asignatura Diseño de Software," *Rev. Iberoam. Tecnol. en Educ. y Educ. en Tecnol.*, vol. 12, no. especial nº 12, pp. 44–52, Apr. 2014.
10. M. I. Lund, "Diseño de Software basado en Casos de Uso. Aplicación práctica dirigida por un proceso de enseñanza-aprendizaje propuesto," in *Proceedings ASSE-JAIIO 2014. en edicion*, 2014.
11. M. I. Lund and M. Herrera, "Curso Diseño de software. Una propuesta para medir la calidad educativa," in *XX CACIC - III Workshop de Innovación en Educación en Informática (WIEI).*, 2014, pp. 11–21.
12. M. Herrera, M. I. Lund, and L. N. Aballay, "Caracterización de alumnos de una materia mediante datos multivariados obtenidos de un cuestionario de satisfacción," in *XXI Congreso Argentino de Ciencias de la Computació*, 2015.
13. L. N. Aballay and M. I. Lund, "Aprendizaje Colaborativo en un curso de Ingeniería de Software: Medición de la Satisfacción," in *CICCSI 2018 - Congreso Internacional de Ciencias de la Computación y Sistemas de Información - ISBN 978-987-45683-6-6.*, 2018.

14. L. N. Aballay and M. I. Lund, "Promoviendo el desarrollo de habilidades blandas en entornos distribuidos y colaborativos. Una estrategia de enseñanza-aprendizaje en tiempos de pandemia," in *49 Jornadas Argenitnas de Informática*, 2020, p. 10.
15. "Educación Presencial o Virtual, ¿cuál es mejor?" *IBERDROLA*. [Online]. Available: https://www.iberdrola.com/talento/educacion-presencial-vs-virtual.
16. "Educación presencial Vs Educación virtual," *InGenio Learning*, 22-Apr-2020. [Online]. Available: https://ingenio.edu.pe/educacion-presencial-vs-educacion-virtual/.
17. D. Baxter Ph.D. and B. Simon Ph.D., "Clash of the Timelines: Lessons Learned from the Front Lines of CS Education," in *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, 2014, pp. 69:1--69:5.
18. J. Mosquera, "Diferencias entre la Enseñanza Remota de Emergencia, la educación on-line y la educación a distancia," *aporrea.org*, 20-Apr-2020. [Online]. Available: https://www.aporrea.org/educacion/a289478.html.
19. J. L. Abreu, "Tiempos de Coronavirus: La Educación en Línea como Respuesta a la Crisis (Times of Coronavirus: Online Education in Response to the Crisis)," *Daena Int. J. Good Conscienc.*, vol. 15, no. 1, pp. 1–15, Jul. 2020.
20. C. Hodges, S. Moore, B. Lockee, T. Trust, and A. Bond, "The Difference Between Emergency Remote Teaching and Online Learning". *EDUCAUSEreview*, Mar. 2020.

Implementación de práctica DevOps en aplicación serverless

Federico Sanchez[1], Ana Laura Molina[1], Lorena Parra[1], Nelson Rodriguez[1], María Murazzo[1]

[1] Departamento de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan. San Juan, Argentina.
fgsanchez@unsj-cuim.edu.ar, amolina@unsj-cuim.edu.com, lorenaparra152@gmail.com, nelson@iinfo.unsj.edu.ar, marite@unsj-cuim.edu.ar

Resumen. Este artículo se refiere a la inclusión de una práctica DevOps, denominada despliegue continuo, en una aplicación web *serverless* que hace uso de tecnología BaaS y FaaS, Amazon Simple Storage Service (S3) y AWS Lambda, respectivamente. El objetivo es analizar el comportamiento de la implementación generada, en relación con los beneficios y limitaciones de despliegue continuo y de aplicaciones *serverless* de forma independiente, para comprobar la utilidad de la propuesta. Los resultados obtenidos exponen una maximización de los beneficios propios del despliegue continuo al ser integrado con esta nueva arquitectura, lo que evidencia la utilidad de la implementación generada.

Palabras clave: DevOps, despliegue continuo, *serverless*, AWS Lambda

1 Introducción

La adopción de DevOps en el desarrollo de software ha incrementado mundialmente, y es posible atribuir su popularidad a que sus prácticas impactan positivamente en la calidad y el éxito del desarrollo de software [1]. DevOps es un enfoque en el cual se mezclan los roles involucrados en la ingeniería de software tradicional y se mejora la comunicación, con el objetivo de mejorar la frecuencia de las implementaciones en producción y mantener la calidad requerida del software. A su vez, la computación *serverless* es una tendencia que crece en la industria de desarrollo de software, con el correspondiente aumento del interés por parte de académicos y aficionados [2]. Su principal beneficio es permitir que el desarrollador se enfoque únicamente en su tarea de desarrollar, sin ocuparse de aspectos de administración del servidor [3].

Adoptar prácticas DevOps en aplicaciones *serverless* pareciera ser un desafío que trae aparejado mejoras en la calidad del software y en la calidad del proceso de desarrollo de software. Este trabajo se centra en implementar una práctica DevOps, denominada

despliegue continuo, en una aplicación web completamente *serverless*. Además, se busca determinar cómo la aplicación se ve afectada por esta práctica, determinando beneficios y desventajas de esta forma de trabajo. Se analizará si se mantienen los beneficios y limitaciones, tanto de despliegue continuo como de *serverless*, de forma individual.

Actualmente, pocos estudios se han enfocado en implementar prácticas de DevOps en aplicaciones *serverless*. En uno de estos trabajos [4] se toma como base un conjunto de 27 prácticas de DevOps y se determina si la arquitectura *serverless* afecta a la implementación de cada una de estas prácticas. Otro trabajo, sirve de manual técnico, detallando paso a paso cómo realizar la implementación de algunas prácticas de DevOps en diferentes proveedores *cloud*.

El presente artículo se desarrolla de la siguiente manera. En la sección 2 se presenta una revisión de la literatura existente sobre DevOps y aplicaciones *serverless*. En la sección 3 se describe la implementación generada. En la sección 4 se refleja el resultado obtenido tras analizar beneficios y limitaciones de la implementación. Por último, en la sección 5 se presentan las conclusiones obtenidas y los posibles trabajos futuros.

2 DevOps

El concepto DevOps surge en el año 2009 [5] como resultado de un movimiento enfocado en fomentar una mayor colaboración entre todos los involucrados en la entrega de software, con el fin de dejar productivo software valioso de manera rápida y confiable. Si bien no existe una única definición de DevOps globalmente aceptada [6], en este trabajo vamos a considerarlo como [7]: "Un conjunto de prácticas que pretenden reducir el tiempo transcurrido entre el *commit* de un cambio en un sistema y el cambio impactado en producción, mientras se asegura una buena calidad". Esta definición permite distinguir dos momentos de tiempo críticos, el primero cuando el desarrollador confirma los cambios realizados en el sistema, que marca el final del desarrollo y el comienzo del proceso de implementación, y el segundo cuando se realiza el pase a producción que permite que los cambios realizados puedan ser accedidos por el usuario final de la aplicación. Por otro lado, la calidad estará asegurada en la medida en la que el sistema cumpla con los requisitos especificados [8].

Poner en producción una nueva versión de un sistema, origina la posibilidad de que surjan ciertos fallos que puedan llevar al incumplimiento de los requisitos y, a su vez, a la insatisfacción del cliente. Si a este escenario se suma la ejecución manual de múltiples tareas, tales como: definición de cuáles serán las funcionalidades a dejar productivas, coordinar el horario del pase a producción, asegurar la compatibilidad de componentes, llevar registro de versiones, entre otros, aumenta aún más la probabilidad de que surjan fallos.

Las acciones que se ejecutan entre el *commit* y el pase a producción pueden ser automatizadas, eliminando así la intervención manual, disminuyendo el riesgo de errores y facilitando el proceso de implementación para cualquier desarrollador del equipo. La

automatización de cada acción da origen a una práctica de DevOps distinta: integración continua, entrega continua, y despliegue continuo. Integración continua (del inglés *continuous integration*) es el proceso que permite, de manera automática, construir los paquetes de código y ejecutar las pruebas necesarias para asegurar el cumplimiento de los criterios de calidad establecidos (pruebas de integración y pruebas de aceptación del usuario) [9]. Entrega continua (del inglés *continuous delivery*) es el proceso que permite, de forma automática, generar un entregable del código que se ha creado o modificado que está preparado para ser desplegado en producción [9]. Despliegue continuo (del inglés *continuous deployment*) es el proceso por el cual se permite que el código escrito se pueda pasar a producción de manera automática [9]. La diferencia entre estas dos prácticas es que el despliegue continuo tiene como objetivo pasar automáticamente a producción cada cambio que se produzca en el código, mientras que, en la entrega continua, el pase a producción está condicionado por una decisión del negocio, ejecutada manualmente, en la cual se determina cuándo y qué se va a implementar. Se puede afirmar entonces que el despliegue continuo incluye a la entrega continua, pero no ocurre lo mismo en el sentido contrario [10].

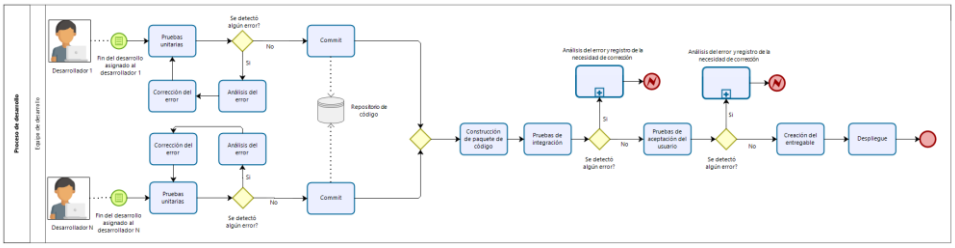


Fig. 1. Diagrama de proceso del desarrollo de software

En la Fig.1, se muestra el diagrama de proceso del desarrollo de software contemplado en este trabajo. Este consiste en una serie de actividades que son ejecutadas entre la finalización del desarrollo asignado a un desarrollador, y el despliegue de la nueva versión del software en el servidor de producción. En este diagrama es posible identificar cuáles son las actividades que se relacionan con cada práctica DevOps mencionada. La automatización de la construcción de paquetes de código, junto a las pruebas de integración y de aceptación, conforman la integración continua. Luego, la automatización de la creación del entregable y el despliegue, forman parte tanto de la entrega continua como del despliegue continuo, la diferencia radica en que el despliegue será desencadenado de forma manual o automática, respectivamente.

3 Aplicaciones Serverless

El término *serverless* no implica la inexistencia de servidores o de lógica de servidor, sino que hace énfasis en la posibilidad que tienen los desarrolladores de delegar a su proveedor *cloud* muchas de las acciones operativas relacionadas al mantenimiento del servidor, tales como actualizaciones de sistemas operativos, tolerancia a fallos, escalabilidad y monitoreo [11]. Este término cubre un amplio rango de técnicas y tecnologías, y es por esto que se divide en dos áreas:

- *Backend as a Service* (BaaS): implica el uso de servicios *backend* ofrecidos por proveedores externos tales como bases de datos, gestores de autenticación, entre otros [12]. BaaS busca reemplazar el desarrollo y la administración de ciertos componentes del servidor con la implementación de componentes de productos externos.
- *Function as a Service* (FaaS): es un modelo de ejecución de cómputo en el *cloud* donde la lógica se ejecuta en contenedores sin-estado que son disparados por eventos y totalmente manejados por terceros [13]. En esta arquitectura, se divide la aplicación en funciones pequeñas y autocontenidas, en lugar de tener una única aplicación monolítica que ejecuta múltiples funciones [3].

Las aplicaciones que hacen uso de Baas, FaaS o ambos, son denominadas aplicaciones *serverless* o aplicaciones con arquitectura *serverless* [13].

En este trabajo, la aplicación *serverless* desarrollada hace uso de FaaS y BaaS. El primer caso a través de AWS Lambda y, el segundo, por medio de Amazon Simple Storage Service (S3), ambos servicios desarrollados por Amazon.

AWS Lambda permite crear funciones que contienen parte de la lógica de la aplicación. Cuando una función Lambda es invocada por un evento, se crea automáticamente un contenedor de acuerdo a la configuración proporcionada. Este contenedor puede reutilizarse para invocaciones posteriores, sin embargo, a medida que la demanda disminuye, el contenedor es dado de baja automáticamente [3].

AWS Lambda permite pagar únicamente por la ejecución de la función. De esta manera, se evita pagar innecesariamente por recursos ociosos. Además, se encarga de asegurar la escalabilidad de estas funciones, es decir, si el tráfico aumenta, escala automáticamente aplicando los recursos requeridos y, cuando el tráfico disminuye, actúa a la inversa. Esto sortea la necesidad de escalar la aplicación completa en ocasiones donde una función concentra la mayor parte del tráfico o del esfuerzo computacional.

S3 es una tecnología *serverless* BaaS, incluso aunque exista desde antes de que el término *serverless* comenzara a usarse [12]. Es un servicio de almacenamiento de recursos estáticos como sitios web HTML o archivos multimedia. Puede utilizarse de tal forma que cuando se cree o elimine un objeto en un bucket, el sistema de notificaciones de eventos propio de S3 dirija los eventos a AWS Lambda. Un ejemplo de uso podría ser un sistema de procesamiento de datos en tiempo real, en el cual se busque disparar automáticamente la ejecución de una función lambda encargada de crear la miniatura de una imagen tras la carga de la imagen original en un bucket.

3 Implementación

La implementación generada consiste en una aplicación web serverless que permite realizar operaciones aritméticas básicas: suma, resta, multiplicación y división. La misma se considera completamente *serverless* debido a que tanto el contenido estático como el *backend* están soportados por servicios Amazon que no requieren de un servidor administrado por el programador para operar. Los servicios de Amazon utilizados en este proyecto son: Amazon Simple Storage Service (S3), Amazon API Gateway y AWS Lambda.

Fig. 2. Arquitectura de la aplicación web *serverless* implementada

El contenido estático de la aplicación web (código HTML, CSS y JavaScript) se aloja en un *bucket* de S3, cuyos permisos de sólo lectura garantizan que cualquier usuario pueda acceder de manera directa a los recursos. El punto de acceso a estos recursos es la URL provista por S3 debido a que no se cuenta con un dominio adquirido.
Para proveer la lógica de negocio se desarrolló un conjunto de funciones Lambda en lenguaje Python. El usuario accede a las funciones Lambda por medio de solicitudes que viajan a través de una API REST desarrollada en Amazon API Gateway, la cual permite manejar el balanceo de carga y el almacenamiento en caché.
En cuanto a la definición de la API, se definieron las rutas y los *endpoints* como parte de la configuración. Cada ruta se asoció a la función Lambda que correspondía en cada caso, tal como se describe en Fig. 3.

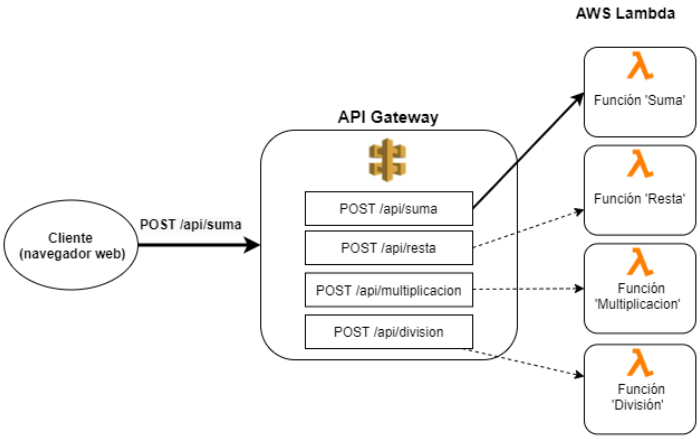


Fig. 3. Mapeo de solicitudes HTTP a funciones Lambda

Esta API recibe solicitudes HTTP, enruta esas solicitudes a una función Lambda que depende de la ruta indicada en la solicitud HTTP y de la configuración de enrutamiento definida, obtiene la respuesta de la función y finalmente devuelve la respuesta al cliente original. Por ejemplo, al recibir una solicitud HTTP "POST /api/suma" invoca a la función Lambda "Suma" correspondiente. En esta implementación, se hizo uso de solicitudes POST para enviar los operandos desde el navegador web para su procesamiento en la función y el posterior retorno del resultado de la operación.

La función Lambda se desarrolló de tal forma que utiliza el cuerpo de la solicitud HTTP, incluida en un objeto JSON, para obtener los operandos de la operación. Luego realiza la operación aritmética y devuelve el resultado en el cuerpo de la respuesta a la solicitud.

Para realizar la implementación de la práctica DevOps de despliegue continuo en esta aplicación, se hizo uso de la herramienta Travis CI.

Cuando los desarrolladores realizan la confirmación del código en GitHub en la rama principal, se dispara la ejecución de Travis CI. Travis CI es una plataforma de integración continua que respalda el proceso de desarrollo mediante la realización de pruebas, creación de entregables y gestión de despliegues de forma automática ante cambios en el código [14].

Además, a fin de mantener seguro el acceso al ambiente de producción, se crearon dos usuarios, uno que permite el despliegue del contenido estático y otro permite el despliegue de las funciones Lamba. La creación de los usuarios se hizo a través del servicio IAM (Identity and Access Management), el cual permite crear usuarios y asignarles diferentes permisos para conceder o negar el acceso a los recursos de AWS. Los dos usuarios creados fueron configurados con los permisos necesarios para realizar el despliegue continuo.

Para lograr el despliegue continuo a través de Travis CI se crearon dos repositorios públicos en GitHub vinculados a Travis CI, uno para el contenido estático de la aplicación web y el otro para alojar las funciones Lamba desarrolladas. Al vincular los repositorios se

añadió entre las variables de entorno las credenciales del usuario correspondiente a cada uno.

En ambos repositorios se incluyó un fichero YAML con todas las configuraciones necesarias para que cada vez que se haga *merge* o *push* en la rama principal del repositorio en el que se esté trabajando, el código fuente de la rama sea desplegado en producción. En la Fig. 4. se muestra una captura de la pantalla visualizada en Travis CI tras realizar el despliegue exitoso de una función Lambda.

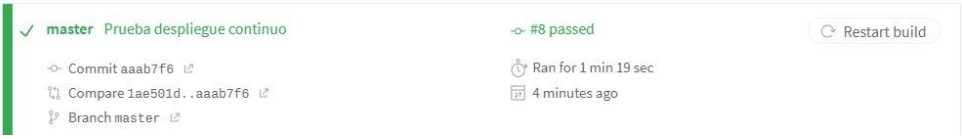


Fig. 4. Captura de la pantalla de Travis CI tras realizar el despliegue exitoso de una función Lambda

4 Análisis de la Implementación

A partir de una revisión bibliográfica sobre los beneficios y desventajas propias de implementar despliegue continuo y de desarrollar aplicaciones *serverless* [3, 7, 9–13, 15, 16], se analizó la permanencia de los mismos en la implementación generada que combina ambas prácticas. El resultado del análisis de beneficios se presenta en la Tabla 1. Finalmente se expone el comportamiento de las limitaciones propias de las aplicaciones *serverless* en la nueva implementación.

Tabla 1. Análisis del impacto de la implementación generada en los beneficios ofrecidos por el despliegue continuo y las aplicaciones *serverless*.

Nº	Beneficio	Comportamiento en la implementación generada (Disminuye, Se mantiene, Aumenta)	Justificación
1	Acuerdos persistentes en el proceso de despliegue	Se mantiene	La implementación de despliegue automático define cómo y cuándo se va a realizar el pase a producción. Ese acuerdo sobre el proceso de pase a producción tiene una persistencia prolongada en el tiempo que va más allá de un único uso. En este caso se definió que, tras realizar un *merge* o un

			push en la rama principal, Travis CI realiza automáticamente el despliegue.
2	Reduce la necesidad de personal dedicado al despliegue	Se mantiene	Al automatizarse el despliegue, se reduce la necesidad de consumir horas del equipo en realizar dicha tarea y permite que los desarrolladores puedan enfocarse en realizar otras acciones como: desarrollar nuevas funcionalidades, ejecutar más cantidad de casos de prueba, adoptar buenas prácticas de programación, etc.
3	Aumenta la probabilidad de mejorar la frecuencia de los despliegues	Aumenta	Cuando el despliegue requiere mucho tiempo y tiene numerosos riesgos asociados, sin mencionar la frustración cada vez que se realiza erróneamente, se tiende a posponer el despliegue tanto como sea posible. Automatizar esta tarea, permite que el lanzamiento del producto se efectúe en minutos, aumentando su frecuencia y reduciendo el tiempo transcurrido entre despliegues. Esto conlleva una reducción considerable en el tiempo que se tarda en lanzar un nuevo producto o servicio al mercado. Adicionalmente, la granularidad fina ofrecida por una aplicación *serverless* con funciones Lambda, facilita la realización de cambios en las funciones y pruebas de dichos cambios, fomentando así la reducción de tiempo entre despliegues.
4	Aumenta la probabilidad de mejorar la calidad del producto	Aumenta	De acuerdo a lo mencionado en el beneficio 2, esta práctica DevOps permite invertir más tiempo del equipo de desarrollo en otras tareas que probablemente aumenten la calidad del software elaborado. Adicionalmente, una aplicación *serverless* permite que los desarrolladores no deban preocuparse por la infraestructura subyacente ni por escribir código que deba gestionar la escalabilidad de la aplicación, ya que de esas tareas se encargan los servicios contratados. Por otro lado, el aumento en la frecuencia entre despliegues (beneficio 3) implica un aumento en la frecuencia de las pruebas de aceptación de usuarios, lo que permite obtener una retroalimentación temprana de la visión del cliente respecto al producto integral elaborado, y asegurar que el producto entregado cumplirá con sus expectativas. Por los motivos expuestos, se puede concluir que la implementación que se

			presenta aumenta la probabilidad de obtener este beneficio.
5	Aumenta la probabilidad de hacer entregas más pequeñas	Aumenta	Aumentar la frecuencia de los despliegues, implica pasar a producción únicamente las funcionalidades desarrolladas y probadas en un período de tiempo más corto. Por lo tanto, la cantidad de funcionalidades incluidas en una entrega disminuye. Por el motivo anteriormente expuesto, como la aplicación desarrollada en este trabajo potencia el beneficio 3, también potencia a este beneficio.
6	Reduce el riesgo asociado a un despliegue	Aumenta	De acuerdo a lo indicado en 5, la cantidad de código modificado disminuye entre despliegues. Por lo tanto, se reduce el riesgo asociado a publicar una implementación en particular. Además, se espera que la automatización garantice la consistencia y mejore la precisión del despliegue, minimizando los errores humanos que pudieran surgir en el proceso. Por los motivos anteriormente expuestos, como la aplicación desarrollada en este trabajo potencia el beneficio 5, también potencia este beneficio.
7	Aumenta la probabilidad de reducir el tiempo requerido para corregir errores detectados en producción.	Aumenta	Mientras menos cantidad de modificaciones implique un despliegue, más probabilidad existe de que el desarrollador detecte rápidamente la causa del error y logre corregirlo antes de que se produzcan daños mayores. Por el motivo anteriormente expuesto, como la aplicación desarrollada en este trabajo potencia el beneficio 5, también potencia este beneficio.
8	Facilita los retrocesos de versiones en caso de detectarse un error en producción	Se mantiene	La publicación frecuente de pequeñas modificaciones permite que, en caso de que se detecte un fallo en producción, se pueda volver a la versión anterior (roll-back) del sistema. En la implementación realizada, el retroceso es además facilitado por el control de versiones del producto completo ofrecido por GitHub.
9	Escalabilidad de la aplicación	Se mantiene	Los proveedores de servicios, en este caso AWS Lambda, permiten escalar la aplicación rápidamente ante aumentos de tráfico, evitando la gran inversión

			de tiempo y dinero que suelen requerir las aplicaciones monolíticas.
10	Reducción de costos	Se mantiene	AWS Lambda cobra sus servicios de acuerdo al tiempo de ejecución requerido en el uso de las funciones. De esta manera, sólo se paga por servicios utilizados, evitando el gasto en servidores ociosos.

A su vez, es importante mencionar que existen ciertas desventajas asociadas al uso de aplicaciones *serverless*, que aplican a esta implementación y que no se han logrado reducir. Por un lado, la implementación del sistema se vuelve dependiente de las características propias del servicio ofrecido y dificulta el cambio de proveedor. Por otro lado, implica un consenso inicial respecto del criterio de división de código en funciones Lambda y una evaluación frecuente de la decisión tomada para realizar ajustes en caso de ser necesario.

5 Conclusiones

A partir del trabajo realizado, es posible evaluar la implementación de despliegue continuo en una aplicación web *serverless*. La principal ventaja consiste en maximizar la probabilidad de mejorar la calidad del producto, ya que el objetivo último de todo desarrollo de software consiste en producir un sistema que cumpla con los requisitos o expectativas del cliente. A su vez, en el desafortunado caso de que se origine un fallo en producción, esta implementación maximiza la probabilidad de reducir el tiempo requerido para corregir errores detectados y mantiene la facilidad de realizar un retroceso de versión en caso de que no se disponga de tiempo para analizar el fallo. Si tuviéramos que indicar algún aspecto susceptible de mejora, el mismo se relaciona con que no se logró reducir ni eliminar las desventajas asociadas a la computación *serverless*, pero se logró mantener y optimizar los beneficios ofrecidos por estas prácticas de forma aislada.

Debido a que los resultados de esta implementación fueron positivos, estudios futuros se van a enfocar en adicionar a esta aplicación la práctica de integración continua que actualmente es ejecutada de forma manual. Automatizar la generación de paquetes de código, pruebas de integración y pruebas de aceptación de usuarios, permitiría aumentar aún más los beneficios ofrecidos por esta solución ya que la automatización de pruebas favorecería la ejecución de más casos de prueba, lo cual resultaría en una detección temprana de errores y en un aumento de calidad del código publicado en producción.

Referencias

1. Mishra A, Otaiwi Z (2020) DevOps and software quality: A systematic mapping. Comput Sci Rev 38:100308
2. Spillner J (2019) Serverless Computing and Cloud Function-based Applications. In: Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing Companion. pp 177–178
3. Bangera S (2018) DevOps for Serverless Applications: Design, deploy, and monitor your serverless applications using DevOps practices. Packt Publishing Ltd
4. Ivanov V, Smolander K (2018) Implementation of a DevOps pipeline for serverless applications. In: International Conference on Product-Focused Software Process Improvement. pp 48–64
5. Kim G, Humble J, Debois P, Willis J (2016) The DevOps Handbook:: How to Create World-Class Agility, Reliability, and Security in Technology Organizations. IT Revolution
6. Erich FMA, Amrit C, Daneva M (2017) A qualitative study of DevOps usage in practice. J Softw Evol Process 29:e1885
7. Bass L, Weber I, Zhu L (2015) DevOps: A software architect's perspective. Addison-Wesley Professional
8. ISO (2015) ISO 9001:2015 Quality management systems — Requirements
9. Mora Perez JJ (2015) DevOps y el camino de baldosas amarillas
10. Shahin M, Babar MA, Zhu L (2017) Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. IEEE Access 5:3909–3943
11. Baldini I, Castro P, Chang K, et al (2017) Serverless computing: Current trends and open problems. In: Research Advances in Cloud Computing. Springer, pp 1–20
12. Roberts M, Chapin J (2017) What Is Serverless? O'Reilly Media, Incorporated
13. Roberts M Serverless architectures. https://martinfowler.com/articles/serverless.html. Accessed 8 Oct 2020
14. Travis CI Core Concepts for Beginners. https://docs.travis-ci.com/user/for-beginners/#builds-stages-jobs-and-phases. Accessed 8 Oct 2020
15. Rossel S (2017) Continuous Integration, Delivery, and Deployment: Reliable and faster software releases with automating builds, tests, and deployment. Packt Publishing Ltd
16. Humble J, Farley D (2010) Continuous delivery: reliable software releases through build, test, and deployment automation. Pearson Education

Comparando un Sistema de Bases de Datos de Grafos con un Sistema de Bases de Datos Relacional

Silvina Migani {silvina.migani@gmail.com} Cristina Vera {civerados@gmail.com}
Departamento e Instituto de Informática, FCEFN-UNSJ

Resumen. Los Sistemas de Gestión de Bases de Datos NoSQL (SGBDs NoSQL) surgieron como una alternativa de solución a problemas no resueltos eficientemente por los tradicionales Sistemas de Gestión de Bases de Datos Relacionales (SGBDRs). El término NoSQL abarca un grupo de distintos tipos de SGBDs, que se distinguen fundamentalmente por el modelo de datos subyacente. Uno de ellos lo constituyen los SGBDs de grafos (SGBDGs), que se presentan como muy convenientes cuando las relaciones de los datos son más importantes que los datos en sí mismos. Este trabajo presenta un ejemplo de aplicación que permite materializar un análisis comparativo entre ambos paradigmas, utilizando PostgreSQL como representante de los sistemas tradicionales y Neo4j por parte de los SGBDGs.

Palabras claves: Bases de Datos - Bases de Datos basadas en Grafos vs Bases de Datos Relacionales – Neo4j vs PostgreSQL

1 Introducción

Los sistemas de gestión de bases de datos relacionales han sido usados ampliamente y con óptimos resultados hace ya muchas décadas. Sin embargo, la continua evolución de la tecnología generó cambios, que en muchas ocasiones no han podido ser abordados adecuadamente por los populares y robustos sistemas de bases de datos relacionales [1] [2]. Como solución a esta problemática surgieron los llamados sistemas de bases de datos NoSQL [2]. Este término agrupa varios tipos de sistemas de bases de datos no relacionales cuya diferencia principal es el modelo de datos subyacente. Las categorías más salientes son los sistemas basados en documentos, columnas, clave-valor y grafos [3]. Éstos últimos prometen ser convenientes para gestionar datos altamente conectados de manera más natural y eficiente que los sistemas de bases de datos tradicionales.

El presente trabajo expone un ejemplo de aplicación implementado en Neo4J Desktop Versión 1.3.4 y en PostgreSQL Version 12, con el objetivo de experimentar y comparar ambos paradigmas. El ejemplo desarrollado se caracteriza por contener datos altamente vinculados, contexto teóricamente ideal para usar sistemas de bases de datos de grafos.

2 Marco Teórico

2.1 Bases de Datos basadas en grafos

Los SGBDGs [4] son sistemas específicamente diseñados para administrar datos estructurados como grafos. Es decir, su modelo de datos lógico es alguna de las variantes de la definición matemática de grafo [5] [7], como por ejemplo: grafos dirigidos, no dirigidos, con vértices y arcos etiquetados o no etiquetados, con propiedades en nodos y arcos, hipergrafos e hipernodos. Las operaciones CRUD (Create-Read-Update-Delete) [5] [6] trabajan sobre grafos y sus elementos (nodos, relaciones, propiedades de nodos y propiedades de los nodos); sin embargo, esto no significa que en su nivel interno, almacenen grafos. Algunos utilizan almacenamiento nativo [4], vale decir, sus estructuras físicas son grafos, pero otros, mapean los grafos a otras estructuras físicas [4], como por ejemplo tablas u objetos. En cuanto a la implementación de las conexiones entre nodos, también se presentan diferentes enfoques, algunos eligen la adyacencia sin índices, y otros no.

Asimismo como cualquier SGBD, además de proveer el almacenamiento persistente, satisfacen las propiedades básicas [1] relativas a integridad, consistencia, control de concurrencia, recuperación de datos ante fallos, etc.

Hoy en día el mercado cuenta con numerosos productos de software que implementan esta tecnología. Algunos de ellos, son Neo4j, Microsoft Azure Cosmos DB, ArangoDB, OrientDB y Virtuoso; enumerados en orden decreciente de popularidad según [9].

3 Trabajo Realizado

Este trabajo presenta un ejemplo de aplicación implementado en PostgreSQL y en Neo4j, como representantes de los SGBDs tradicionales y de grafos respectivamente. El ejemplo elegido fue motivado por la situación vivida respecto a la aparición de la pandemia de COVID19 en San Juan. Una vez que empezaron a surgir los casos positivos, el gobierno de la provincia trató de encontrar el nexo epidemiológico para conocer el tipo de circulación del virus, las entidades públicas o privadas frecuentadas por personas con la enfermedad, entre otro tipo de información afín a la problemática. Es importante mencionar que si bien la situación planteada es auténtica y conocida por todos, los datos y las consultas planteadas son ficticias, pero siempre tratando de mantener una correspondencia lo más cercana posible a la realidad.

A continuación se explica brevemente la experiencia realizada.

3.1 Modelado Conceptual

Antes de realizar la implementación en los gestores mencionados, se generó el modelo conceptual correspondiente a la porción del minimundo considerado.

- MODELO ENTIDAD RELACIÓN EXTENDIDO (ERE)

El modelo ERE correspondiente al ejemplo se expone abajo (Figura 1).

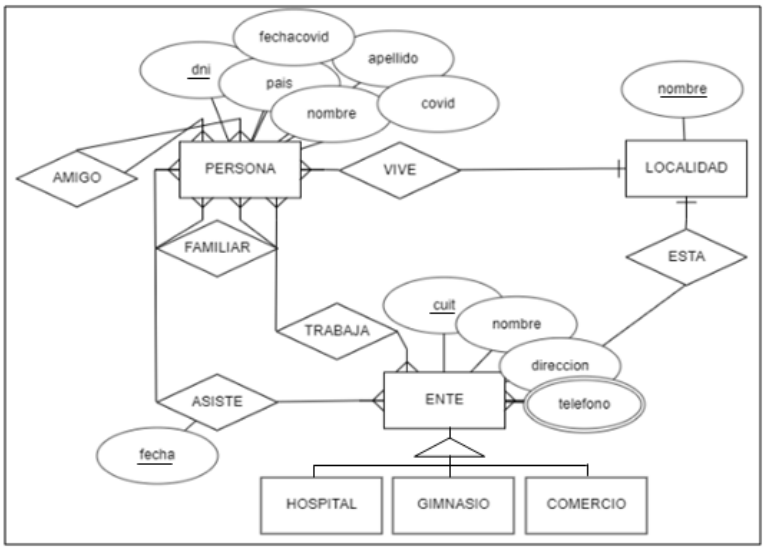


Figura 1: Modelo ERE COVID 2020SJ

La relación ASISTE mantiene la fecha en la que la persona fue a ese ente o entidad (pública o privada). Mientras que para la relación TRABAJA no se registra la fecha dado que se consideró que la persona concurre al ente todos los días hábiles.

- MODELO DE GRAFOS

El modelo de grafos de la Figura2 corresponde al resultado de la instrucción CALL db.schema.visualization() en Neo4j, luego de haber creado cada uno de los nodos y aristas. Com refiere a entes comercios, Gim a entes gimnasios, y Hospital a entes hospitales.

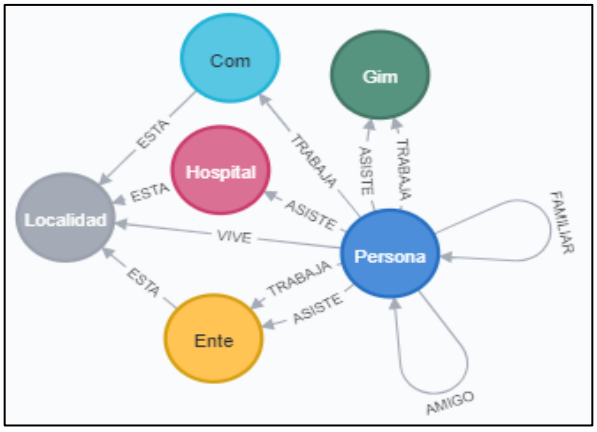


Figura2: Modelo de Grafos COVID 2020SJ

3.2 Creación de las Bases de Datos

Se implementó la base de datos relacional en PostgreSQL y la de grafos en Neo4j. Cabe mencionar, que las relaciones AMIGO y FAMILIAR están cargadas en un solo sentido en ambos motores.

- PostgreSQL

Se crearon las siguientes tablas: PERSONA (clave primaria {dni,pais}), LOCALIDAD, AMIGO, FAMILIAR, ENTE (Entidades), TELEFONOS, ASISTE y TRABAJA. También se insertaron algunas tuplas en todas las tablas.

- NEO4J

Se crearon los nodos y relaciones haciendo uso del lenguaje Cypher de Neo4j [8]. La base de datos completa se ve en la

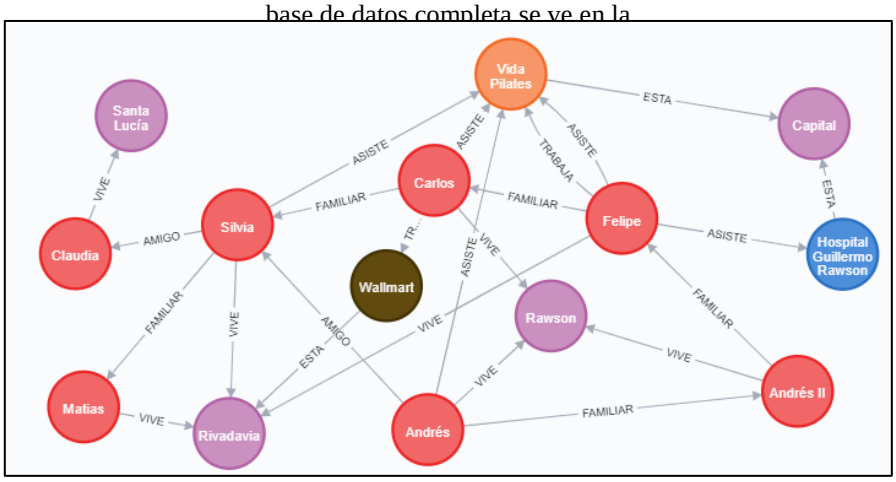


Figura 3.

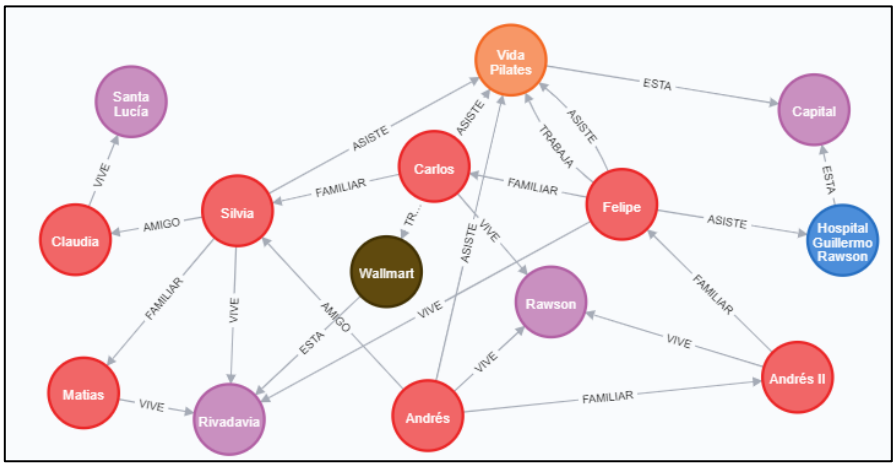


Figura 3: Base de Datos en Neo4J

3.3 Implementación de Consultas

Algunas de las consultas que se consideraron oportunas para el dominio de aplicación, se muestran a continuación. Se muestra el enunciado de la consulta y luego la implementación en ambos SGBDs (PostgreSQL y Neo4j)..

1. Personas (dni, pais, nombre, apellido y localidad) con COVID positivo junto a la localidad donde viven.

- POSTGRESQL

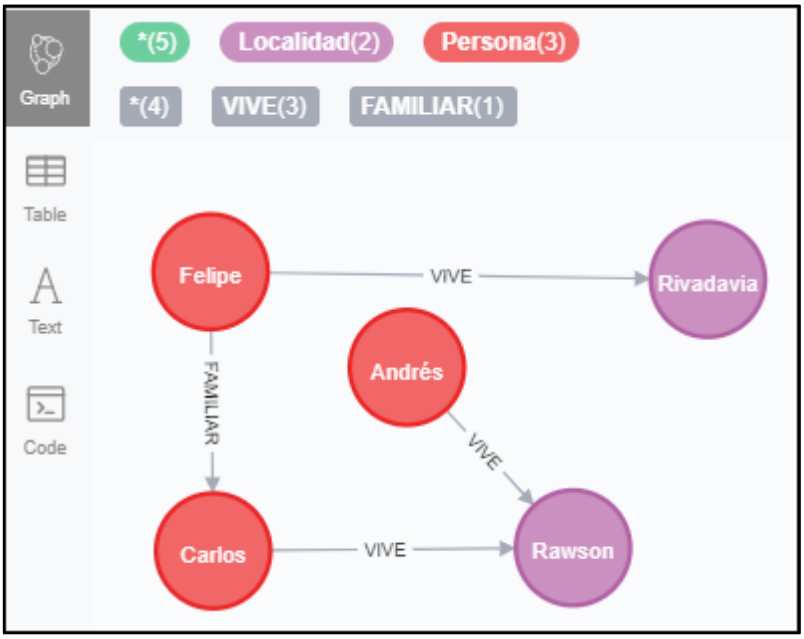
```
SELECT  dni, pais, nombre,  apellido, localidad
FROM  persona natural join localidad
WHERE covid='Positivo';
```

Data Output | Explain | Messages | Notifications

	dni integer	pais	nombre	apellido	localidad localidad
1	30900800	Argentina	Andrés	Tapia	(Rawson)
2	30900800	Chile	Carlos	Morales	(Rawson)
3	15900800	Argentina	Felipe	Morales	(Rivadavia)

- NEO4J

```
MATCH r=(a:Localidad)-[:VIVE]-(b:Persona {covid: 'Positivo'})
RETURN r
```



2. Contactos directos (dni, país, nombre y apellido) de Andrés con dni 30900800 y nacionalidad Argentina.
   - POSTGRESQL

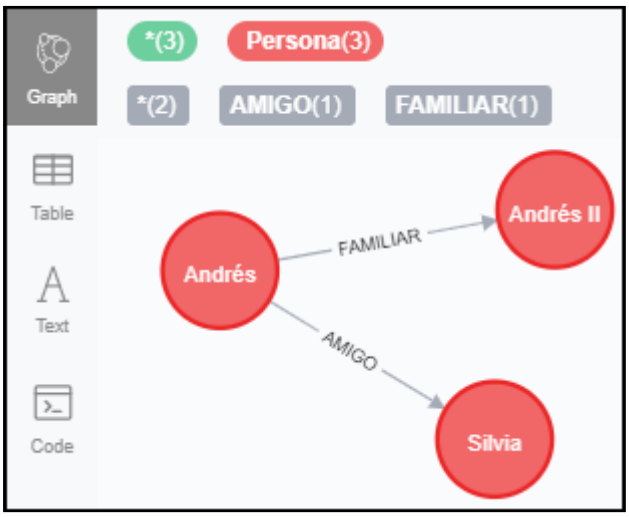
```
Select dni, pais, nombre, apellido
From Persona natural join(
Select dni2 dni,pais2 pais
From Amigo  Where dni1 = 30900800  AND pais1='Argentina'
UNION
Select dni1 dni,pais1 pais
From Amigo  Where dni2 = 30900800  AND pais2='Argentina'
UNION
Select dni2 dni,pais2 pais
From Familiar Where dni1 = 30900800  AND pais1='Argentina'
UNION
Select dni1 dni,pais1 pais
From Familiar Where dni2 = 30900800  AND pais2='Argentina')aux;
```

Data Output | Explain | Messages | Notificatio

	dni [PK] integer	pais [PK] chara	nombre character varyi	apellido character v
1	30900801	Chile	Silvia	Chiri
2	13900801	Argentina	Andrés II	Tapia

- NEO4J

```
MATCH r=(a:Persona)-[:AMIGO|:FAMILIAR]->(b:Persona)
WHERE a.dni = 30900800  AND  a.pais='Argentina'
RETURN r
```



3. Contactos (dni, país, nombre y apellido) directos e indirectos grado 2 de de Andrés con dni 30900800 y nacionalidad Argentina.

- POSTGRESQL

```
Select dni, pais, nombre, apellido
From Persona natural join(
Select dni2 dni, pais2 pais
From Familiar Where dni1 = 30900800  AND pais1='Argentina'
UNION
Select dni1 dni, pais1 pais
From Familiar Where dni2 = 30900800  AND pais2='Argentina'
UNION
Select F2.dni2 dni, F2.pais2 pais
From Familiar F  join Familiar F1 on (F.dni1=F1.dni1 and F.pais1=F1.pais1)
join Familiar F2 on (F1.dni2=F2.dni1 and F1.pais2=F2.pais1)
Where F.dni1 = 30900800  AND F.pais1='Argentina'
UNION
Select F2.dni1 dni, F2.pais1 pais
From Familiar F  join Familiar F1 on (F.dni1=F1.dni2 and F.pais1=F1.pais2)
join Familiar F2 on (F1.dni2=F2.dni1 and F1.pais2=F2.pais1)
Where F.dni1 = 30900800  AND F.pais1='Argentina'
UNION
```

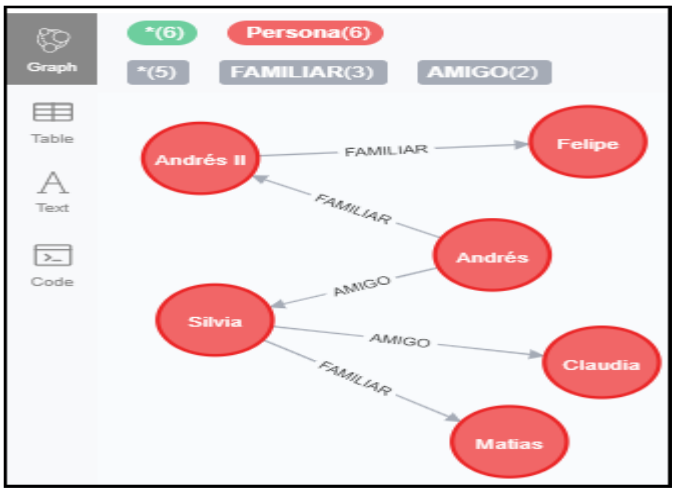
```
Select F2.dni1 dni, F2.pais1 pais
From Familiar F  join Familiar F1 on (F.dni2=F1.dni1 and F.pais2=F1.pais1)
join Familiar F2 on (F1.dni1=F2.dni2 and F1.pais1=F2.pais2)
Where F.dni2 = 30900800  AND F.pais2='Argentina'
UNION
Select F2.dni2 dni, F2.pais2 pais
From Familiar F  join Familiar F1 on (F.dni2=F1.dni2 and F.pais2=F1.pais2)
join Familiar F2 on (F1.dni2=F2.dni1 and F1.pais2=F2.pais1)
Where F.dni2 = 30900800  AND F.pais2='Argentina'
union
Select dni2 dni, pais2 pais
From amigo Where dni1 = 30900800  AND pais1='Argentina'
UNION
Select dni1 dni, pais1 pais
From amigo  Where dni2 = 30900800  AND pais2='Argentina'
UNION
Select F2.dni2 dni, F2.pais2 pais
From amigo F  join amigo F1 on (F.dni1=F1.dni1 and F.pais1=F1.pais1)
join amigo F2 on (F1.dni2=F2.dni1 and F1.pais2=F2.pais1)
Where F.dni1 = 30900800  AND F.pais1='Argentina'
UNION
Select F2.dni1 dni, F2.pais1 pais
From amigo F  join amigo F1 on (F.dni1=F1.dni2 and F.pais1=F1.pais2)
join amigo F2 on (F1.dni2=F2.dni1 and F1.pais2=F2.pais1)
Where F.dni1 = 30900800  AND F.pais1='Argentina'
UNION
Select F2.dni1 dni, F2.pais1 pais
From amigo F  join amigo F1 on (F.dni2=F1.dni1 and F.pais2=F1.pais1)
join amigo F2 on (F1.dni1=F2.dni2 and F1.pais1=F2.pais2)
Where F.dni2 = 30900800  AND F.pais2='Argentina'
UNION
Select F2.dni2 dni, F2.pais2 pais
From amigo F  join amigo F1 on (F.dni2=F1.dni2 and F.pais2=F1.pais2)
join amigo F2 on (F1.dni2=F2.dni1 and F1.pais2=F2.pais1)
Where F.dni2 = 30900800  AND F.pais2='Argentina')aux;
```

Data Output | Explain | Messages | Notifications

	dni [PK] integer	pais [PK] character	nombre character var	apellido character va
1	30900801	Chile	Silvia	Chiri
2	13900801	Argentina	Andrés Il	Tapia
3	15900800	Argentina	Felipe	Morales
4	18000000	Argentina	Claudia	Lopez

- NEO4J

```
MATCH r=(a:Persona)-[:FAMILIAR*1..2]->(b:Persona)
WHERE a.dni = 30900800   AND  a.pais='Argentina'
RETURN r
```



4. Contactos directos e indirectos (todos) de Andrés con dni 30900800 y nacionalidad Argentina.

- POSTGRESQL (consulta recursiva)

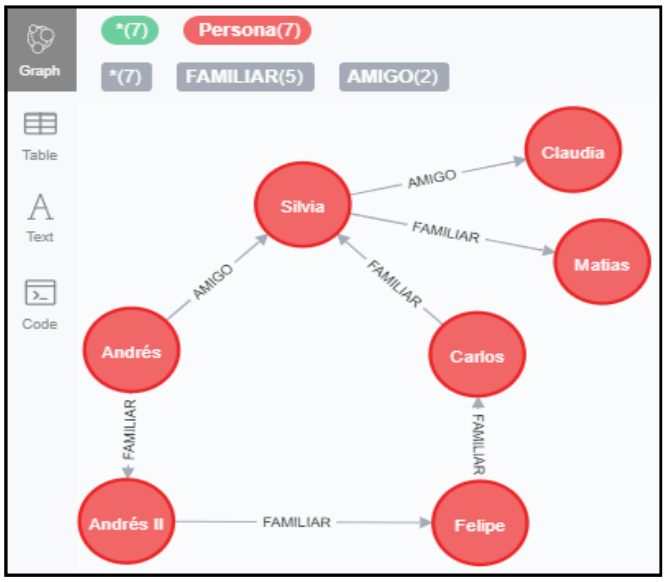
```
SELECT dni2, pais2 from (
SELECT * FROM
((WITH RECURSIVE PARIENTES AS (
  SELECT *
  FROM  familiar
  WHERE   dni1 = 30900800  AND pais1='Argentina'
  UNION ALL
  SELECT familiar.*
  FROM familiar JOIN PARIENTES
   ON (familiar.dni1 = PARIENTES.dni2
          and familiar.pais1 = PARIENTES.pais2))
 SELECT *
FROM PARIENTES)
UNION
(WITH RECURSIVE AMIGOS AS (
  SELECT *
  FROM  amigo
  WHERE   dni1 = 30900800  AND pais1='Argentina'
```

```
  UNION ALL
  SELECT amigo.*
  FROM amigo JOIN AMIGOS
   ON (amigo.dni1 = AMIGOS.dni2
          and amigo.pais1 = AMIGOS.pais2))
 SELECT *
FROM AMIGOS)) AUX) AUXI;
```

	dni2 integer	pais2 character varyi
1	13900801	Argentina
2	30900801	Chile
3	18000000	Argentina
4	30900800	Chile
5	30900801	Chile
6	15900800	Argentina
7	40900800	Argentina

- NEO4J

```
MATCH r=(a:Persona{dni:30900800,pais:'Argentina'})-[*]->(b:Persona)
RETURN r
```



5. Localidades donde viven casos positivos, mostrando el nombre de la localidad, la cantidad de casos y el dni de los infectados.

- POSTGRESQL

```
SELECT l.nombreloc,COUNT(*)
FROM persona NATURAL JOIN localidad l WHERE covid='Positivo'
GROUP BY l.nombreloc;
```

Data Output | Explain | Messages | Notifications

	nombreloc [PK] character varying (40)	count bigint
1	Rawson	2
2	Rivadavia	1

Cabe hacer notar, que la consulta 5 implementada arriba, muestra la cantidad de casos positivos por localidad, pero no la lista de los dni correspondiente a los infectados. Esto se debe a que cuando en SQL se agrupo por un atributo, el grupo queda resumido a una tupla, y por ende, en este caso, resulta imposible mostrar los dni,pais correspondientes.
Abajo, se muestra otra implementación haciendo uso de una función que permite generar un JSON con la lista de los dni,pais solicitados.

Data Output | Explain | Messages | Notifications

	nombreloc [PK] character var	cantidad bigint	infectados json
1	Rawson	2	[{"dni" : 30900800, "pais" : "Argentina"}, {"dni" : 30900800, "pais" : "Chile"}]
2	Rivadavia	1	[{"dni" : 15900800, "pais" : "Argentina"}]

- NEO4J

```
MATCH (a:Localidad)<-[:VIVE]-(b:Persona {covid: 'Positivo'})
WITH  a, count(b) AS cantidad, collect({dni:b.dni,pais:b.pais}) AS infectados
WHERE cantidad >0
RETURN a.nombre, cantidad, infectados
```

Table | Text | Code

"a.nombre"	"cantidad"	"infectados"
"Rawson"	2	[{"dni":30900800,"pais":"Argentina"},{"dni":30900800,"pais":"Chile"}]
"Rivadavia"	1	[{"dni":15900800,"pais":"Argentina"}]

6. Localidad donde viven más casos positivos, mostrando el nombre de la localidad, la cantidad de casos y el dni de los infectados.

- POSTGRESQL

```
SELECT l.nombreloc,COUNT(*) cantidad,
JSON_AGG(JSON_BUILD_OBJECT('dni',dni,'pais',pais)) AS infectados
FROM persona NATURAL JOIN localidad l
WHERE covid='Positivo'
GROUP BY l.nombreloc
ORDER BY cantidad DESC
LIMIT 1;
```

Data Output | Explain | Messages | Notifications

	nombreloc [PK] character varying (40)	cantidad bigint	infectados json
1	Rawson	2	[{"dni" : 30900800, "pais" : "Argentina"}, {"dni" : 30900800, "pais" : "Chile"}]

- NEO4J

```
MATCH (a:Localidad)<-[:VIVE]-(b:Persona {covid: 'Positivo'})
WITH  a, count(b) AS cantidad, collect({dni:b.dni,pais:b.pais}) AS infectados
WHERE cantidad >0
RETURN a.nombre,cantidad, infectados
ORDER BY cantidad desc
LIMIT 1
```

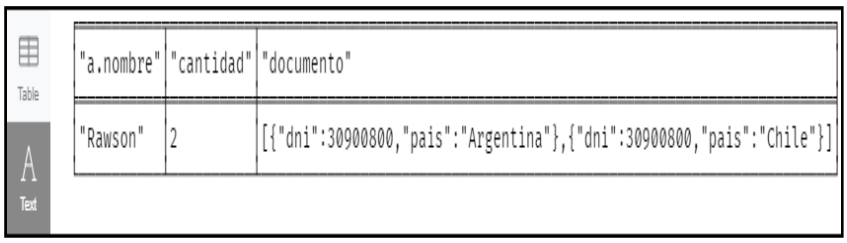


Table | Text

"a.nombre"	"cantidad"	"documento"
"Rawson"	2	[{"dni":30900800,"pais":"Argentina"},{"dni":30900800,"pais":"Chile"}]

7. Todas las personas que podrían tener nexo epidemiológico con la persona dni 30900800 de Argentina.

- POSTGRESQL

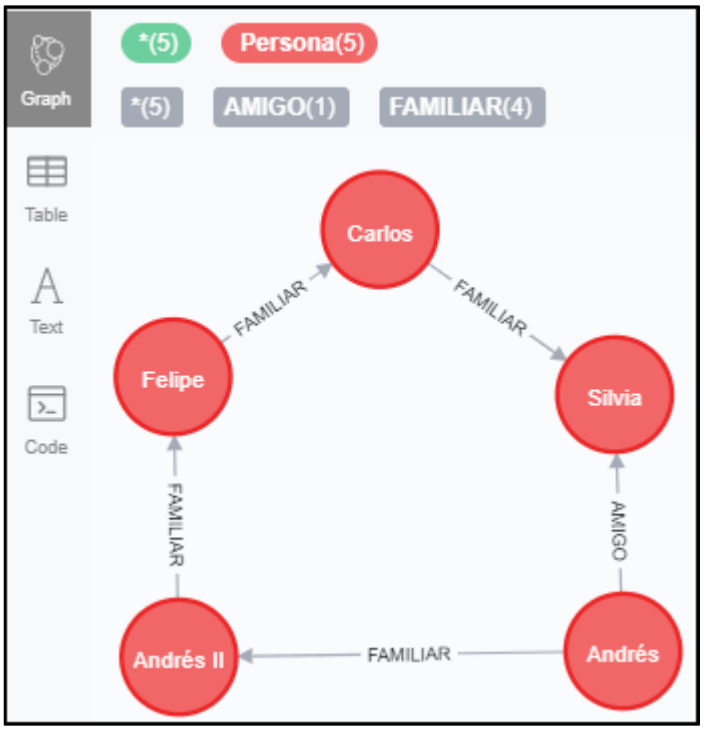
```
SELECT dni,pais,nombre, apellido
FROM Persona
NATURAL JOIN (SELECT dni2 dni,pais2 pais FROM Amigo A
WHERE dni1 = 30900800  AND pais1='Argentina'
UNION
SELECT dni1 dni,pais1 pais FROM Amigo
WHERE dni2 = 30900800  AND pais2='Argentina'
UNION
SELECT dni2 dni,pais2 pais FROM Familiar
WHERE dni1 = 30900800  AND pais1='Argentina'
UNION
SELECT dni1 dni,pais1 pais FROM Familiar
WHERE dni2 = 30900800  AND pais2='Argentina'
UNION
SELECT A2.dni,A2.pais FROM Trabaja A1 JOIN Trabaja A2 ON (A1.cuit=A2.cuit)
WHERE A1.dni = 30900800  AND A1.pais='Argentina' AND (A2.dni=A1.dni AND
A2.pais <> A1.pais OR A2.dni<>A1.dni )
UNION
SELECT A1.dni,A1.pais FROM Trabaja A1 JOIN Trabaja A2 ON (A1.cuit=A2.cuit)
WHERE A2.dni = 30900800  AND A2.pais='Argentina' AND (A2.dni=A1.dni AND
A2.pais<>A1.pais OR  A2.dni<>A1.dni )
UNION
SELECT A2.dni,A2.pais FROM Asiste A1 JOIN Asiste A2 ON (A1.cuit=A2.cuit)
WHERE A1.fecha=A2.fecha AND A1.dni = 30900800  AND A1.pais='Argentina' AND
(A2.dni=A1.dni AND A2.pais <> A1.pais OR A2.dni<>A1.dni )
UNION
SELECT A1.dni,A1.pais FROM Asiste A1 JOIN Asiste A2 ON (A1.cuit=A2.cuit)
WHERE A1.fecha=A2.fecha AND A2.dni = 30900800  AND A2.pais='Argentina' AND
(A2.dni=A1.dni AND A2.pais<>A1.pais OR A2.dni<>A1.dni ))aux;
```

Data Output | Explain | Messages | Notificatior

	dni [PK] integer	pais [PK] charact	nombre character va	apellido character va
1	30900800	Chile	Carlos	Morales
2	30900801	Chile	Silvia	Chiri
3	13900801	Argentina	Andrés II	Tapia
4	15900800	Argentina	Felipe	Morales

- NEO4J

```
MATCH (a:Persona{dni:30900800,pais:'Argentina'})-[:FAMILIAR|:AMIGO]-(b:Persona)
RETURN a as p, b as contactos
UNION
MATCH(a:Persona{dni:30900800,pais:'Argentina'})-[r:TRABAJA]-(:Ente)<-[s:TRABAJA]
-(b:Persona)
 RETURN a as p, b as contactos
UNION
MATCH      (a:Persona{dni:30900800,pais:'Argentina'})-[r:ASISTE]->(:Ente)<-[s:ASISTE]
-(b:Persona)  WHERE r.fecha=s.fecha
RETURN a as p, b as contactos
```



4 Conclusiones

La experiencia realizada permitió advertir, en forma preliminar, las virtudes del lenguaje Cypher de Neo4j. Es un lenguaje declarativo, potente e intuitivo, que hace muy sencilla la tarea de codificación, fundamentalmente cuando se necesitan recorrer vinculaciones.

Dentro de las ventajas de Neo4j respecto a PostgreSQL, se mencionan:

- Flexibilidad de las Estructuras: En Neo4j se agregó la propiedad covid y fechaCovid en forma dinámica, y no a todos los nodos persona, solamente a los que fue necesario. Al contrario, en PostgreSQL se debieron definir ambas columnas en la tabla Persona y sólo las tuplas correspondientes a casos positivos tuvieron valores no nulos.
- Visualización de los datos. Poder ver los datos como grafos, permite entender más rápida y claramente las relaciones entre ellos.
- Simplicidad, Claridad y Expresividad del Lenguaje: Como se puede notar, la implementación con Chyper de varias consultas implementadas, es mucho mas clara, simple y potente que en SQL de PostgreSQL. Esto se magnifica cuando se necesita navegar por varias relaciones, y a cualquier nivel de profundidad. Capacidad no ofrecida por todos los sistemas relacionales. Y en aquellos que lo soportan, resultan códigos más complejos y con un consumo alto de recursos.

Este trabajo debiera ser continuado para arribar a conclusiones definitivas e incluir dentro de la comparación, los tiempos de respuesta de las consultas realizadas.

Referencias

[1] Silberschatz, Korth, Sudarshan (2011). Database System Concepts. Sexta Edicion.
[2] Bhogal, Choksi (2015).Handling Big Data using NoSQL.
[3] Cattell, R. (2011). Scalable SQL and NoSQL data stores. ACM Sigmod Record, 39(4), 12-27.
[4] Angles, Gutierrez (2008). Survey of graph database models.
[5] Renzo Angles (2018). The Property Graph Database Model.
[6] Angles, Gutierrez (2017). An introduction to Graph Data Management.
[7] Ghrab, Romero, Skhiri, Vaisman, Zimányi (2016). GRAD: On Graph Database Modeling.
[8] Robinson, Webber, Eifrem (2013). Graph Databases.
[9] DB-Engines Ranking of Graph DBMS. https://db-engines.com/en/ranking/graph+dbms. Ultima visita (1/10/2020).

A maturity model for the Integrated Agile Transformation Model™

Fernando Pinciroli

Research Institute, Champagnat University, Mendoza, Argentina
pincirolifernando@uch.edu.ar

Abstract. Almost twenty years after the Agile Manifesto has appeared, it is more than clear that agility should not be limited only to the field of software development, but must reach all areas of organizations. Our professional activity led us to learn about different approaches to achieve this: Enterprise Scrum, SAFe and the Heart of Agile. But the most important experience we had was with the Integrated Agile Transformation Model™. Along the use of this model, we encountered a series of opportunities for improvement, among which is the maturity model that we propose in this work. As we describe in this article, it is important to have a structured framework that ensures success not only in the implementation of IATM but, above all, in the objectives it pursues, which consists of the agile transformation of organizations, directing them towards a process of continuous improvement in Agile.

Keywords: agile transformation; scaling agile; agnostic agile; Integrated Agile Transformation Model.

1 Introduction

Almost twenty years after the Agile Manifesto has appeared, it is more than clear that agility should not be limited only to the field of software development, but must reach all areas of organizations. Thus, the approach offered by "Management 3.0" [1] is a starting point to advance on the organization completely.

In particular, our professional activity led us to learn about different approaches to achieve this: Mike Beedle, co-author of the Agile Manifesto, taught us the principles of Enterprise Scrum [2] and we also had to work with SAFe [3]. We also have an affinity to the "Heart of Agile" [4], both because of our close friendship with Alistair Cockburn, co-author of the Agile Manifesto and creator of this approach, as well as for having been the founders of the "Heart of Agile" community in Argentina.

But the most important experience we had was with IATM, the Integrated Agile Transformation Model [5]. Arie van Bennekum, its creator and another co-author of the Agile Manifesto, trained us on this model and we completed the certification. But the most extraordinary thing was being able to become part of their team of international coaches, having had the opportunity to apply it together with its creator in different companies in Europe.

During the use of this model, we encountered a series of opportunities for improvement, among which is the maturity model that we propose in this work. As we describe in this article, it is important to have a structured framework that ensures success not only in the implementation of IATM but, above all, in the objectives it pursues, which consists of the agile transformation of organizations, directing them towards a process of continuous improvement in Agile.

In sections 2, 3 and 4 we briefly describe IATM, going through the approach, the types of coaching and the roles it proposes, respectively. In section 5 we will focus on the assessment phase, for which we have developed a maturity model, which is described in section 6. Finally, in section 7, we provide the conclusions and future work.

2 The approach and its elements

IATM is based on an iterative-incremental process (Figure 1). Every iteration is called a "wave", and every wave is made up of different phases: "Assessment & foundations", "Synchronize", "Follow", "Lead" and "Hand over".

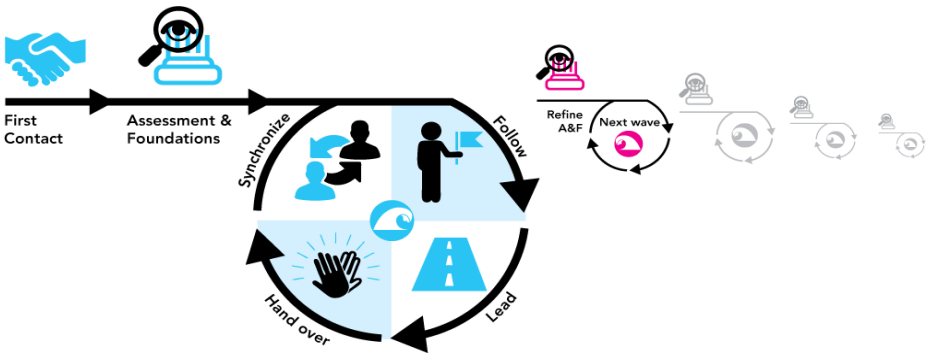


Figure 1: Iterations in IATM™.

The first phase, "Assessment & foundations", sets the basis for the whole process and is the moment to define the roadmap. It starts with the assessment of eight competences: "Agile supporting culture and management style", "Foundations content and process", "Create and sustain a participatory agile environment", "Guide teams to appropriate and matching solutions", "Build and maintain professional knowledge", "Agile mindset and attitude", "Supportive architecture" and "Development practices".

This assessment is conducted via interviews, document study, observations, walks, group sessions and management sessions. The assessment outcomes are the main input to define the roadmap, which is embodied by means of the transformation backlog. Every item of this backlog must be a prioritized SMART-style objective.

Finally, the transformation teams are organized, together with the working agreements and governance criteria.

The second phase, "Synchronize", aims to align all those involved in the agile transformation by training the teams (including a coach academy), creating team foundations, designing and implement team spaces and visuals, creating reporting process and visuals, planning team rituals.

In the "Follow" phase, the goal is to ensure the quality of the rituals, making the teams follow the coaches. During this phase, internal trainers and coaches are actively involved, and if works on discipline and quality of the rituals. Coaching is role modeling mostly on team level to get team rituals in place and tools identified and being used for the first time. Rhythms are installed. Management sessions on agile leadership and first community activities are initiated.

After starting the team activities, the phase "Lead" seeks to coach on the individual level, for instance, providing feedback in retrospective to individual professionals and the team's leads. Other coaching can be performed: on software engineering practices, on scaling, on management levels, oriented to the teams. Continuous learning and improvement, based on feedback afterwards is implemented.

The last phase, "Hand over", is the place to prepare the environment for the next wave and the future. All the internal coaches know what to do after this point. It is conducted a retrospective about the current wave, feeding the transformation backlog for the next wave. Transformation foundations are refined and, finally, success is celebrated.

3 Types of coaching

The approach foresees four types of coaching: management, organization, team and development coaching [5].

In the context of management coaching, the coach trains and guides managers on all current management levels in the organization. The focus is on aspects such as facilitating the teams and individuals, breaking current processes, overcoming existing paradigms and creating new agile paradigms in the organization, since managers are essential in the embedding of new paradigms and teams taking responsibilities.

During the organization coaching, the coach forms the bridge between people and teams who are already in transformation and those who are not. He helps overcome the standard gap during transformation between traditional ways of working (still present in the organization) and the agile stack implemented at the client.

In the team coaching, the coach works day to day with and in the teams to get the new way in the team in place. He is present in the team space and, especially during the first weeks of a wave, he demonstrates the rituals, gives both team and individuals feedback for team and individual refinement.

Finally, the development coach is able to take technical people by the hand and work together with them on agile practices, implementing pipelines for continuous delivery, continuous integration, etc. He helps them overcome hindering opinions and actually demonstrates new ways of working for developers and other technical people.

4 The snowman

The governance of the process is carried out by representatives of three domains: business, agile and technical, colored in black, blue and green, respectively, in Figure 2. The model was drawn with a shape that resembles a snowman, so it is colloquially called that way.

Figure 2: Roles in the governance model [5].

These are the roles present in the snowman:

- **Transformation sponsor:** owns the business case; ensures ongoing viability in line with the business case, funds and resources and decision-making process for escalations.
- **Business visionary:** owns wider implications in business change; defines, communicates and promotes business vision; monitors progress in line with the vision; contributes to requirements, design and review sessions; approves changes to high level requirements; ensures business resources and collaboration across business areas; promotes the vision into the working practice; arbitrates in disagreements between team members.
- **Agile transformation manager:** communicates with senior management and project governance; ensures high level planning; monitors progress against baselined plans; manages risks and issues, escalates when needed; manages overall project configuration and business involvement; motivates and

facilitates the team; handles problems escalating from the Solution Development Team and coaches this team.

- **Technical coordinator:** agrees and controls technical architecture; determines technical environment; identifies and owns related risks, escalates when needed; ensures non-functional requirements are achievable and met and adherence to technical standards and best practices; controls technical configuration of the solution; manages technical issues of go live; resolves technical disputes.
- **Solution Development Team:** is composed by the Agile Team Leader, Business Ambassador, Business Analyst, Solution Developer, Solution Tester, Business Advisor and Technical Advisor roles.

5 The assessment phase

The "Assessment & foundations" phase begins with an assessment in order to know the organization and the aspects that can favor the agile transformation or those that could be considered threats for this purpose. This is achieved through traditional interviews, documentation review, observations, and analysis of all the information obtained.

Eight competences are assessed and the results obtained and the actions to be taken based on them depend on the abilities of those who are conducting the agile transformation.

For this reason, we want to offer a maturity model to achieve several objectives:

1) Collaborate with those who carry out the transformation, facilitating their work.
2) To achieve a uniform criterion among the different consultants who carry out the assessment.
3) Identify recurring situations in different organizations, in order to enable the use of previously selected solutions for the same problems and, thus, enable the definition of transformation patterns.
4) Promote the continuous improvement of the evaluation criteria with all the opportunities that are detected with the successive applications of this maturity model.
5) Have a small roadmap to improve organizational skills throughout the successive transformation waves.

Arie van Bennekum has come up with a scheme [6] to make the assessment. The results obtained are allocated in the transformation backlog that will then be prioritized to guide the order in which each item will be addressed in the successive waves of transformation.

The transformation backlog items should also be written following the SMART model, with the required actions correctly defined and based on outcomes aligned to the organization goals. The backlog must be located on a central and visual place on the wall and presented to all stakeholders and interviewed persons.

6 The maturity model

Van Bennekum's assessment scheme was extended by us and used for the first time in the consultancy that we carried out together with Arie for an important company located in Sweden and Germany. Both, Arie's scheme and ours, are extensive, so it is not possible to include them in full in this work, so we present here the maturity model only for the first of the competencies of the first phase of IATM.

Those who wish to see the full maturity model and diagnostic schematics can find it here: [7].

Van Bennekum's diagnostic form is composed of three levels. We have added a fourth level in order to be more accurate with the assessment and to help those who carry out the transformation.

So, the maturity model that we elaborated for the first competence is as follows:

1. Agile supporting culture and management style				
	1.1. Management style			
	1.1.1. Practice of a delegating style of management or open to it			
#	**Practice**	**Level 1**	**Level 2**	**Level 3**
1	How does shared leadership coexist with the leadership of the functional structure?	There is a person or area is incorporated among the management and the operational levels that acts as a "buffer".	Management levels coordinate among themselves to facilitate cross-cutting processes.	The organization is completely horizontal.
2	Is there a person or area that acts as a buffer between the commands that go down the functional structure and the initiatives that come from the teams up?	This "buffer" belongs to the agile transformation team.	This "buffer" is recognized as an area of the organization.	This "buffer" are the same middle managers.
3	If the members of the teams must respond to different managers, how are the conflicting commands managed?	The managers agree with each other.	There are service level agreements.	Managers do not need to intervene and teams interact freely with each other.

4	How the participation and initiatives of the teams are encouraged?	The area responsible for agile transformation is in charge of promoting these initiatives.	Direct managers are in charge of promoting these initiatives.	These initiatives are promoted throughout the entire management structure of the organization.
5	A well-founded answer, whether yes or no, is given to the initiatives that arise from the teams?	Yes, but at the request of the teams.	Yes, always and without the need for the teams to request the answer.	Initiatives are prioritized in a backlog to continue to be considered in the future.

1.1.2. Contracting, management and monitoring on a team level or open to it

	Practice	**Level 1**	**Level 2**	**Level 3**
1	Are there metrics that track the performance of each team?	Yes, but kept within the scope of each team.	Yes, and they are public.	The organization stores them and offers a benchmark of the overall performance in the organization.
2	Is the variability of the performance of each equipment known?	Yes, but kept within the scope of each team.	Yes, and they are public.	The organization stores them and offers a benchmark of the overall variability in the organization.
3	In addition to the technical aspects, is the learning capacity or the adoption of new values of each team evaluated?	Only on some teams.	Yes, in all teams.	This practice is promoted by the organization.
4	Is the exchange of lessons learned among teams promoted?	Occasionally.	Yes, and they are public.	The organization coordinates their dissemination and provides the instruments so that they can be used.
5	Is the formation of groups promoted by specialties or roles among teams?	Only for specific technical specialties	Yes, for all roles	The organization is structured that way.

1.1.3. Management role models value and principles of the Agile Manifesto and Agile management or is open to it

Practice	**Level 1**	**Level 2**	**Level 3**

1	Is there an explicit plan to adopt principles and values and to improve them?	Yes, at the team level.	Yes, at the level of the entire IT area.	Yes, at the level of the entire organization.
2	The existing principles point to the provision of software or solutions?	Point to the provision of software and IT services.	Point to the provision of all kind of products and services.	Point to the provision of quality from the customer point of view.
3	Are the values and principles known only by technical teams or shared by business areas and different management layers?	At team level.	At IT area level.	At the entire organization level.
	1.2. Business participation			
	1.2.1. Analyzes organizational environment			
	Practice	**Level 1**	**Level 2**	**Level 3**
1	Are relationships between stakeholders analyzed?	Stakeholders are detected and profiled.	Variants are detected for each type of stakeholder.	Tools are used to monitor stakeholders.
2	Are stakeholder expectations analyzed?	Yes, but informally.	Tools such as “persona grid” are used.	Tools such as “user journey” are used.
3	Is the degree of influence of each stakeholder analyzed to understand how they should be kept informed or actively participating?	Yes, but informally.	Tools such as “stakeholder mapping” are used.	Tools are used to monitor stakeholders.
	1.2.2. Diagnoses client need			
	Practice	**Level 1**	**Level 2**	**Level 3**
1	Is there a distinction between business and software needs?	Yes, but business needs are managed informally.	Yes, there are two tracks: discovery and development tracks, with their own backlogs.	In addition, a distinction is made between business, user and software requirements.
2	Are empathy-based techniques used to understand client needs?	Yes, but informally.	Tools such as empathy maps are used.	Techniques such as design thinking are used.
	1.2.3. Creates appropriate designs to achieve intended outcomes			
	Practice	**Level 1**	**Level 2**	**Level 3**

1	Are the products defined based on client-oriented techniques?	The Product Owner has his own techniques to define the product.	Tools such as user journey, innovator compass are used.	There is a well-defined process for product design.
2	Is the risk or impact of customer needs analyzed?	Risks and impacts are only identified.	Tools such as “feedback grid”, “assumption grid”, “idea prioritization”, etc. are used.	Risks and opportunities are measured and managed.

1.2.4. Predefines a quality product and outcomes with client

	Practice	Level 1	Level 2	Level 3
1	The quality is understood as what the client perceives?	Yes, at the client level.	Quality is identified for client and customer.	Quality in use is managed.
2	Is it clear that what the client does not perceive as quality is waste?	Outcomes are identified.	Outcomes are prioritized and valued.	Outcomes are monitored after the solutions are deployed.
3	Is the difference between output and outcome understood?	Yes, it is.	There is an intention to minimize outputs and maximize outcomes.	Regular sessions are held to identify and assess the results.

1.3. Transparency

1.3.1. A clear scope and foundations for projects

	Practice	Level 1	Level 2	Level 3
1	Is the scope established with intermediate objectives or MVP?	Yes, it is.	An MVP roadmap is established and the vision is shared.	The MVP roadmap is periodically reviewed.
2	Is it clear that when one of these aspects: delivery date, scope and budget is fixed, the other two must be variable?	Yes, it is.	More emphasis is placed on iterative and incremental methods for fixed time challenges	A model like Cynefin is used to establish the way to approach work.

1.3.2. Used to work with rolling wave adaptive planning

1	Is there a positive feedback between product increment deliveries and the evolution of product design?	Yes, at the end of each iteration.	Also, through the heartbeats.	The results of the solutions are monitored after they are deployed.
2	Is there a "rolling forecast" concept?	Yes, it is.	The rolling forecast is reviewed at the end of each iteration.	All involved carry out empirical project management.

1.3.3. IATM is perceived as a fundament or open to it

1	What are the characteristics of IATM that are considered to be able to make a greater contribution?	IATM is perfectly understood within the agile transformation team.	IATM is perfectly understood throughout the whole IT area.	IATM is perfectly understood and fully accepted at all management levels of the organization.
	1.3.4. Use or are open to have high frequent business and stakeholder participation			
1	Is stakeholder participation established by joint agreement rules?	Work agreements are established and respected by all.	Service level agreements are established and respected by all.	All the stakeholders proactively participate in the periodic review of work agreements.
2	Are there retrospective sessions with the business and stakeholders?	Product Owners participate in team retrospectives.	Business-side roles participate in heartbeats.	There are retrospectives sessions dedicated to specific roles on the business side.
3	Is there a business-side role development plan?	Product Owners are trained and developed.	The discovery team is trained and developed.	All the "snowman" roles are trained and developed.

7 Conclusions

Having a model as a framework is always useful as a guide and to ensure that we are on a safe path [8] [9]. The dynamics of organizations and changes in work teams often have as a consequence the disparity of criteria among the people who use an instrument. Without a doubt it is very important to maintain a clear direction when dealing with a challenge as great as it is to lead an organization towards a complete transformation and, furthermore, on a new paradigm that points to the same structures on which organizations are traditionally based.

This is why we saw the need to help maintain this coherence and the correct direction, especially at the very time when the strategy and roadmap to be followed in the framework of the agile transformation is decided.

Thus, we are convinced that a maturity model such as the one we have proposed in this article will help the members of the transformation team to carry out their work more easily, reducing the bias among them, allowing the identification of similar situations that favor the application of the same solutions, having a small roadmap to improve organizational skills throughout the successive transformation waves, and more.

As an obvious future work, we see not only the possibility but even the need to continue updating this model, both because of the opportunities for improvement that may arise, and because of the new tools, techniques and approaches that continue to appear permanently. And, fundamentally, we hope to continue improving the model from our real experience of putting it into practice in the next agile transformations with IATM.

Acknowledgement

I want to thank Arie van Bennekum, co-author of the Agile Manifesto and the main person responsible for the definition of the IATM, for the confidence that he gave me and for the enormous generosity with which he transmitted me his enormous knowledge and experience, while honoring me with being part of his work team in the agile transformation of organizations.

References

1. Appelo, J. Management 3.0: leading agile developers, developing agile leaders. Addison-Wesley, 2011.
2. Beedle, M. "Enterprise Scrum". [Online]. Available: http://www.enterprisescrum.com/. [Accessed: 12-Oct-2020].
3. Leffingwell, D. "Scaled Agile por Lean Enterprises 5.0" [Online]. Available: https://www.scaledagileframework.com/. [Accessed: 12-Oct-2020].
4. Cockburn, A. "Heart of Agile" [Online]. Available: https://heartofagile.com/. [Accessed: 12-Oct-2020].
5. van Bennekum, A. "Integrated Agile Transformation Model™. The Complexity in the simplicity of Agile", 2017.
6. van Bennekum, A. "Agile Maturity Diagnostics v.5.2", 2019.
7. Pinciroli, F. "A maturity model for the Integrated Agile Transformation Model™. Complementary material", https://uch.edu.ar/investigacion, 2020.
8. Sherehiy, B., Karwowski, W. and Layer, J. "A review of enterprise agility: concepts, frameworks, and attributes". International Journal of Industrial Ergonomics, vol. 37, no. 5, 2007, pp. 445-460.
9. Pinciroli, F. "Necesidad del empleo de herramientas estándares en XP". In: Proceedings of the XI Congreso Argentino de Ciencias de la Computación, Entre Ríos, 2005.

Beyond Random Split for Assessing Statistical Model Performance

Carlos Catania, Jorge Guerra, Juan Manuel Romero, and Gabriel Caffaratti and Martin Marchetta

Universidad Nacional de Cuyo.
Facultad de Ingeniería. LABSIN.
Mendoza. Argentina

Abstract. Even though a train/test split of the dataset randomly performed is a common practice, could not always be the best approach for estimating performance generalization under some scenarios. The fact is that the usual machine learning methodology can sometimes overestimate the generalization error when a dataset is not representative or when rare and elusive examples are a fundamental aspect of the detection problem. In the present work, we analyze strategies based on the predictors' variability to split in training and testing sets. Such strategies aim at guaranteeing the inclusion of rare or unusual examples with a minimal loss of the population's representativeness and provide a more accurate estimation about the generalization error when the dataset is not representative. Two baseline classifiers based on decision trees were used for testing the four splitting strategies considered. Both classifiers were applied on CTU19 a low-representative dataset for a network security detection problem. Preliminary results showed the importance of applying the three alternative strategies to the Monte Carlo splitting strategy in order to get a more accurate error estimation on different but feasible scenarios.

Keywords: Sampling strategies · Population representativeness · Dataset splitting

1 Motivation

An experimental design is a fundamental part of the machine learning workflow. In the particular case of prediction problems, part of the design includes estimating the model's generalization performance. Estimating this performance is a critical aspect of developing a model since it gives an idea of whether it can deal with future (not seen) scenarios reasonably well.

The standard experimental design for evaluating the performance of a machine learning model is well known. As depicted in Figure 1 a dataset is split usually in a 70/30 ratio. 30% of the data, called testing set, should be left aside and ideally never used until model tuning has finished. On the other hand, 70% of the data, referred as training set, could be used to train and optionally validate the model or conduct a hyperparameter search for model tuning.

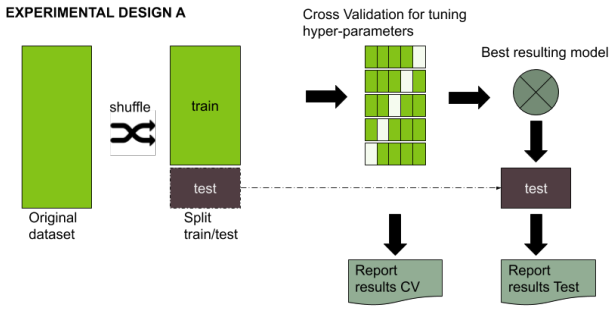


Fig. 1: Standard experimental design for evaluating the performance of a machine learning model.

Train and test datasets need some degree of similarity (both need to follow the same distribution). Otherwise, it would be impossible for the model to achieve a decent performance on test. However, if the examples are too similar in both datasets, then it is not possible to assure an accurate generalization performance for the model. Moreover, the model overall generalization performance could be overestimated.

Even though a train/test split of the dataset randomly performed is a common practice, is not always the best approach for estimating performance generalization under some scenarios. A common situation is when predicting patient outcomes. In these cases, the model should be constructed using certain patient sets (e.g., from the same clinical site or disease stage), but then need to be tested on a different sample population [4]. Another situation is the fact that it is not always possible to have access to a representative sample. Detecting a non-representative sample is possible through the application several techniques, such as cross-validation, learning curves, and confidence intervals, among others. Unfortunately, in many cases, a non-representative sample is all we have to generate a machine learning model. In those cases when a sample does not follow the same population distribution, a random split might not provide the required level of representativeness for rare or elusive examples in a testing set. As a result, the standard error metrics could overestimate the performance of the model. In classification problems, it is possible to deal with the lack of representativeness using a stratification strategy. However, when rare examples are not labeled, a predictor-based sampling strategy will be necessary [9, 1]

In the present work, we analyze several strategies based on the predictors' variability to split in training and testing sets. Such strategies aim at guaranteeing the inclusion of rare or unusual examples with a minimal loss of the population's representativeness. The hypothesis is that by including rare examples during model evaluation a more accurate performance estimates will be obtained.

The contributions of the present article are:

- The analysis of four splitting strategies with different distributions for training and testing sets.
- The evaluation of two different tree-based baseline classifiers over four different splitting strategies.

2 Splitting Strategies

2.1 Monte Carlo

The usual strategy for model evaluation consists of taking an uniformly random sample without replacement of a portion of the data for the training set, while all other data points are added to the testing set. Such strategy can be thought as special case of the Monte Carlos Cross Validation (MCCV) [10] with just one resampl instance. The Monte Carlo (MC) splitting strategy guarantees the same distribution across not only response but also predictor variables for training and testing sets.

In comparison with Monte Carlo, the remaining splitting strategies provide steps to create different test sets that include rare or elusive examples while maintaining similar properties across the predictor space as the training set.

2.2 Dissimilarity-based

Maximum dissimilarity splitting strategies were proposed by [1] and [9]. The simplest method to measure dissimilarity consist of using the distance between the predictor values for two samples. The larger the distance between points, the larger indicative of dissimilarity. The application of dissimilarity during data splitting requires a set initialized with a few samples. Then, the dissimilarity between this set and the rest of the unallocated samples can be calculated. The unallocated sample that is most dissimilar to the initial set would then be added to the test set.

Dissimilarity splitting proved to be useful over chemical databases' splitting [6, 5]. Nevertheless, this method strongly depends on the initial set used to calculate the dissimilarity of the rest of the samples, prompting problems in cases of small datasets where the initial set is not representative enough [11].

2.3 Informed-based

A well-known non-random split strategy consists of using some kind of grouping information from the data to restrict the set of samples used for testing. The general idea after splitting the data is that members of a group present in training set should not be included in the testing set. Such strategies are well-known in areas such as Medicine and Finances [4], where testing should be conducted on a different patient group or, in the finances field, where the model should be tested on a time series from a different time period.

2.4 Clustering-based

The clustering split strategy follows the same principle of the informed split. However, there could be situations where no grouping information can be extracted from the samples to perform an informed split. In these cases, the application of a clustering algorithm could be used to replace the missing information. The labels generated by this procedure will be then used for performing a group split similarly to the informed split strategy.

3 Application to a Network Security Dataset

The four splitting strategies described in section 2 were applied to a network security dataset for botnet detection conformed by nineteen network captures published by the stratosphere IPS research group at CTU [7]. Specifically, the dataset has fourteen botnet captures and five normal captures including traffics like DNS, HTTPS and P2P. In total, all captures represents 20866 connections having 19271 labeled as "Botnet" and 1595 labeled as "Normal". All these captures were gathered between 2013 and 2017.

The first ten predictors of the CTU19 dataset summarize each flow of connections with same IP, protocol and destination port into a 10-dimensional numerical vector:

$$\mathbf{v}_{feat} =< x_{sp}, x_{wp}, x_{wnp}, x_{snp}, x_{ds}, x_{dm}, x_{dl}, x_{ss}, x_{sm}, x_{sl} > \quad (1)$$

wherein the first four dimensions of the numerical vector represent the periodicity predictors (strong periodicity (x_{sp}), weak periodicity (x_{wp}), weak non periodicity (x_{wnp}) and strong non periodicity (x_{snp})), the next three refer to the duration predictors (duration short (x_{ds}), duration medium (x_{dm}) and duration large (x_{dl}), respectively), and the last three represent the size predictors (size short (x_{ss}), size medium (x_{sm}), size large (x_{sl})). The vector for a given connection is generated considering the cumulative frequency of the corresponding values associated with the behavioral of each predictor.

In addition to the information provided by the flow-based predictors, the CTU19 dataset includes information related with the flow like source IP, destination IP, protocol, port and the source linked with each capture. However, the present study will focus only on the information provided by the flow-based predictors as discussed in [2, 8].

3.1 Initial Exploratory Analysis

Fig. 2 represents a 2D projection of the CTU19 dataset considering the first 2 principal components (PCA was applied to the flow-based predictors). In addition, the figure includes the distribution for each predictor. As depicted by the box plot, botnet and normal classes present a different distribution over the flow-based predictors. However, both classes show a partial overlap at the time of projecting them into a 2D space.

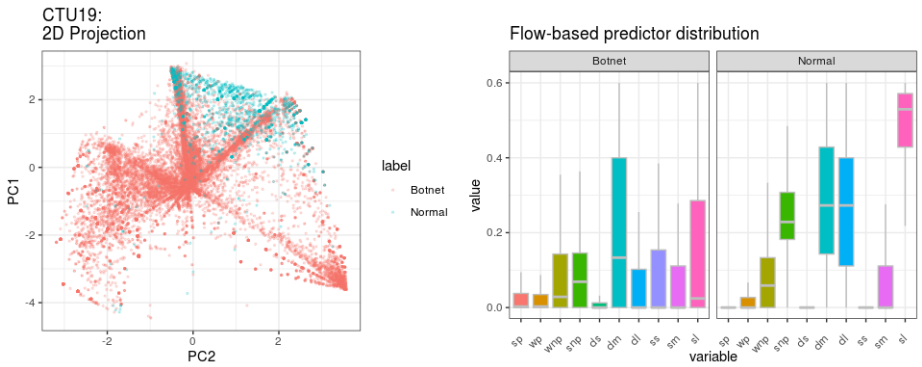


Fig. 2: Boxplot and 2D projection of the flow-based predictors in the CTU19 dataset. Botnet traffic in red and Normal in blue.

Fig. 3 decomposes the CTU19 dataset by traffic capture with the same information as provided by Fig 2. In general, the patterns from the 2D projection normal captures are different compared with Botnet captures. Normal captures are concentrated while Botnet captures spread along the 2D predictor space. When analysing Normal captures, most of them overlap the same predictor space. In other words, each capture has some examples on every portion of the Normal predictor subspace, which suggests the adequate representativeness. Nevertheless, in the case of Botnet, there are several cases of captures having only a limited presence on the class predictor subspace (see captures 2014-02-07-win3 and 2014-01-25-win3). Such lack of representativeness observed by some Botnet captures could difficult the classification model's performance estimation.

3.2 Training and Testing Sets Creation

Fig. 4 describes the process used for the creation of the training and testing sets according to each splitting strategy. A subset of 3000 data points from CTU19 were used for each strategy. 2000 data points for conforming the training set and 1000 for the testing set. The generation of the training and testing sets will depend on the splitting strategy applied as discussed in section 2. The procedure is repeated 25 times. Therefore 25 pairs of training and testing sets are generated for each splitting strategy.
The differences between each splitting strategy are observed in Fig 5 considering data points from the 25 samples. The 2D projection using the first two principal components confirms the similarities between training and testing sets when the baseline Monte Carlo splitting is applied. Moreover, both datasets follow the same pattern in the predictor space which corresponds with the similarity observed in the predictor distributions (see box plots below). However, different patterns are observed between training and testing sets in the remaining splitting strategies. In particular, dissimilarity-based and informed-based splitting

Fig. 3: On the top, the scatter plots on the top depicts the 2D projection using PCA for each one of the 19 captures conforming the CTU19 dataset (Botnet traffic in red and Normal in blue). On the bottom, the corresponding distributions for the 10 flow-based predictors on each capture.

present the most different patterns. Nevertheless, the same pattern is observed in both datasets for the clustering-based splitting case, although with a small displacement from the axis.

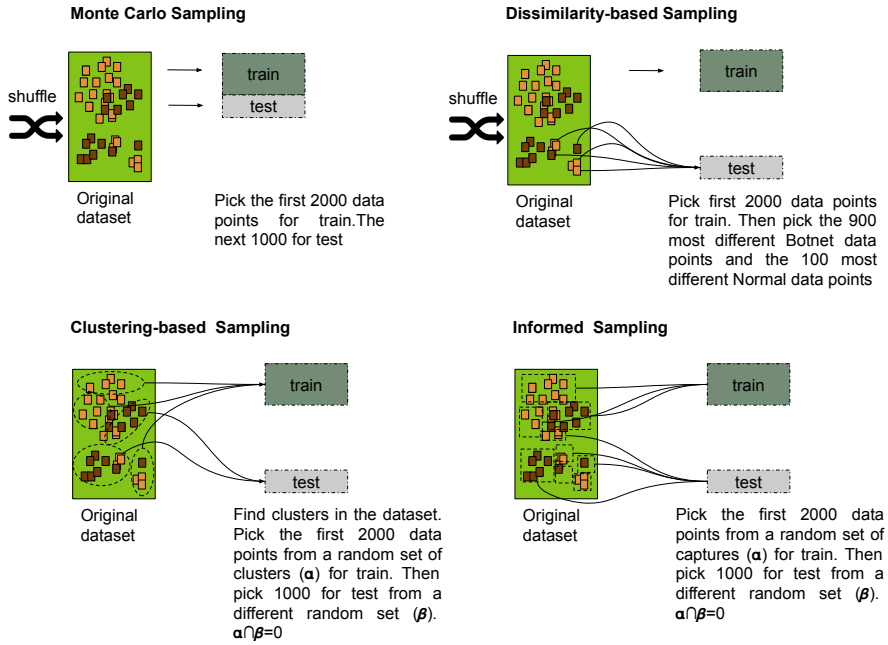


Fig. 4: Procedure used for generating the training and testing sets for the four splitting strategies: Monte Carlo, Dissimilarity-based, clustering-based and informed-based. In all the cases a training set with 2000 data points and a testing set with 1000 were generated.

When projecting many predictors into two dimensions, intricate predictor relationships could mask regions within the predictor space where the model will inadequately predict new samples.

An algorithmic approach as described in [3] is applied to get more detailed information about the similarities between training and testing datasets. The general idea is to create a new dataset randomly permuting the predictors from the training set and then, row-wise concatenate to the original training set labeling original and permuted samples. A classification model is run on the resulting dataset to predict the probability of new data being in the class of the training set. Table 1 shows average percentage of samples from testing set not recognized as part of the training set.

As expected, Monte Carlo splitting exhibits the lower error and standard deviation, confirming the similarities between training and testing set observed in the 2D projection from Fig.5. Both dissimilarity and informed-based strategies has the same average error. However, informed-based shows a higher variation. Finally, the clustering-based strategy shows the biggest differences between training and testing sets.

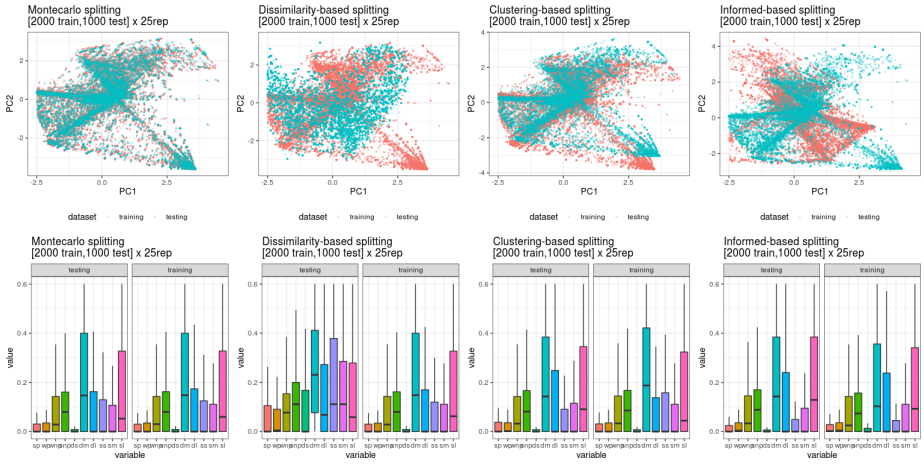


Fig. 5: 2D projection and boxplot distribution for the 25 pairs of training and testing datasets for each splitting strategy.

Table 1: Average Percentage and standard deviation of test samples not recognized as part of the training set for the four splitting strategies

Avg Err %	sd	Splitting Strategy
0.06	0.05	informed-based
0.01	0.01	monte carlo
0.17	0.13	cluster-based
0.06	0.02	dissimilarity-based

4 Error Estimation on Baseline Classifiers

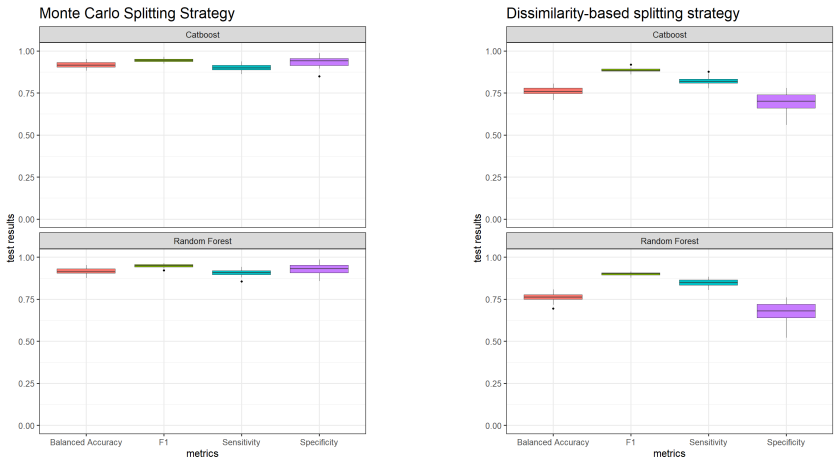
The impact of the different splitting strategies discussed in section 2 is measured on Random Forest (RF) and CatBoost (CB) classifiers. Both CB and RF are two well-known classifiers providing acceptable results on tabular data without conducting a hyper parameters tuning. Both baseline classifiers were executed with default parameters. Nevertheless, downsampling technique is applied to training set to deal with classes imbalance issues.

4.1 Metrics

Several standard metrics for model classification assessment were used to evaluate baseline classifiers performance on the different train/test datasets. The metrics correspond to True Positive Rate (TPR) and False Positive Rate (FPR). The **Sensitivity** measures the proportion of positives that are correctly identified (TPR), and the **Specificity** measures the proportion of negatives that are correctly identified ($1 - FPR$).

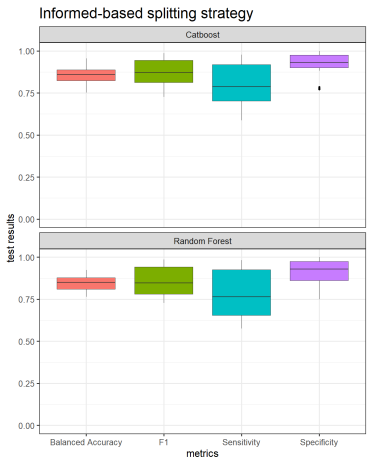
Additional metrics were used to deal with class imbalance: **F1-Score** and **Balanced Accuracy**. F1-Score is computed as the weighted average between TPR and the total numbers of malicious connections in the dataset. Balanced Accuracy is calculated as the average of correctly classified proportion of each class individually.

4.2 Results

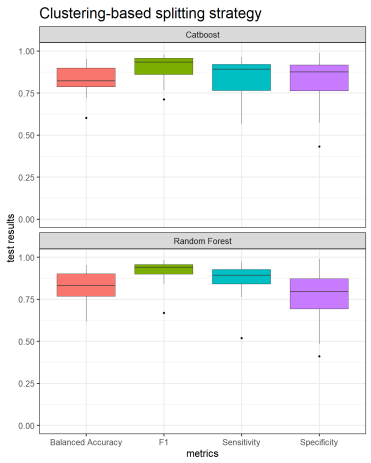


(a) Monte Carlo splitting Strategy

(b) Dissimilar splitting Strategy



(c) Informed-based splitting Strategy



(d) Clustering-based splitting Strategy

Fig. 6: Random Forest and CatBoost models' performance result for each splitting strategy

Fig. 6 depicts the overall performance results for baseline classifications models using the splitting strategies described in Section 2. Both models were used to predict "Botnet" or 'Normal" connections. Since no significant difference is observable in the performance between both baseline models, the results discussed in this section corresponds only to the catboost algorithm.

Fig. 6a displays testing set's results for RF (below) and CB (above) using the MC splitting strategy. Prediction performance in terms of Balanced Accuracy median is around 0.91. The remaining metrics obtained similar values for both baseline models. A low variability is observed in all the considered metrics. For the Dissimilarity-based splitting strategy (see Fig. 6b), all considered metrics decrease in comparison to MC strategy. The Balanced Accuracy median decreases to 0.75 while the specificity decreases to 0.70. The F1 and Sensitivity median maintained higher values, although lower than the observed in the MC strategy.

Fig. 6c states the results for the Informed-based splitting strategy. In this case, Balanced Accuracy and F1 show a median value of 0.85, 0.87, respectively. On the other hand, in terms of Specificity and Sensitivity, the median values are around 0.78 and 0.93, respectively. Notice that despite the good performance in terms of median values, a considerable variation is observed for the F1 and Sensitivity metrics. Finally, Fig. 6d presents the models' results using the Clustering splitting strategy. The Balanced Accuracy median value decreased to 0.82 while F1 increases to 0.93 compared with the informed-splitting strategy. The rest of the metrics, Sensitivity and Specificity, shown similar results with median around 0.89 and 0.87, respectively.

5 Discussion

In general both baseline models presented acceptable performance for predicting "Botnet" and "Normal" classes since the median Balanced Accuracy metric was over 0.75 in most of cases. As expected, the Monte Carlo splitting strategy showed the smallest estimation error (a high balanced accuracy). The similarities between training and testing sets was already observed in Table 1, where only 1% of the testing set was not recognized as part of training set. When the size of not recognized samples increments only to a 6%, such as in the case of dissimilarity strategies, the estimation error increases considerably (from 0.91 to 0.75 median Balanced Accuracy). Since Monte Carlo and Dissimilarity-based splitting strategies use the same procedure for generating the training set, their analysis can provide a valuable error estimation under an anomalous but certainly feasible set of samples.

In the case of informed and clustering based splitting strategies, the difference between training and testing sets is not only larger than Monte Carlo but also both show a larger variation (see Table 1). A considerable variation is also observed in the performance of both baseline classifiers. However, both splitting strategies still provide information about the robustness of the models on sets with different representativeness. Previous statement is particularly valid for the informed-based splitting strategy, where captures with very different repre-

sentiveness levels are used for building training and testing sets. For instance, informed-based Balanced Accuracy range provides ad-hoc information about the expected values when not so representative sets are used for training. Moreover, it is possible to infer that a 0.75 value for Balanced Accuracy could be the worst performance scenario observed by the model. Such value is still suitable under some real-life situations.

The clustering-based strategy provides a more extreme scenario than informed-based for estimating the model performance under not so representative sets. Under the clustering-base strategy a concentrated portion of the predictor space present in the testing set is excluded from training set, whereas in the informed-based it is possible to find datapoints spread along the whole predictor space. Consequently, it is possible to observe higher variation in the baseline models performance.

6 Concluding Remarks and Future Work

Despite being the standard splitting strategy, Monte Carlo can overestimate the results when dataset is not representative. Other splitting techniques based on dissimilarity, information present in the dataset and the application of a clustering algorithm can help in the estimation under different low-representativeness scenarios.

Multiple training and testing sets where generated using the different strategies on the CTU19 Botnet dataset. Small differences between training and testing sets were corroborated using the algorithm proposed by [3] in all the four techniques. As expected Monte Carlo showed the smallest differences while clustering-based showed the biggest.

Two baseline classifiers were used for evaluating each splitting strategy in the error estimation process. The Dissimilarity-based splitting strategy provided a valuable error estimation under an anomalous but certainly feasible set of samples. On the the other hand, informed-based strategy offers ad-hoc information about the expected values when sets not so representive are used for training, while clustering-based strategy emerges as an alternative to informed-based for estimating the model performance under low representativeness sets with more pessimist estimation.

Preliminary results showed the importance of applying the three alternative strategies to the Monte Carlo splitting strategy in order to get a more accurate error estimation on different but possible situations. However, given the particular low-representativeness nature of the botnet detection problem, a deeper analysis and evaluation on other datasets should be conducted.

7 Acknowledgments

The authors would like to thank the financial support received by SIIP-UNCuyo during this work. In particular the projects 06/B363 and 06/B374. In addition,

we want to gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

1. Clark, R.D.: OptiSim: An extended dissimilarity selection method for finding diverse representative subsets. Journal of Chemical Information and Computer Sciences **37**(6), 1181–1188 (1997). https://doi.org/10.1021/ci970282v
2. Guerra, J.L., Veas, E., Catania, C.A.: A study on labeling network hostile behavior with intelligent interactive tools. In: 2019 IEEE Symposium on Visualization for Cyber Security (VizSec). pp. 1–10 (2019). https://doi.org/10.1109/VizSec48167.2019.9161489
3. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media (2009)
4. Kuhn, M., Johnson, K.: Applied Predictive Modeling with Applications in R. Springer (2013)
5. Martin, T.M., Harten, P., Young, D.M., Muratov, E.N., Golbraikh, A., Zhu, H., Tropsha, A.: Does rational selection of training and test sets improve the outcome of QSAR modeling? Journal of Chemical Information and Modeling **52**(10), 2570–2578 (2012). https://doi.org/10.1021/ci300338w
6. Snarey, M., Terrett, N.K., Willett, P., Wilton, D.J.: Comparison of algorithms for dissimilarity-based compound selection. Journal of Molecular Graphics and Modelling **15**(6), 372–385 (1997). https://doi.org/10.1016/S1093-3263(98)00008-4
7. Stratosphere IPS Project: The CTU-19 dataset, Malware Captures. https://mcfp.felk.cvut.cz/publicDatasets/ (October 2017), [Online; accessed Jun-2020]
8. Torres, J.L.G., Catania, C.A., Veas, E.: Active learning approach to label network traffic datasets. Journal of Information Security and Applications **49**, 102388 (2019). https://doi.org/https://doi.org/10.1016/j.jisa.2019.102388, http://www.sciencedirect.com/science/article/pii/S2214212618304344
9. Willett, P.: Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. Journal of Computational Biology **6**(3-4), 447–457 (1999). https://doi.org/10.1089/106652799318382
10. Xu, Q.S., Liang, Y.Z.: Monte carlo cross validation. Chemometrics and Intelligent Laboratory Systems **56**(1), 1–11 (2001)
11. Yang, Y., Ye, Z., Su, Y., Zhao, Q., Li, X., Ouyang, D.: Deep learning for in vitro prediction of pharmaceutical formulations. Acta Pharmaceutica Sinica B **9**(1), 177–185 (2019). https://doi.org/10.1016/j.apsb.2018.09.010

Estudio de eficiencia de redes de sensores para Inteligencia Ambiental

Adrián Puebla, Rosana Gimenez, and Pablo Sandez y Miguel Mendez-Garabetti

Laboratio de Invstigación en Ciencia y Tecnología(LabICyT-UDA)
Facultad de Ciencias Sociales y Adinistrativas, Universidad del Aconcagua
adrianep@live.com.ar

Abstract. La evolución de la Inteligencia Ambiental apunta a enfrentar problemas ambientales como el cambio climático, la contaminación del aire y el control de plagas, entre otros. Este concepto utiliza diferentes tomas de información y datos de alta calidad, provenientes de observaciones y monitoreo de aguas, suelo, aire, biodiversidad, etc. de una región determinada. Las redes inalámbricas de sensores (Wireless Sensor Networks) resuelven en gran parte esta toma de información posibilitando la creación de redes con la obtención de variables (Humedad, temperatura, etc.) en diferentes lugares y al mismo tiempo. En este proyecto se evaluará la eficiencia de una gran red inalámbrica de sensores (Wireless Sensor Networks) para Inteligencia Ambiental, simulando con Cupcarbon en zonas de la provincia de Mendoza, Argentina. En este caso se utilizará similar topología de red, pero con escenarios de diferentes protocolos como WiFi, LoRa y ZigBee. Se buscará la determinación del consumo de energía de cada nodo en distintos escenarios, realizando una comparación con los resultados obtenidos y determinando cuál red con WiFi, LoRa o ZigBee, es más eficiente en su consumo.

Keywords: Inteligencia Ambiental · Wireless Sensor Networks · Cupcarbon · Simulación en la provincia de Mendoza.

1 Introducción

1.1 Inteligencia Ambiental(IE)

El concepto de inteligencia ambiental abarca muchos ámbitos, actualmente el enfoque más pensado es en la Domótica, mejorando los lugares donde habita el ser humano y evitando que esté pendiente de muchas tareas en su hogar, dejándolo libre para enfocarse en otras cosas. En el cuidado de personas ancianas o con enfermedades que necesiten de una atención permante, se podría realizar monitoreo inteligente y ante indicadores de importancia dar avisos de alertas.

En general los Sistemas toman información de áreas en particular para alimentar bases de datos, también captan información en tiempo real para realizar acciones necesarias.

Es pertinente citar un trabajo de investigación donde los científicos de la Administración Nacional Oceánica y Atmosférica de Estados Unidos (NOAA) ocupan Inteligencia Ambiental[1]

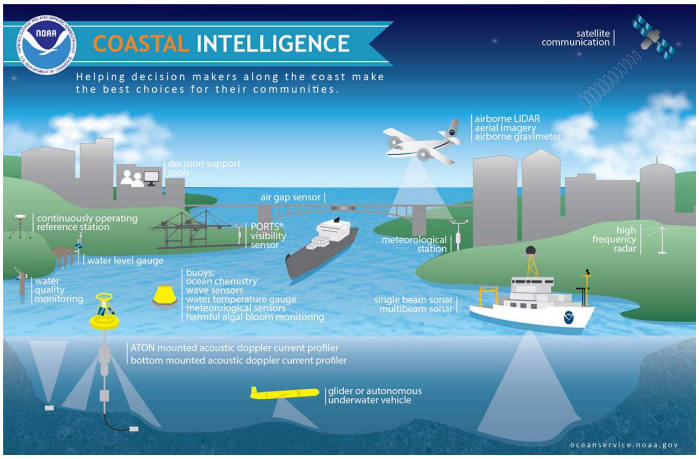


Fig. 1. Red de sensores ayudando a la Navegación, figura obtenida en el sitio web de National Ocean Service, www.oceanservice.noaa.gov/tools/coastalintelligence/

En este ejemplo la información obtenida por los sensores indica las condiciones en la zona costera. Las personas en la comunidad marítima confían en esta inteligencia para una variedad de decisiones, desde cuánta carga almacenar en un barco hasta elegir la ruta más eficiente y segura entre dos puntos marítimos.

La inteligencia ambiental ha tomado gran importancia que puede verse también en otros sistemas como la detección de contaminación ambiental, estudios climáticos, enfermedades en cultivos, etc.

1.2 Redes Inalámbricas de Sensores WSN

Las redes inalámbricas de sensores WSN están formadas por dispositivos que se comunican entre nodos, que son capaces de obtener información de una zona, procesarla de forma local y enviarla a través de enlaces inalámbricos hasta un nodo central; los nodos actúan como elementos de infraestructura de las comunicaciones entre el nodo central y el nodo más lejano.
La red de sensores está formada por muchos dispositivos distribuidos espacialmente, que utilizan sensores para controlar diversas condiciones en distintos puntos, entre ellos están la humedad, temperatura, sonido, la presión y movimiento; en un escenario normal los sensores pueden ser fijos o móviles.
Las redes WSN pueden ser muy económicas y en general de una electrónica de

fácil implementación. Se puede utilizar la electrónica de Arduino que está a la mano de cualquier persona.
Se tomará como ejemplo una red de sensores inalámbricos para el monitoreo del arroyo de agua, la que tiene una estructura de seis nodos que se encargarán de sensar las condiciones ambientales presentes en su recorrido.[2]

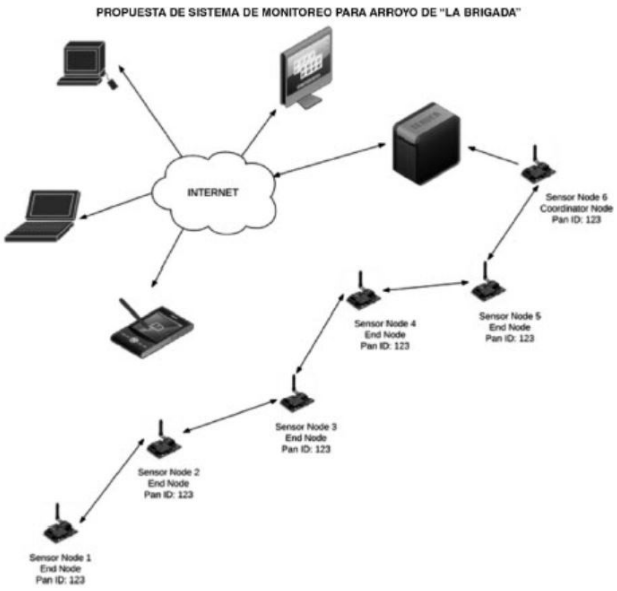


Fig. 2. Red sensando el estudio de un arroyo, figura obtenida en el sitio web de Scientific Electronic Library Online(Chile), http://dx.doi.org/10.4067/S0718-33052016000400005

Las redes de sensores inalámbricos desarrollarán un amplio crecimiento, entre sus ámbitos se encuentran:

• Monitoreo climático.
• Servicio en agricultura de precisión: control de clima, humedad, recolección de datos sobre el terreno, control de riegos. Además medición de niveles de contaminantes, o de sustancias no autorizadas, determinación de productos químicos.
• Servicios de Automatización: control de invernaderos, alimentación de animales, etc.
• Medición de características físicas de pacientes en entornos hospitalarios o a distancia.
• Servicio de seguridad de zonas ante posibles intrusos.

1.3 Estandar ZigBee, WiFi y LoRa

En este proyecto se utilizarán los estándares ZigBee, WiFi y LoRa. A continuación se describe de forma muy breve cada uno.

ZigBee Este estándar fue creado por la Alianza ZigBee[4], que tiene gran cantidad de compañías, desde fabricantes de semiconductores hasta desarrolladores de software. Es una organización sin fines de lucro que surge en el año 2002 y es muy utilizado en la actualidad.
ZigBee es el estándar más aceptado hoy para usar en redes de sensores y actuadores que deban operar a batería.
Este estándar es para el armado de redes Inalámbricas de corta distancia y baja velocidad de datos.

Ventajas del estándar ZigBee:[5][6]

• Bajo consumo que permita usar equipos a batería.
• Bajo costo de dispositivos, de instalación y mantenimiento.
• De corto alcance, menor a 75mts.
• Optimizado para ciclo efectivo de transmisión menor a 0.1 (porc).
• Velocidad de transmisión menor que 250 kbps. Típica: menor que 20 kbps.
• El hardware necesario para poder crear un dispositivo es mucho más reducido comparado con Bluetooth o WiFi.
• Opera en las bandas de 868 MHz, 915 MHz y 2.4 GHz y puede transferir datos hasta 250Kbps.

Wifi Este estándar es una tecnología de red inalámbrica a través de la cual los dispositivos como teléfonos, computadoras, impresoras , etc., pueden interactuar en una red local de datos o con internet.
Wi-Fi esta administrada por la "Wi-Fi Alliance"[7], organización comercial que cumple con los estándares 802.11 y certifica el hardware y software para este estándar. La conectividad de una red de wifi se logra a través de un router inalámbrico que administra la red. En la Tabla 1 se ejemplifican algunas características de los estándares WIFI[8][9].

Table 1. Características de estándares Wifi mas usados

Características	802.11	802.11a	802.11b	802.11g	802.11n	802.11ac	802.11ad
Vel Teórica	2 Mbit/s	54 Mbit/s	11 Mbit/s	54 Mbit/s	600 Mbit/s	6.93 Gbps	7.13 Gbit/s
Vel Práctica	1 Mbit/s	22 Mbit/s	6 Mbit/s	22 Mbit/s	100 Mbit/s	100 Mbit/s	6 Gbit/s
Frec	2,4 Ghz	5,4 Ghz	2,4 Ghz	2,4 Ghz	2,4 y 5,4 Ghz	5,4 Ghz	60 Ghz
AB	22 MHz	20 MHz	22 MHz	20 MHz	20/40 MHz	80 a 160 MHz	2 MHz
Dist	330 m	390 m	460 m	460 m	820 m	460m	300 m
Año	1997	1999	1999	2003	2009	2013	2012

LoRa Este estandar de tecnología inalámbrica similar a Bluetooth, WiFi, SigFox o Zigbee. LoRa utiliza un tipo de modulación en radiofrecuencia patentado por Semtech, una importante empresa fabricante de chips de radio.

La tecnología de modulación se llama Chirp Spread Spectrum(CSS), y se usa en comunicaciones militares. La gran ventaja de la misma es que puede lograr comunicaciones a largas distancias y tiene gran solidez frente a las interferencias.

LoRa es una tecnología originalmente creada por Semtech, pero actualmente está administrada por la "LoRa Alliance"[3]. Los fabricantes deben estar certificados en esta entidad para poder producir cualquier hardware.
Ventajas de LoRa[10]
• Gran tolerancia a las interferencias
• Gran sensibilidad para recibir datos
• Basado en modulación "chirp"
• Bajo Consumo
• Distancia alcance 10 a 20km
• Poca transferencia de datos
• Conexión punto a punto
• Frecuencias de trabajo: 868 Mhz en Europa, 915 Mhz en América, y 433 Mhz en Asia

Es una tecnología ideal para conexiones a grandes distancias y para redes que se necesiten sensores que no dispongan de corriente eléctrica de la red. Aplicable a ciudades inteligentes, lugares con poca cobertura celular cómo por ejemplo zonas agrícolas o ganaderas en el campo, o para construir redes privadas de sensores y/o actuadores.

1.4 Simuladores Cupcarbon

En este proyecto se utilizará el simulador de red Cupcarbon.

Cupcarbon es un simulador[1] de red de sensores inalámbricos. Sirve para diseñar, visualizar, depurar, validar algoritmos distribuidos para monitoreo y crear escenarios ambientales con incendios, lluvias, gas. Es muy útil para proyectos educativos y científicos. Ayuda a explicar visualmente los conceptos básicos de las redes de sensores y su funcionamiento. También es muy intuitivo para científicos que necesitan evaluar topologías inalámbricas
Cuenta con dos entornos de simulación: El primer entorno de simulación permite el diseño de escenarios de movilidad y la generación de eventos naturales como incendios y gas, así como la simulación de móviles como vehículos. El segundo entorno de simulación se enfoca en una simulación de eventos discretos de redes WSN que tiene en cuenta el escenario diseñado sobre la base del primer entorno.

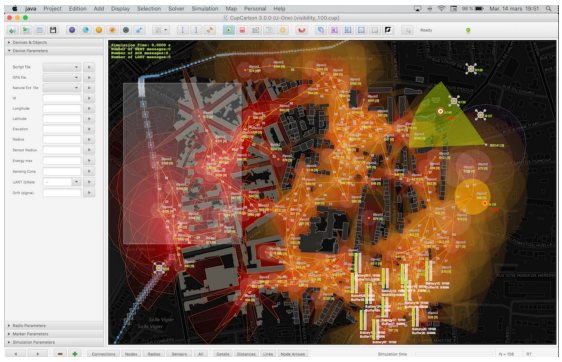


Fig. 3. Vista del simulador Cupcarbon, figura obtenida en el sitio web de CupCarbon, https://cupcarbon.com/

Las redes se pueden diseñar y crear simulaciones en mapas reales utilizando el marco de OpenStreetMap (OSM) para implementar sensores directamente ubicados en el mapa. Se puede programar y configurar cada nodo de forma individual. Una vez creado y probado el código se pueden configurar plataformas de hardware como Arduino/XBee. Pero por el momento sólo puede generar códigos para redes y algoritmos simples.
La versión actual de CupCarbon permite configurar dinámicamente los nodos para poder realizar cambios y unir o separar distintas redes, evaluándolos en tiempo real. Además, se puede estimar el consumo de energía que se puede calcular y mostrar en función del tiempo. Esto permite aclarar la estructura, viabilidad e implementación realista de una red antes de su despliegue real. Incluyen los protocolos LoRa, WiFi y ZigBee.

2 Simulación y análisis de datos

2.1 Escenario modelo a simular

Si bien es importante aclarar que este estudio se enfoca en la red WSN y en sus protocolos, es posible imaginar la gran utilidad que puede brindar la implementación de una red como la propuesta. En este ejemplo se creará una red de sensores donde se medirá la contaminación de sectores en Mendoza, que estarán divididos en zonas más transitadas por personas a pie o en bicicleta. Una de las finalidades podría ser que los datos alimentan un concentrador de información donde transmite a una aplicación celular e indica a los usuarios el estado de la contaminación en tiempo real, ayudándolo a elegir por qué camino circular o cuándo es el mejor momento de transitar o realizar deportes y así evitar problemas de salud.

Zonas seleccionadas:

Se han seleccionado seis zonas de sondeos en la provincia de Mendoza, que podrian verse como Seis sub redes. Como se mencionó anteriormente lo que interesa, son lugares transitados por gran cantidad de personas y donde se realizan prácticas deportivas.

Zonas de sondeo(Fig. 4):

- Cercanías al lago de Parque General San Martin
- Plaza Independencia y Peatonal
- Espacios verdes alrededor de Casa de Gobierno y Poder judicial
- Plaza Godoy Cruz y Espacio Verde Luis Menotti Pescarmona
- Boulevard Dorrego
- Sector del camino peatonal del Barrio Bancario de G. Cruz

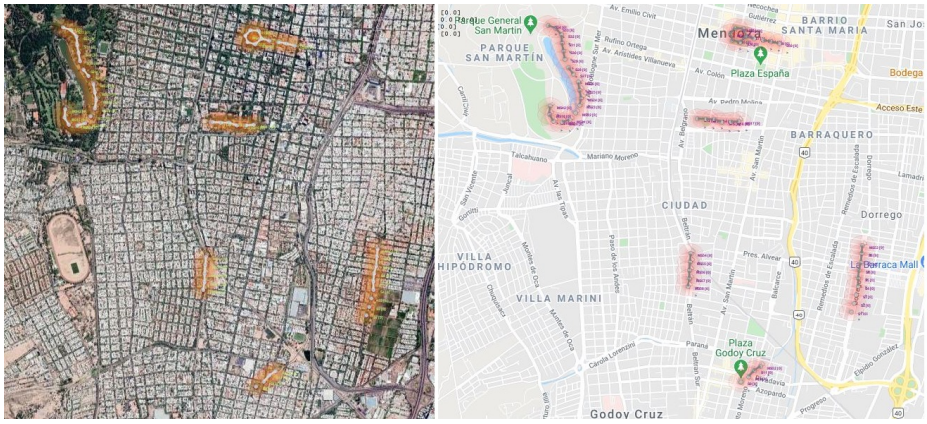


Fig. 4. Zonas de sondeo de contaminación ambiental, figura capturada en el simulador CupCarbon

2.2 Configuración del simulador

Se configura individualmente cada uno de los nodos y se indica mediante script cómo se va a enviar la información. También se configura el consumo energético de cada nodo de acuerdo al dispositivo seleccionado y el protocolo a utilizar en cada escenario.

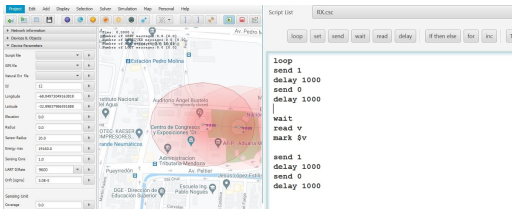


Fig. 5. Configuración de nodo en Cupcarbon, figura capturada en el simulador CupCarbon

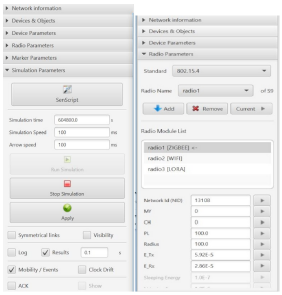


Fig. 6. Configuración de Simulación y Protocolos de red, figura capturada en el simulador CupCarbon

2.3 Simulación y análisis los datos

Simulación en los escenarios propuestos Se simula en los escenarios con protocolos WIFI, LORA y ZIGBEE. En la siguiente imagen se visualiza la simulación en acción:

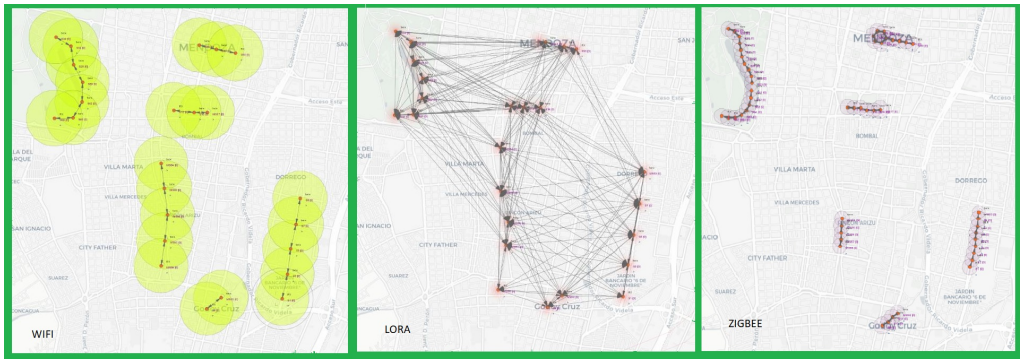


Fig. 7. Simulando los escenarios, figura capturada en el simulador CupCarbon

En el escenario con WIFI se utilizaron 27 nodos que conforman 6 redes, las cuales deben conectarse entre sí con otra tecnología debido a problemas de limitación de alcance en la señal. El mismo problema de alcance afecta a ZIGBEE, que tiene distancia mucho menor y se deben utilizar 52 nodos. En cambio en LORA al tener gran alcance se puede tener toda la red conectada con la misma tecnología, también utiliza 27 nodos, pero por la necesidad de ubicación de los sensores.

Análisis de la red y los datos obtenidos En cada escenario se pudo obtener las gráficas de consumo evaluando la duración de la batería con una sola carga:

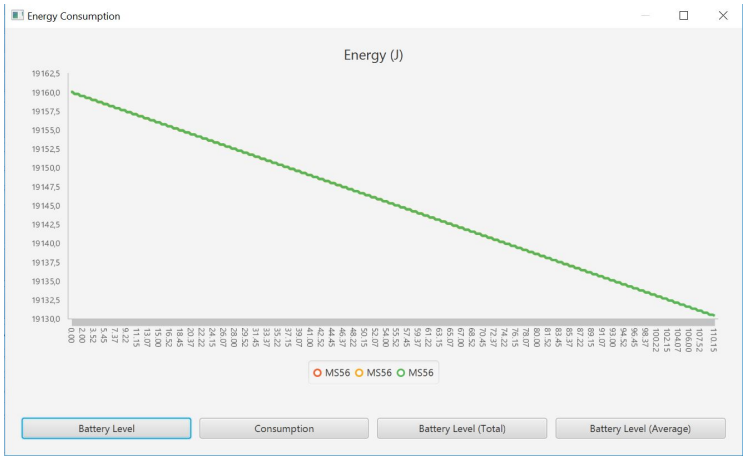


Fig. 8. Gráficas de consumo de un nodo en WIFI, figura capturada en el simulador CupCarbon

3 Cálculos de consumo de acuerdo a las gráficas

Con los datos obtenidos en las gráficas se pueden realizar los cálculos de consumo energético de cada nodo.

Wifi:

E(Total)= La energía inicial de baterías en el simulador es 19169J
E(Medida)= Energía medida en la simulación a los 110 segundos

10

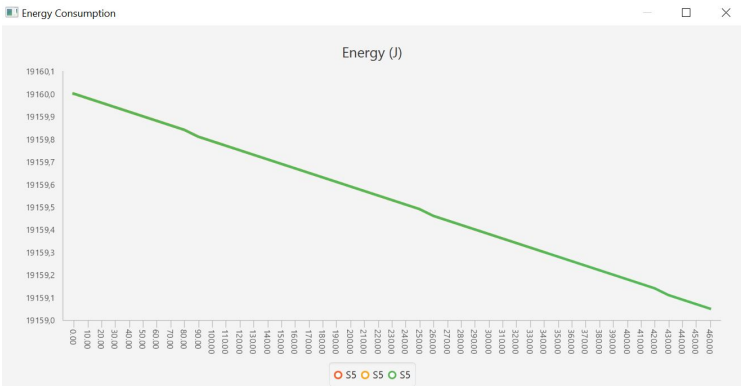


Fig. 9. Gráficas de consumo de un nodo en LORA, figura capturada en el simulador CupCarbon

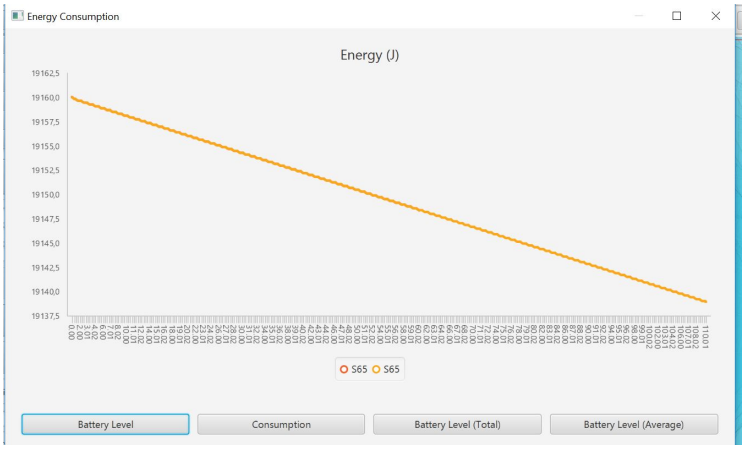


Fig. 10. Gráficas de consumo de un nodo en ZIGBEE, figura capturada en el simulador CupCarbon

E(Total) - E(Medida)= 19169 - 19130=39J
P=E/T=39J/110,15s = 0,35W
P=IxV entonces I=P/V – 0,35W/5V =0,070A= 70mA
P=E/T entonces T=E/P=19169/0,35=54768,6 s
Regla de tres:
86400s – 1 día
54768,6s – 0,63 días duraría la batería con una sola carga

LORA:

E(Total)= La energía inicial de baterías en el simulador es 19169J

Table 2. Consumos medios según fabricante

Tecnología	WIFI	LORA	ZIGBEE
Corriente TX	300mA	35mA	120mA
Corriente RX	NC	0,003mA	0,0018mA
Tensión	3 a 5v	3 a 5v	3 a 5v

E(Medida)= Energía medida en la simulación a los 110 segundos
E(Total) - E(Medida)= 19169 - 19159,8=9,2J
P=E/T=9,2J/110s = 0,08(J/s)
P=IxV entonces I=P/V – 0,08W/5V =0,016A= 16mA
P=E/T entonces T=E/P=19169/0,08=239.612,5 s
Regla de tres:
86400s – 1 día
239.612,5s – 2,77 días duraría la batería una sola carga

Zigbee:

E(Total)= La energía inicial de baterías en el simulador es 19169J
E(Medida)= Energía medida en la simulación a los 110 segundos
E(Total) - E(Medida)= 19169 - 19137,5=31,5j
P=E/T=31,5J/110s = 0,29(J/s)
P=IxV entonces I=P/V – 0,29W/5V =0,058A= 58mA
P=E/T entonces T=E/P=19169/0,29=66100 s
Regla de tres:
86400s – 1 día
66100s – 0,76 días duraría la batería una sola carga

Table 3. Consumos calculados del Simulador

Tecnología	WIFI	LORA	ZIGBEE
Corriente TX/RX	70mA	16mA	58mA
Tensión	5v	5v	5v

Calculando el consumo que arrojan las Fig 8, Fig 9 y Fig 10 se puede tomar el valor total de energía en la batería y el valor consumido a los 110 segundos, determinando un valor de consumo aproximado.
Evaluando la Tabla 2 "Consumos según fabricante" en comparación con los valores obtenidos en la simulación Tabla 3 "Consumos calculados del Simulador" se obtienen variaciones, debido a que en el simulador el consumo influyen de la cantidad de mensajes transmitidos que se programan en los scripts de cada nodo

sumados al consumo de los sensores de dióxido de carbono.

4 Conclusiones

Una vez evaluados los escenarios de WIFI, LORA y ZIGBEE, los resultados indican que la tecnología con menor consumo energético es LORA, también se puede aprovechar este beneficio agregando más cantidad de sensores por cada nodo, mejorando así los valores censados, o elegir esta tecnología en zonas alejadas o de difícil acceso donde la vida de la batería, al ser más duradera, requiere un menor mantenimiento.

Referencias

1. National Oceanic and Atmospheric Administration U.S, Department of Commerce, (2020) https://oceanservice.noaa.gov/tools/coastalintelligence/.
2. Alejandro Cama-Pinto, Melisa Acosta-Coll, Gabriel Piñeres-Espitia, José Caicedo-Ortiz, Ronald Zamora-Musa, Sepulveda-Ojeda. (2016) Titulo: Diseño de una red de sensores inalámbricos para la monitorización de inundaciones repentinas en la ciudad de Barranquilla, Colombia Facultad de Ingeniería, Universidad de la Costa, Colombia, http://dx.doi.org/10.4067/S0718-33052016000400005.
3. Web oficial de Lora Alliance (2020), https://lora-alliance.org/about-lorawan.
4. Zigbee Alliance (2020), https://zigbeealliance.org/es/.
5. Javier Martín Moreno, Daniel Ruiz Fernández,(2007) Título: Protocolo ZigBee , $https://rua.ua.es/dspace/bitstream/10045/1109/7/InformezigBee.pdf$.
6. Maria Agusta Loyola Pinos, Paul Alfonso Becerra Delgado, (2015) Pretesis: Manual para la aplicación de la tecnología zigbee para edificios inteligentes en la ciudad de cuenca, Univ. Politécnica Salesiana, $https://rua.ua.es/dspace/bitstream/10045/1109/7/InformezigBee.pdf$.
7. Web oficial de Wifi Alliance (2020), https://www.wi-fi.org/.
8. Raúl Martínez Rivas,(2014) Título: Evaluación de características técnicas para la implementación de un sistema automático de trazabilidad de bienes académicos multimedia, basado en estándares inalámbricos, San Salvador, El Salvador $http://isa.uniovi.es/docencia/SIGC/pdf/wifi2.pdf$.
9. Antonio Bernier Moreno, Víctor M. Vega García, Diego Martínez Lomas, Título: Seguridad Wifi, $https://dspace.ups.edu.ec/bitstream/123456789/7986/1/UPS-CT004855.pdf$.
10. Silvia Hernandez, Jose López Vicario, Xavi Vilajosana Guillen, (2020) Título: Estudio en detalle de Lorawan $http://openaccess.uoc.edu/webapps/o2/bitstream/10609/106369/7/shernandezc6TFM0120memoria.pdf$
11. Web oficial de Cup Carbon 2020, https://cupcarbon.com/.

3D to 2D Projections: A Preliminary Study of Dental Segmentation using Watershed Algorithm

J.D. Tamayo-Quintero[1], J.B. Gómez-Mendoza[1], and S.V. Guevara-Pérez[2]

[1] Universidad Nacional de Colombia, Manizales, Colombia.
{jdtamayoq, jbgomezm}@unal.edu.co
[2] Universidad Nacional de Colombia, Bogotá, Colombia.
svguevarap@unal.edu.co

Abstract. Teeth segmentation is an essential step in computer-aided orthodontic because dentists base their work on dental measurements to make the diagnosis. In this paper, we propose 3D to 2D projections to obtain an automatic teeth segmentation method based on the watershed technique. Our methodology allows projecting the three-dimensional digital models into six types of two-dimensional array: range image, accumulation image, and four-types of curvature images. The projections are used to slice the 3D dental image and created a 2D binary mask to employ as an input image of the watershed algorithm. The preliminary results show automatic segmentation using the watershed technique. However, the methodology needs refinement.

Keywords: 3D dental model · 3D to 2D projection · automatic segmentation · orthodontic

1 Introduction

Several areas of medicine are undergoing a change in the methodologies for the assisted and automated evaluation of systems. Such techniques aim for the optimization of diagnosis, treatment, and prognosis of pathologies [4,17,5]. A concrete case is Digital Dentistry that has adopted these new developments with the immersion of the CAD/CAM, reducing the complexity of the steps in the realization of crowns, prosthetic implants, and improving their efficiency and quality [3]. For example, in the United States, it was estimated that 19.000 CAD/CAM units were sold in 2017 and around 15 % of dentists used technology in digital dentistry [1].

Digital dentistry needs computer algorithms for orthodontic diagnostics and treatment that usually require 3D digital models. It often needs to extract, move, and rearrange teeth for simulating the treatment outcome [11]. In that context, teeth segmentation is an important tool in many automated dental software packages. The digital models of patients often have problems with teeth agglomeration, misalignment, irregular shapes, and difficult dental contours. In addition, some artifacts are added at the time of digitization what makes segmentation not an easy task [18].

Several segmentation techniques use surface curvature to identify dental boundaries [10,6]; principal curvatures, including minimum principal curvature [9,8]; and mean principal curvature [18].

A fully automatic algorithm for tooth segmentation from 3D digitized image captures by a laser scanner was introduced by Kondo *et al.* [7]. The author avoids the complexity of directly processing 3D using a projection to 2D in which he uses two range images:(i) to obtain the dental arch from the plan-view image, and (ii) to compute a panoramic image of the dental model based on the dental arch. The interstices between the teeth are detected separately in the two range images, and results from both views are combined for a determination of interstice locations and orientations.

In this work, we explore other alternatives to projection such as the curvature and accumulation to enrich the 2D data, and we believe that can be convenient for viewing and processing data using the entire large collection of image processing tools available.

This paper is structured as follows: In Section 2 we describe the basic structure behind our proposal. Section 3 presents the results achieved by applying the methodology on digital dental models. Finally, the conclusions of our work are shown in Section 4.

2 Materials and methods

Fig. 1 shows our implementation methodology. First of all, the 3D dental models are sliced to avoid irrelevant information (section 2.2). Second, the 3D information is projected on accumulation image, range image, and fourth types of curvature images (section 2.3). Third, we created a binary mask image using the enrich 2D information: the accumulation image is used to define borders, the range image to conserve depth information, and the curvature to preserve abrupt data changes. Finally, the watershed algorithm is applied to the binary image.

2.1 Data

The data used in this study were acquired using the Autoscan 3D EX scanner from four dental gypsum model and one ceramic bridge by the dental specialist in 3D Jean-Michel. (The data are available in `https://sketchfab.com/jean-michel3d`).

Table 2.1 shows five dental models (see Fig. 2). the first column shows four models correspond to dental digital models, and one digital dental bridge with 14 teeth. The second column shows the vertices, and the third column contains the triangles of each model.

2.2 3D slice

The digitized models from the plaster models are composed of two parts: the first part is the anatomy, and the second part is the gingiva, as seen in Fig. 3.

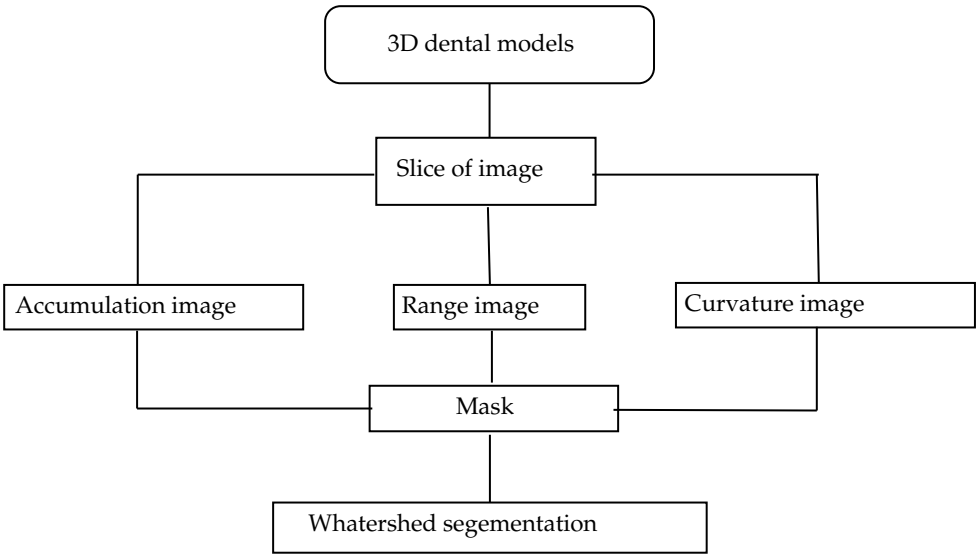


Fig. 1. Work-flow of our proposed methodology

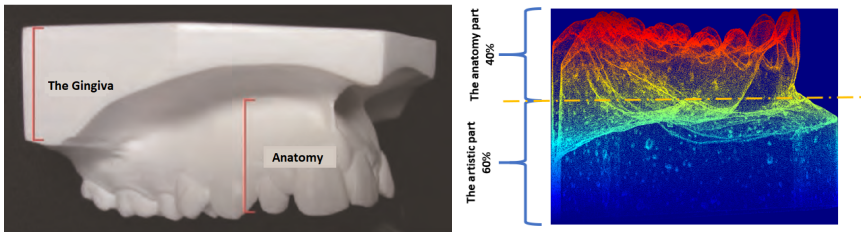
(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4 (e) Model 5

Fig. 2. Data acquired with AutoScan-DS-EX

Model	Vertices	Triangles
Model 1	334.7k	670.3k
Model 2	365.1k	730.6k
Model 3	160.6k	330.7k
Model 4	56.9k	114.1k
Model 5	123k	246k

Table 1. Data: five dental models

The gingiva part does not provide relevant information to orthodontists [13]; taking advantage of this particularity, we make a cut in the digitized model to eliminate irrelevant information.



(a) The Gypsum model (b) Digital dental model

Fig. 3. Dental model by slice

2.3 3D to 2D projections

Our method projects 3D point clouds to 2D images because processing a 2D image is much faster than processing the 3D points directly, as well, at the end of the process, we can reproject onto the 3D space. Also, 2D images are suitable structures for visualizing and processing data taking advantage of existing image processing techniques such as OpenCV, NumPy, PILLOW, scikit-image. Furthermore, Morphological Image Processing can be used. In our case study, we enrich 2D space with three types of projection: range images, accumulation image, and curvature images.

Range images a range image is a 2D structure that contains height information at each pixel. Range images are generated projecting 3D points to an XY-plane using a virtual camera placed on the horizontal plane with normal vector, and crossing the lowest point in the Z-coordinate axis $(0, 0, zmin)$. It projects several points on the same pixel (See Fig 4).

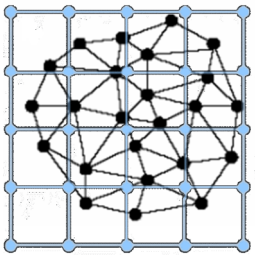
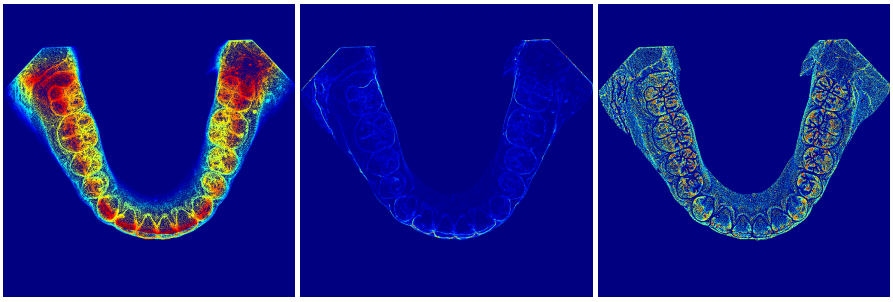


Fig. 4. Example of 3D to 2D projection using an unstructured data

Thus, we define the range image:

– Using the spatial average of the point distribution: stores the average depth among all projected points on the same pixel (Fig. 5(a)).

Other alternatives can be used such as maximum, minimum, or median. However, we choose the average because it is faster than the before techniques mentioned.



(a) Range image (b) Acumulation (c) Minimo Curv

Fig. 5. Images by slices

Accumulator point Accumulation image is a 2D structure that contains point density values at each pixel. Accumulator points are generated by projecting 3D points to the XY-plane as a range image, but in this case, the vertices in the same pixel or grid are accumulated and codify as pixel intensity. (Fig. 5(b)).

These accumulation image are used to define edges and contours. For example Spreitze [16] used similar methodology for precise mapping of large wood accumulations in fluvial systems and Serna [14] for segmentation and classification of 3D urban objects.

Curvature image A curvature image is a 2D structure that contains curvature information at each pixel. In this case, the curvature images are generated in a 3d space, projecting 3D $(x,\ y,\ Curv)$ to an XY-plane using a virtual camera located on the horizontal plane with normal vector, and crossing the lowest point in the Curv-coordinate axis (0, 0, $curvmin$); where $Curv$ is the curvature value at each point. In general, several $Curv$ can be applied. In this work, we used the vtkCurvatures function [2], which takes a polydata input and computes the curvature of the mesh at each point. Four possible methods are available:

Gauss Curvature: discrete Gauss curvature (K) computation

$$K(vertex\ v) = 2\pi - \sum_{\text{facet neighbs f of v}} (angle_f\ at\ v).$$

The contribution of every facet is for the moment weighted by $Area(facet)/3$. The curvatures estimation will lie in the range -1 to 0.2 with some large values like 20 to 40. Thus we need to enhance the values in range (-1 to 0.2).

Mean Curvature:
$H(vertex\ v) =$ average over edges neighbors e of $H(e)$

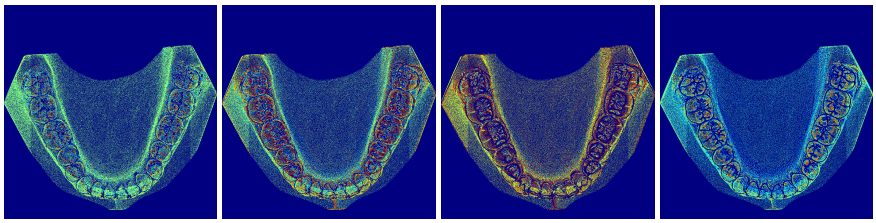
$$H(edge\ e) = length(e) * dihedral_angle(e)$$

NB:$dihedral_angle$ is the oriented angle between $-PI$ and PI; this means that the surface is assumed to be orientable, and the computation creates the orientation.

Maximum and Minimum:
Maximum ($k_\max$) and Minimum ($k_\min$) Principal Curvatures

$$k_\max = H + \sqrt{H^2 - K}$$
$$k_\min = H - \sqrt{H^2 - K}$$

Excepting spherical and planar surfaces which have equal principal curvatures, the curvature at a point on a surface varies with the direction one ´´sets off" from the point. For all directions, the curvature will pass through two extrema: a minimum ($k_\min$) and a maximum ($k_\max$) which occur at mutually orthogonal directions to each other.



(a) Gaus (b) Mean (c) Maximum (d) Minimum

Fig. 6. Examples: curvature images with spatial pixel size of 0.1mm

These Curvature images are used to define features in dental morphology like cusps, pits, and grooves [15,8,12].

Spatial resolution size

Spatial resolution is the size of the grid to be projected on a pixel and it is the only free available parameter during the projection (see Fig 7). It has to be carefully chosen. On one hand, if the pixel size is too large Fig 7(d) and (g), it does not preserve fine details because it would project multiple points on the same pixel. Too small pixel size implies connectivity difficulties and higher image sizes Fig 7 (b) and (e). To avoid connectivity obstacles and loss of information, the spatial resolution (i.e pixel size) determined based upon properties that lie in the data.

In our experiments, we analyze several options and we use the spatial resolution of the commercial digital caliper that corresponds to 0.1 mm and additionally with to 0.05 mm.

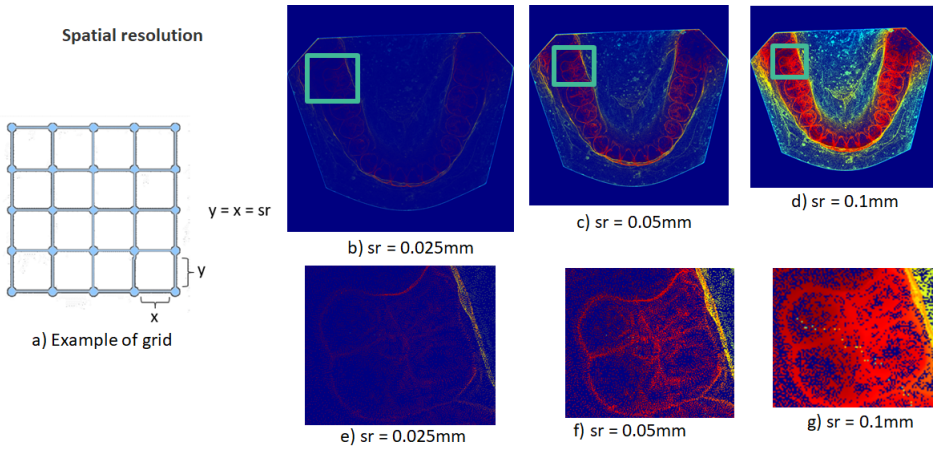


Fig. 7. Spatial resolution size

2.4 Mask

A mask is a binary image consisting of zero- and non-zero values. If a mask is applied to another binary or to a gray-scale image of the same size, all pixels which are zero in the mask are set to zero in the output image. The new image (mask) is used as input image of the watershed algorithm.

2.5 Watershed segmentation

A well-known technique to separate objects in contact in binary images uses the distance transformation and watershed method (see Fig 8).

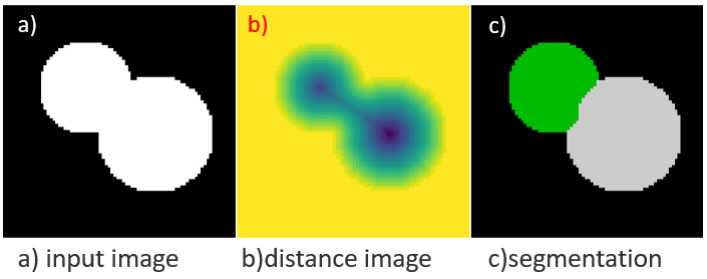


Fig. 8. A example of Watershed segmentation

The idea is to create boundaries at the edges of the overlapping objects. This strategy works very well on rounded objects. The watershed transformation

treats the image as a topographic map, with the brightness of each point representing its height, and finds the lines that run along the tops of ridges.

3 Results and Discussion

We tested our method using the data described before. An important parameter is the spatial resolution, which is the only free parameter. We simplify the selection of this parameter considering the square of the pixel assuming $x = y = sr$, where sr was 0.1 mm and 0.05 mm (see Fig 7). Also we tested different spacial resolutions in terms of time using $sr = 0.005$, $sr = 0.010$, $sr = 0.025$, $sr = 0.05$, $sr = 0.075$, $sr = 0.1$, $sr = 0.25$, $sr = 0.5$, $sr = 0.5$ and $sr = 1$ (Fig 9 and Fig. 10).

Table 2.1 show the data obtained with different spatial resolution size. The smaller sr has a critical effect on the calculation time: the larger the image, the slower the calculation. Note that the number of pixels increases proportionally to sr (Fig 9).

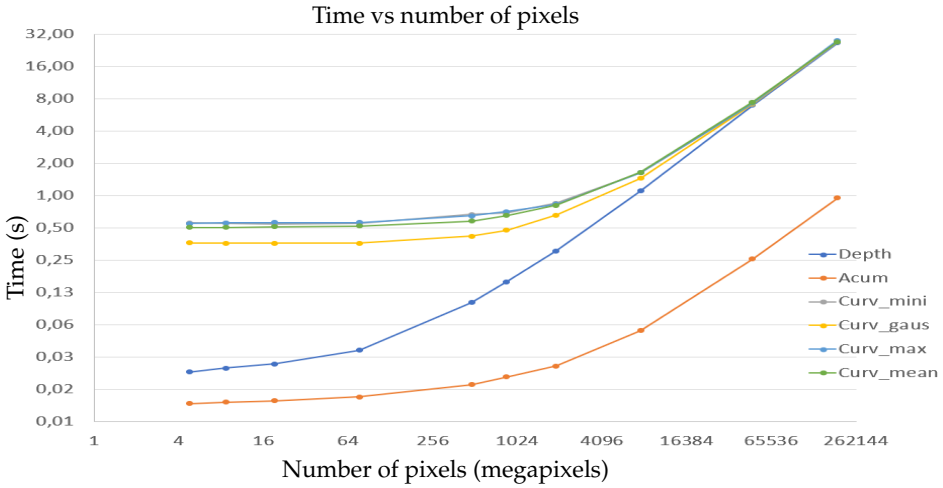


Fig. 9. Time VS pixels

Fig 7 correspond to a test used to select a spatial resolution. The idea is to choose a properly sr where fine details are preserved and neighbor connectivity. (b) and (e) presuppose connectivity problems (c) and (f) apparently it is the best choice to our case, and (d) and (g) are too small, fine details are not preserved.

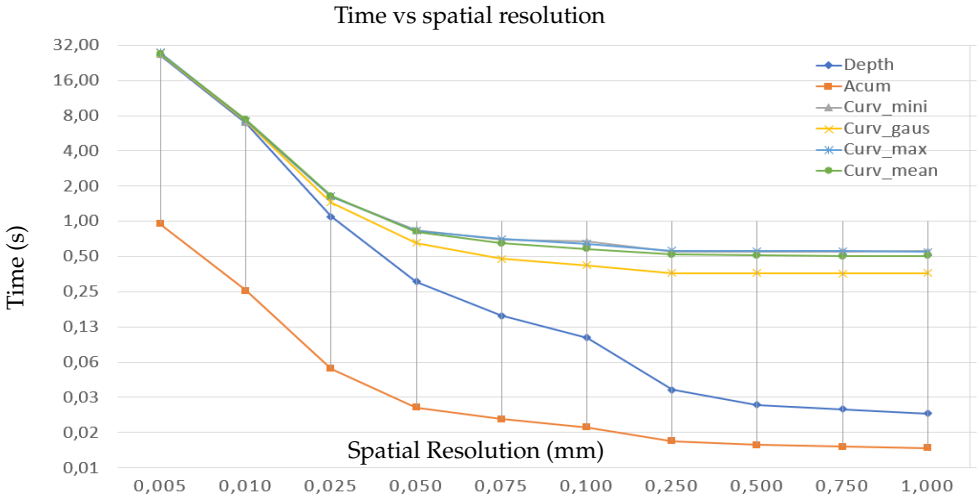


Fig. 10. Time Vs Spatial Resolution size

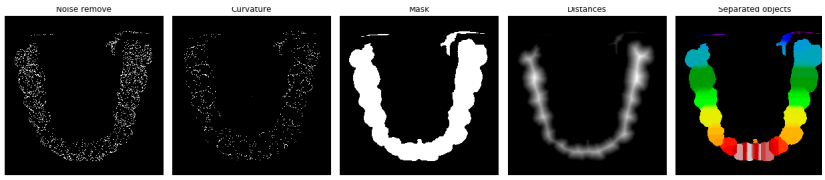
In Fig 9 and 10 the projection image sizes are 6949x6949 pixels for $sr = 0.01$, and 695x695 pixels for $sr = 0.1$. Note that large image presuppose connectivity problems Fig.7(b) and high computational cost as can be seen in the Fig.10. On the other hand small image avoid fine details Fig.7(g).

Resolution (px/mm)	1	1,33	2	4	10	13,33	20	40	100	200
Pixels	69	93	139	278	695	927	1390	2780	6949	13899
S. Resolution (sr)	1	0,75	0,5	0,25	0,1	0,075	0,05	0,025	0,01	0,005
Depth (seg)	0,023	0,025	0,027	0,036	0,101	0,156	0,303	1,100	6,903	26,386
Acum (seg)	**0,012**	**0,012**	**0,012**	**0,013**	**0,017**	**0,020**	**0,026**	**0,055**	**0,255**	**0,948**
Curv_mini (seg)	0,554	0,550	0,545	0,553	0,668	0,692	0,845	1,631	7,073	26,911
Curv_gaus (seg)	0,362	0,359	0,360	0,360	0,419	0,476	0,655	1,446	7,148	27,348
Curv_max (seg)	0,546	0,556	0,558	0,562	0,643	0,705	0,828	1,634	7,213	27,785
Curv_mean (seg)	0,504	0,503	0,512	0,520	0,578	0,649	0,811	1,648	7,375	26,932

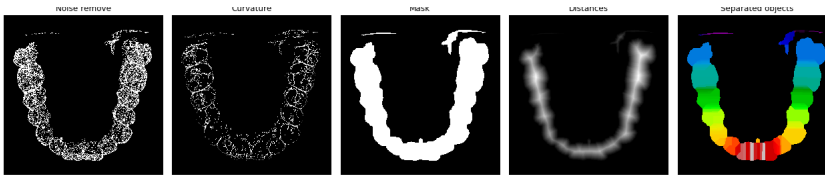
Table 2. Projection Performance - Time, Resolution, and Pixels: Experimental results were conducted 30 times to assess the consistency of results on a 2.6 GHz Core i9-9750H Intel Workstation MSI WE75 CPU with 32GB RAM, QUADRO RTX, the results shown are the average of the 30 data obtained.

Methodology

The results got with our method are shown below Fig. 11, we used two different spatial resolutions sr =0.05 mm and sr =0.1 mm. For the first case, the digital dental model (Fig.2 (a)) corresponds to an upper-jaw.



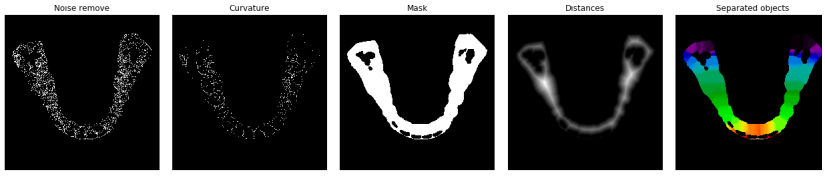
(a) Spatial resolution size $sr = 0.05$ mm



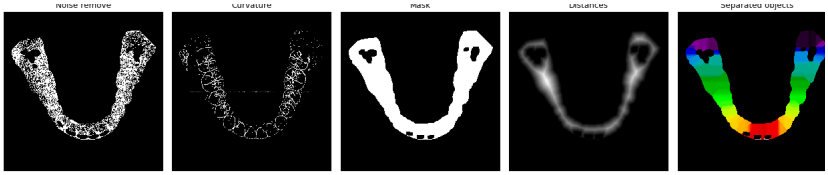
(b) Spatial resolution size $sr = 0.1$ mm

Fig. 11. Watershed segmentation: upper jaw spatial resolution size (a) $sr = 0.05$ mm (b) $sr = 0.1$ mm

Note that when the spatial resolution is greater ($sr = 0.1$ mm), the amount of neighborhood information increases. The two options work well, however, it is important to preserve the characteristics to perform later analyzes for example to detect cusps, pits, and other anatomical features.



(a) Spatial resolution size $sr = 0.05$ mm



(b) Spatial resolution size $sr = 0.1$ mm

Fig. 12. Watershed segmentation: lower-jaw spatial resolution size (a) $sr = 0.05$ mm (b) $sr = 0.1$ mm

The second case corresponds to the lower jaw Fig. 12, the results are not as good as the previous case and this is because of the curvature of the jaw (curve

of Spee). In (a) the spatial resolution $sr = 0.05$ mm shows worse results than in (b) and this is because of neighborhood problems. This preliminary result leads us to suggest a multi-scale approach where initial structures are detected at low scales using a fast approach and then a refinement may be carried out in large images to get fine details.

4 Conclusions

The proposed method presents a technique for the segmentation of digital dental models. The preliminary results attempt for the segmentation of the upper jaw and lower jaw. However, the methodology needs refinement. We believe that the segmentation of the lower jaw could be improved by using a non-orthographic projection based on the curve of the jaw.

In this work, we explore other alternatives to projection such as the curvature and accumulation to enrich the 2D data, and we believe that can be convenient for viewing and processing data.

During the projection, the spatial resolution size (sr) is the only free parameter. It controls the projection image size and must be carefully selected. one way to select the size is using a multi-scale approach where initial structures are detected at low scales using a fast approach (large sr), and then a refinement may be carried out in large images (small sr) to get more accurate results.

The gingiva part of the digital dental model has an impact since it can unnecessarily increase the projection time and it can introduce noise. To filter out not relevant information, we slice the 3d model into two slices, and the anatomical part are conserved

References

1. Afshari, F.S., Sukotjo, C., Alfaro, M.F., McCombs, J., Campbell, S.D., Knoernschild, K.L., Yuan, J.C.C.: Integration of digital dentistry into a predoctoral implant program: Program description, rationale, and utilization trends. Journal of dental education **81**(8), 986–994 (2017)
2. Avila, L., Kitware, I.: The VTK User's Guide. Kitware (2010), `https://books.google.com.co/books?id=6IxSewAACAAJ`
3. Cruz-Olivo, E.: Digital dentistry: The future is now. Revista Estomatología **25** (07 2018). https://doi.org/10.25100/re.v25i2.6484
4. Graber, L., Vanarsdall, R., Vig, K., Huang, G.: Orthodontics - E-Book: Current Principles and Techniques. Elsevier Health Sciences (2016), `https://books.google.com.co/books?id=N0SwDAAAQBAJ`
5. Gupta, R., Luthra, R., Kukreja, S.: Diagnostic aids used in prosthodontics-a review. global journal for research analysis **6**(10) (2018)

6. Ji, Z., Liu, L., Chen, Z., Wang, G.: Easy mesh cutting. In: Computer Graphics Forum. vol. 25, pp. 283–291. Wiley Online Library (2006)
7. Kondo, T., Ong, S.H., Foong, K.W.C.: Tooth segmentation of dental study models using range images. IEEE Transactions on Medical Imaging **23**(3), 350–362 (March 2004). https://doi.org/10.1109/TMI.2004.824235
8. Kronfeld, T., Brunner, D., Brunnett, G.: Snake-based segmentation of teeth from virtual dental casts. Computer-Aided Design and Applications **7**(2), 221–233 (2010)
9. Kumar, Y., Janardan, R., Larson, B., Moon, J.: Improved segmentation of teeth in dental models. Computer-Aided Design and Applications **8**(2), 211–224 (2011)
10. Lee, Y., Lee, S., Shamir, A., Cohen-Or, D., Seidel, H.P.: Mesh scissoring with minima rule and part salience. Computer Aided Geometric Design **22**(5), 444–465 (2005)
11. Liao, S.H., Liu, S.J., Zou, B.j., Ding, X., Liang, Y., Huang, J.h.: Automatic tooth segmentation of dental mesh based on harmonic fields. BioMed Research International **2015**, 1–10 (08 2015). https://doi.org/10.1155/2015/187173
12. Mairaj, D., Wolthusen, S.D., Busch, C.: Teeth segmentation and feature extraction for odontological biometrics. In: 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. pp. 323–328. IEEE (2010)
13. Mariaca, P., Garces, A., Trujillo, N., Ramirez, A., Florez, E., Liz, D.: Manual para realización de historia clínica odontológica del escolar. Universidad Cooperativa de Colombia (2007), https://books.google.com.co/books?id=89OZXQM067MC
14. Serna, A., Marcotegui, B.: Detection, segmentation and classification of 3d urban objects using mathematical morphology and supervised learning. ISPRS Journal of Photogrammetry and Remote Sensing **93** (07 2014). https://doi.org/10.1016/j.isprsjprs.2014.03.015
15. Smith, J.B., Vann, D.R., Dodson, P.: Dental morphology and variation in theropod dinosaurs: implications for the taxonomic identification of isolated teeth. The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology **285**(2), 699–736 (2005)
16. Spreitzer, G., Tunnicliffe, J., Friedrich, H.: Large wood (lw) 3d accumulation mapping and assessment using structure from motion photogrammetry in the laboratory. Journal of Hydrology **581**, 124430 (2020). https://doi.org/https://doi.org/10.1016/j.jhydrol.2019.124430, http://www.sciencedirect.com/science/article/pii/S0022169419311655
17. White, S.C., Pharoah, M.J.: The evolution and application of dental maxillofacial imaging modalities. Dental Clinics of North America **52**(4), 689 – 705 (2008). https://doi.org/https://doi.org/10.1016/j.cden.2008.05.006, http://www.sciencedirect.com/science/article/pii/S0011853208000487, contemporary Dental and Maxillofacial Imaging
18. Wu, K., Chen, L., Li, J., Zhou, Y.: Tooth segmentation on dental meshes using morphologic skeleton. Computers & Graphics **38**, 199–211 (2014)

3D to 2D Projections: A Preliminary Study of Dental Segmentation using Watershed Algorithm

J.D. Tamayo-Quintero[1], J.B. Gómez-Mendoza[1], and S.V. Guevara-Pérez[2]

[1] Universidad Nacional de Colombia, Manizales, Colombia.
{jdtamayoq, jbgomezm}@unal.edu.co
[2] Universidad Nacional de Colombia, Bogotá, Colombia.
svguevarap@unal.edu.co

Abstract. Teeth segmentation is an essential step in computer-aided orthodontic because dentists base their work on dental measurements to make the diagnosis. In this paper, we propose 3D to 2D projections to obtain an automatic teeth segmentation method based on the watershed technique. Our methodology allows projecting the three-dimensional digital models into six types of two-dimensional array: range image, accumulation image, and four-types of curvature images. The projections are used to slice the 3D dental image and created a 2D binary mask to employ as an input image of the watershed algorithm. The preliminary results show automatic segmentation using the watershed technique. However, the methodology needs refinement.

Keywords: 3D dental model · 3D to 2D projection · automatic segmentation · orthodontic

1 Introduction

Several areas of medicine are undergoing a change in the methodologies for the assisted and automated evaluation of systems. Such techniques aim for the optimization of diagnosis, treatment, and prognosis of pathologies [4,17,5]. A concrete case is Digital Dentistry that has adopted these new developments with the immersion of the CAD/CAM, reducing the complexity of the steps in the realization of crowns, prosthetic implants, and improving their efficiency and quality [3]. For example, in the United States, it was estimated that 19.000 CAD/CAM units were sold in 2017 and around 15 % of dentists used technology in digital dentistry [1].

Digital dentistry needs computer algorithms for orthodontic diagnostics and treatment that usually require 3D digital models. It often needs to extract, move, and rearrange teeth for simulating the treatment outcome [11]. In that context, teeth segmentation is an important tool in many automated dental software packages. The digital models of patients often have problems with teeth agglomeration, misalignment, irregular shapes, and difficult dental contours. In addition, some artifacts are added at the time of digitization what makes segmentation not an easy task [18].

Several segmentation techniques use surface curvature to identify dental boundaries [10,6]; principal curvatures, including minimum principal curvature [9,8]; and mean principal curvature [18].

A fully automatic algorithm for tooth segmentation from 3D digitized image captures by a laser scanner was introduced by Kondo *et al.* [7]. The author avoids the complexity of directly processing 3D using a projection to 2D in which he uses two range images:(i) to obtain the dental arch from the plan-view image, and (ii) to compute a panoramic image of the dental model based on the dental arch. The interstices between the teeth are detected separately in the two range images, and results from both views are combined for a determination of interstice locations and orientations.

In this work, we explore other alternatives to projection such as the curvature and accumulation to enrich the 2D data, and we believe that can be convenient for viewing and processing data using the entire large collection of image processing tools available.

This paper is structured as follows: In Section 2 we describe the basic structure behind our proposal. Section 3 presents the results achieved by applying the methodology on digital dental models. Finally, the conclusions of our work are shown in Section 4.

2 Materials and methods

Fig. 1 shows our implementation methodology. First of all, the 3D dental models are sliced to avoid irrelevant information (section 2.2). Second, the 3D information is projected on accumulation image, range image, and fourth types of curvature images (section 2.3). Third, we created a binary mask image using the enrich 2D information: the accumulation image is used to define borders, the range image to conserve depth information, and the curvature to preserve abrupt data changes. Finally, the watershed algorithm is applied to the binary image.

2.1 Data

The data used in this study were acquired using the Autoscan 3D EX scanner from four dental gypsum model and one ceramic bridge by the dental specialist in 3D Jean-Michel. (The data are available in `https://sketchfab.com/jean-michel3d`).

Table 2.1 shows five dental models (see Fig. 2). the first column shows four models correspond to dental digital models, and one digital dental bridge with 14 teeth. The second column shows the vertices, and the third column contains the triangles of each model.

2.2 3D slice

The digitized models from the plaster models are composed of two parts: the first part is the anatomy, and the second part is the gingiva, as seen in Fig. 3.

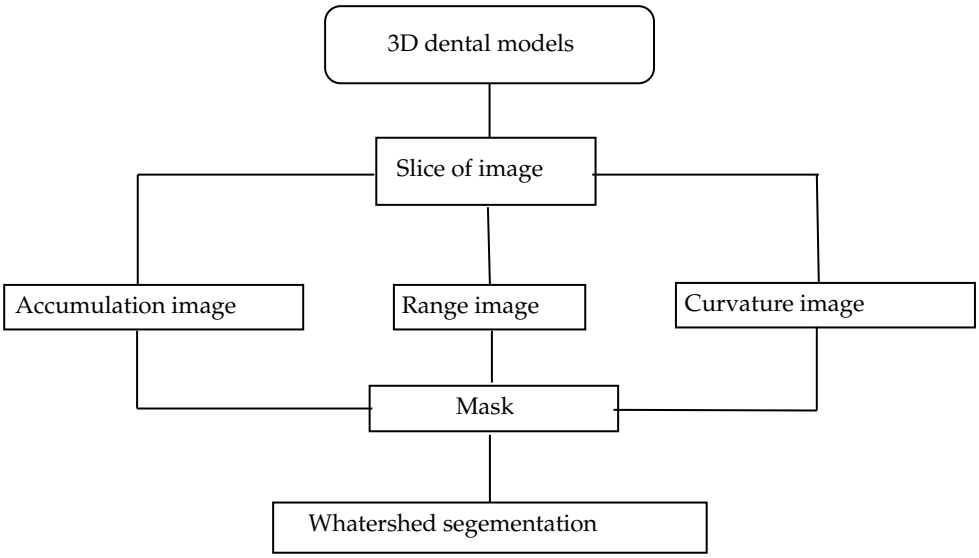


Fig. 1. Work-flow of our proposed methodology

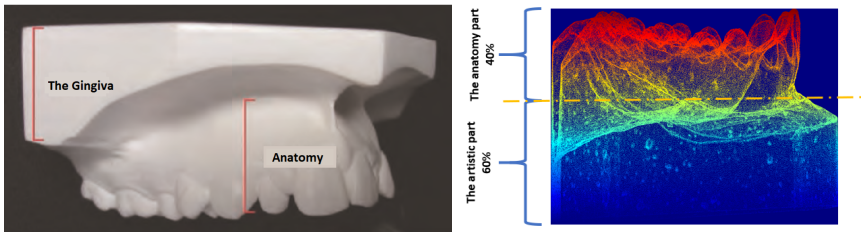
(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4 (e) Model 5

Fig. 2. Data acquired with AutoScan-DS-EX

Model	Vertices	Triangles
Model 1	334.7k	670.3k
Model 2	365.1k	730.6k
Model 3	160.6k	330.7k
Model 4	56.9k	114.1k
Model 5	123k	246k

Table 1. Data: five dental models

The gingiva part does not provide relevant information to orthodontists [13]; taking advantage of this particularity, we make a cut in the digitized model to eliminate irrelevant information.



(a) The Gypsum model (b) Digital dental model

Fig. 3. Dental model by slice

2.3 3D to 2D projections

Our method projects 3D point clouds to 2D images because processing a 2D image is much faster than processing the 3D points directly, as well, at the end of the process, we can reproject onto the 3D space. Also, 2D images are suitable structures for visualizing and processing data taking advantage of existing image processing techniques such as OpenCV, NumPy, PILLOW, scikit-image. Furthermore, Morphological Image Processing can be used. In our case study, we enrich 2D space with three types of projection: range images, accumulation image, and curvature images.

Range images a range image is a 2D structure that contains height information at each pixel. Range images are generated projecting 3D points to an XY-plane using a virtual camera placed on the horizontal plane with normal vector, and crossing the lowest point in the Z-coordinate axis $(0, 0, zmin)$. It projects several points on the same pixel (See Fig 4).

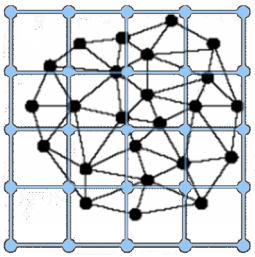
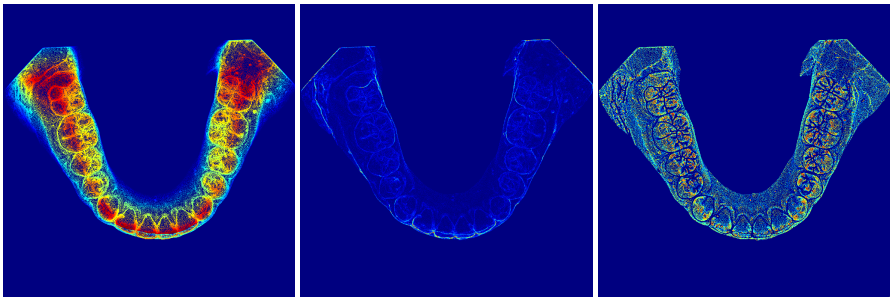


Fig. 4. Example of 3D to 2D projection using an unstructured data

Thus, we define the range image:

– Using the spatial average of the point distribution: stores the average depth among all projected points on the same pixel (Fig. 5(a)).

Other alternatives can be used such as maximum, minimum, or median. However, we choose the average because it is faster than the before techniques mentioned.



(a) Range image (b) Acumulation (c) Minimo Curv

Fig. 5. Images by slices

Accumulator point Accumulation image is a 2D structure that contains point density values at each pixel. Accumulator points are generated by projecting 3D points to the XY-plane as a range image, but in this case, the vertices in the same pixel or grid are accumulated and codify as pixel intensity. (Fig. 5(b)).

These accumulation image are used to define edges and contours. For example Spreitze [16] used similar methodology for precise mapping of large wood accumulations in fluvial systems and Serna [14] for segmentation and classification of 3D urban objects.

Curvature image A curvature image is a 2D structure that contains curvature information at each pixel. In this case, the curvature images are generated in a 3d space, projecting 3D $(x,\ y,\ Curv)$ to an XY-plane using a virtual camera located on the horizontal plane with normal vector, and crossing the lowest point in the Curv-coordinate axis (0, 0, $curvmin$); where $Curv$ is the curvature value at each point. In general, several $Curv$ can be applied. In this work, we used the vtkCurvatures function [2], which takes a polydata input and computes the curvature of the mesh at each point. Four possible methods are available:

Gauss Curvature: discrete Gauss curvature (K) computation

$$K(vertex\ v) = 2\pi - \sum_{\text{facet neighbs f of v}} (angle_f\ at\ v).$$

The contribution of every facet is for the moment weighted by $Area(facet)/3$. The curvatures estimation will lie in the range -1 to 0.2 with some large values like 20 to 40. Thus we need to enhance the values in range (-1 to 0.2).

Mean Curvature:
$H(vertex\ v) =$ average over edges neighbors e of $H(e)$

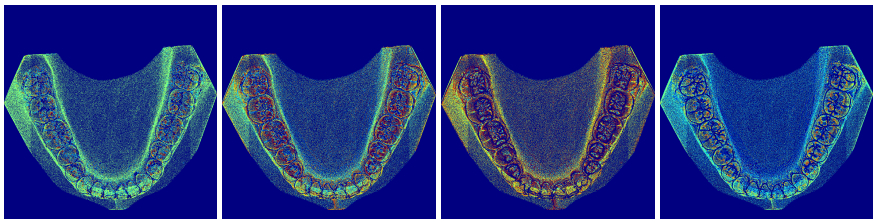
$$H(edge\ e) = length(e) * dihedral_angle(e)$$

NB:$dihedral_angle$ is the oriented angle between $-PI$ and PI; this means that the surface is assumed to be orientable, and the computation creates the orientation.

Maximum and Minimum:
Maximum ($k_\max$) and Minimum ($k_\min$) Principal Curvatures

$$k_\max = H + \sqrt{H^2 - K}$$
$$k_\min = H - \sqrt{H^2 - K}$$

Excepting spherical and planar surfaces which have equal principal curvatures, the curvature at a point on a surface varies with the direction one ´´sets off" from the point. For all directions, the curvature will pass through two extrema: a minimum ($k_\min$) and a maximum ($k_\max$) which occur at mutually orthogonal directions to each other.



(a) Gaus (b) Mean (c) Maximum (d) Minimum

Fig. 6. Examples: curvature images with spatial pixel size of 0.1mm

These Curvature images are used to define features in dental morphology like cusps, pits, and grooves [15,8,12].

Spatial resolution size

Spatial resolution is the size of the grid to be projected on a pixel and it is the only free available parameter during the projection (see Fig 7). It has to be carefully chosen. On one hand, if the pixel size is too large Fig 7(d) and (g), it does not preserve fine details because it would project multiple points on the same pixel. Too small pixel size implies connectivity difficulties and higher image sizes Fig 7 (b) and (e). To avoid connectivity obstacles and loss of information, the spatial resolution (i.e pixel size) determined based upon properties that lie in the data.

In our experiments, we analyze several options and we use the spatial resolution of the commercial digital caliper that corresponds to 0.1 mm and additionally with to 0.05 mm.

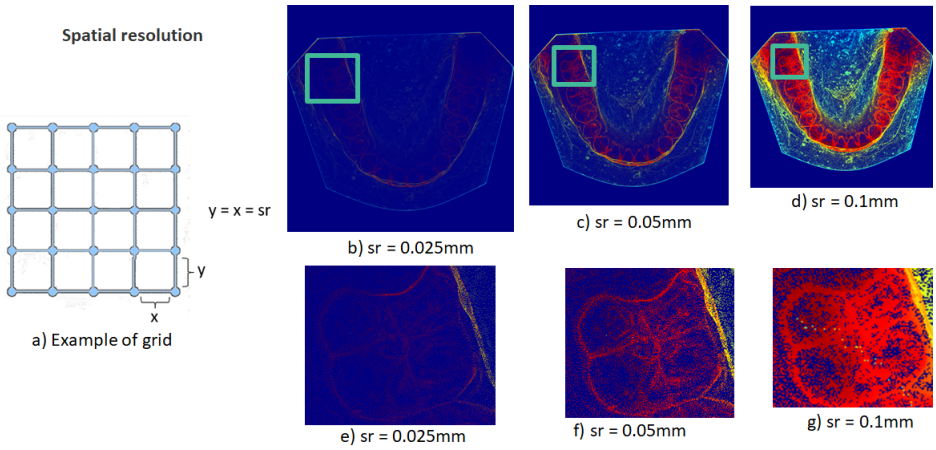


Fig. 7. Spatial resolution size

2.4 Mask

A mask is a binary image consisting of zero- and non-zero values. If a mask is applied to another binary or to a gray-scale image of the same size, all pixels which are zero in the mask are set to zero in the output image. The new image (mask) is used as input image of the watershed algorithm.

2.5 Watershed segmentation

A well-known technique to separate objects in contact in binary images uses the distance transformation and watershed method (see Fig 8).

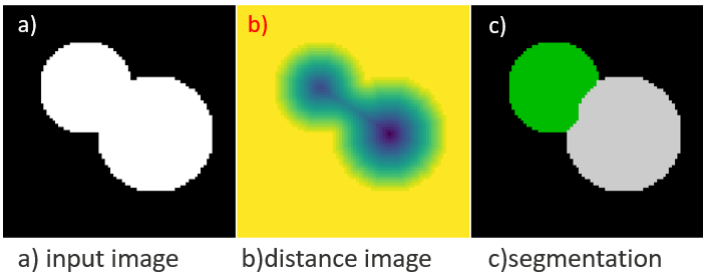


Fig. 8. A example of Watershed segmentation

The idea is to create boundaries at the edges of the overlapping objects. This strategy works very well on rounded objects. The watershed transformation

treats the image as a topographic map, with the brightness of each point representing its height, and finds the lines that run along the tops of ridges.

3 Results and Discussion

We tested our method using the data described before. An important parameter is the spatial resolution, which is the only free parameter. We simplify the selection of this parameter considering the square of the pixel assuming $x = y = sr$, where sr was 0.1 mm and 0.05 mm (see Fig 7). Also we tested different spacial resolutions in terms of time using $sr = 0.005$, $sr = 0.010$, $sr = 0.025$, $sr = 0.05$, $sr = 0.075$, $sr = 0.1$, $sr = 0.25$, $sr = 0.5$, $sr = 0.5$ and $sr = 1$ (Fig 9 and Fig. 10).

Table 2.1 show the data obtained with different spatial resolution size. The smaller sr has a critical effect on the calculation time: the larger the image, the slower the calculation. Note that the number of pixels increases proportionally to sr (Fig 9).

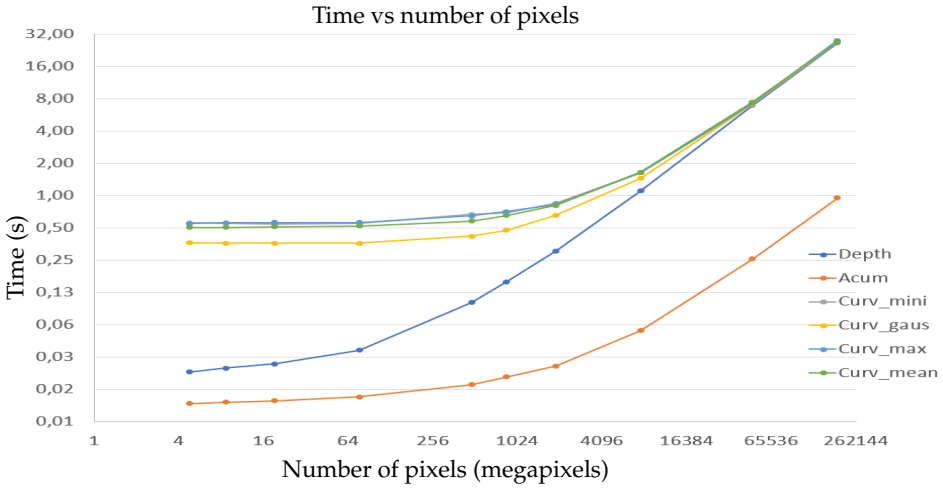


Fig. 9. Time VS pixels

Fig 7 correspond to a test used to select a spatial resolution. The idea is to choose a properly sr where fine details are preserved and neighbor connectivity. (b) and (e) presuppose connectivity problems (c) and (f) apparently it is the best choice to our case, and (d) and (g) are too small, fine details are not preserved.

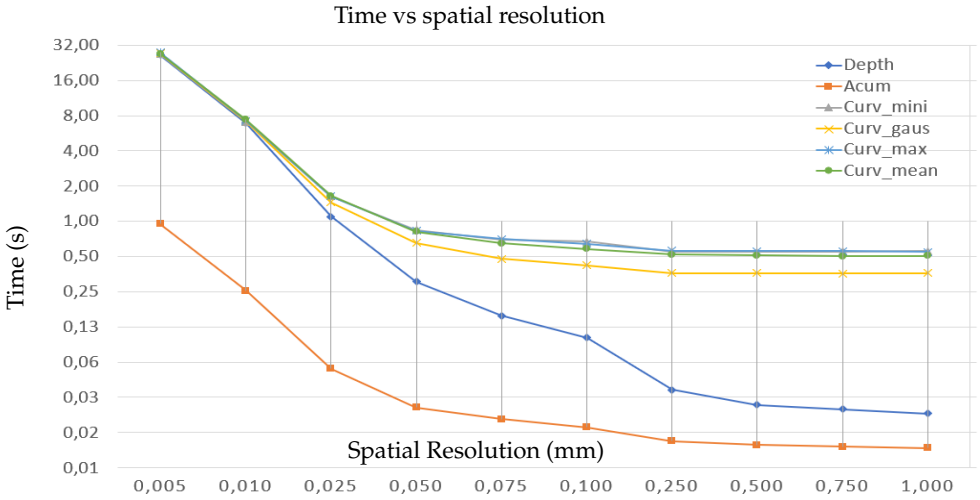


Fig. 10. Time Vs Spatial Resolution size

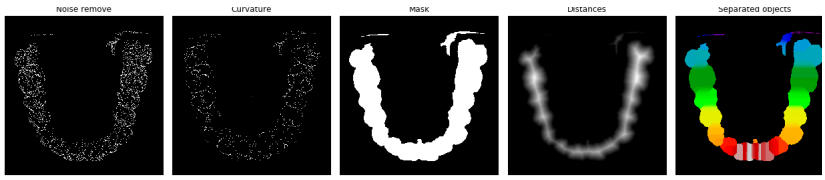
In Fig 9 and 10 the projection image sizes are 6949x6949 pixels for $sr = 0.01$, and 695x695 pixels for $sr = 0.1$. Note that large image presuppose connectivity problems Fig.7(b) and high computational cost as can be seen in the Fig.10. On the other hand small image avoid fine details Fig.7(g).

Resolution (px/mm)	1	1,33	2	4	10	13,33	20	40	100	200
Pixels	69	93	139	278	695	927	1390	2780	6949	13899
S. Resolution (sr)	1	0,75	0,5	0,25	0,1	0,075	0,05	0,025	0,01	0,005
Depth (seg)	0,023	0,025	0,027	0,036	0,101	0,156	0,303	1,100	6,903	26,386
Acum (seg)	**0,012**	**0,012**	**0,012**	**0,013**	**0,017**	**0,020**	**0,026**	**0,055**	**0,255**	**0,948**
Curv_mini (seg)	0,554	0,550	0,545	0,553	0,668	0,692	0,845	1,631	7,073	26,911
Curv_gaus (seg)	0,362	0,359	0,360	0,360	0,419	0,476	0,655	1,446	7,148	27,348
Curv_max (seg)	0,546	0,556	0,558	0,562	0,643	0,705	0,828	1,634	7,213	27,785
Curv_mean (seg)	0,504	0,503	0,512	0,520	0,578	0,649	0,811	1,648	7,375	26,932

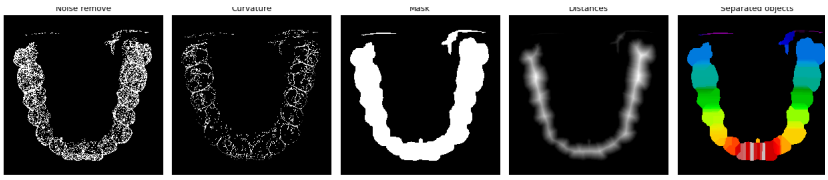
Table 2. Projection Performance - Time, Resolution, and Pixels: Experimental results were conducted 30 times to assess the consistency of results on a 2.6 GHz Core i9-9750H Intel Workstation MSI WE75 CPU with 32GB RAM, QUADRO RTX, the results shown are the average of the 30 data obtained.

Methodology

The results got with our method are shown below Fig. 11, we used two different spatial resolutions sr =0.05 mm and sr =0.1 mm. For the first case, the digital dental model (Fig.2 (a)) corresponds to an upper-jaw.



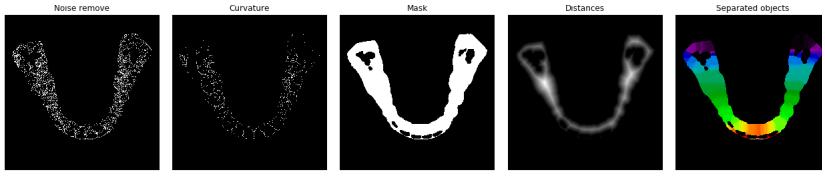
(a) Spatial resolution size $sr = 0.05$ mm



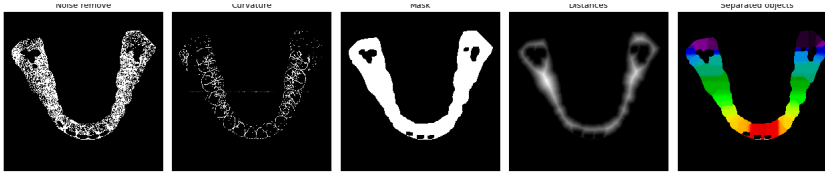
(b) Spatial resolution size $sr = 0.1$ mm

Fig. 11. Watershed segmentation: upper jaw spatial resolution size (a) $sr = 0.05$ mm (b) $sr = 0.1$ mm

Note that when the spatial resolution is greater ($sr = 0.1$ mm), the amount of neighborhood information increases. The two options work well, however, it is important to preserve the characteristics to perform later analyzes for example to detect cusps, pits, and other anatomical features.



(a) Spatial resolution size $sr = 0.05$ mm



(b) Spatial resolution size $sr = 0.1$ mm

Fig. 12. Watershed segmentation: lower-jaw spatial resolution size (a) $sr = 0.05$ mm (b) $sr = 0.1$ mm

The second case corresponds to the lower jaw Fig. 12, the results are not as good as the previous case and this is because of the curvature of the jaw (curve

of Spee). In (a) the spatial resolution $sr = 0.05$ mm shows worse results than in (b) and this is because of neighborhood problems. This preliminary result leads us to suggest a multi-scale approach where initial structures are detected at low scales using a fast approach and then a refinement may be carried out in large images to get fine details.

4 Conclusions

The proposed method presents a technique for the segmentation of digital dental models. The preliminary results attempt for the segmentation of the upper jaw and lower jaw. However, the methodology needs refinement. We believe that the segmentation of the lower jaw could be improved by using a non-orthographic projection based on the curve of the jaw.

In this work, we explore other alternatives to projection such as the curvature and accumulation to enrich the 2D data, and we believe that can be convenient for viewing and processing data.

During the projection, the spatial resolution size (sr) is the only free parameter. It controls the projection image size and must be carefully selected. one way to select the size is using a multi-scale approach where initial structures are detected at low scales using a fast approach (large sr), and then a refinement may be carried out in large images (small sr) to get more accurate results.

The gingiva part of the digital dental model has an impact since it can unnecessarily increase the projection time and it can introduce noise. To filter out not relevant information, we slice the 3d model into two slices, and the anatomical part are conserved

References

1. Afshari, F.S., Sukotjo, C., Alfaro, M.F., McCombs, J., Campbell, S.D., Knoernschild, K.L., Yuan, J.C.C.: Integration of digital dentistry into a predoctoral implant program: Program description, rationale, and utilization trends. Journal of dental education **81**(8), 986–994 (2017)
2. Avila, L., Kitware, I.: The VTK User's Guide. Kitware (2010), `https://books.google.com.co/books?id=6IxSewAACAAJ`
3. Cruz-Olivo, E.: Digital dentistry: The future is now. Revista Estomatología **25** (07 2018). https://doi.org/10.25100/re.v25i2.6484
4. Graber, L., Vanarsdall, R., Vig, K., Huang, G.: Orthodontics - E-Book: Current Principles and Techniques. Elsevier Health Sciences (2016), `https://books.google.com.co/books?id=N0SwDAAAQBAJ`
5. Gupta, R., Luthra, R., Kukreja, S.: Diagnostic aids used in prosthodontics-a review. global journal for research analysis **6**(10) (2018)

6. Ji, Z., Liu, L., Chen, Z., Wang, G.: Easy mesh cutting. In: Computer Graphics Forum. vol. 25, pp. 283–291. Wiley Online Library (2006)
7. Kondo, T., Ong, S.H., Foong, K.W.C.: Tooth segmentation of dental study models using range images. IEEE Transactions on Medical Imaging **23**(3), 350–362 (March 2004). https://doi.org/10.1109/TMI.2004.824235
8. Kronfeld, T., Brunner, D., Brunnett, G.: Snake-based segmentation of teeth from virtual dental casts. Computer-Aided Design and Applications **7**(2), 221–233 (2010)
9. Kumar, Y., Janardan, R., Larson, B., Moon, J.: Improved segmentation of teeth in dental models. Computer-Aided Design and Applications **8**(2), 211–224 (2011)
10. Lee, Y., Lee, S., Shamir, A., Cohen-Or, D., Seidel, H.P.: Mesh scissoring with minima rule and part salience. Computer Aided Geometric Design **22**(5), 444–465 (2005)
11. Liao, S.H., Liu, S.J., Zou, B.j., Ding, X., Liang, Y., Huang, J.h.: Automatic tooth segmentation of dental mesh based on harmonic fields. BioMed Research International **2015**, 1–10 (08 2015). https://doi.org/10.1155/2015/187173
12. Mairaj, D., Wolthusen, S.D., Busch, C.: Teeth segmentation and feature extraction for odontological biometrics. In: 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. pp. 323–328. IEEE (2010)
13. Mariaca, P., Garces, A., Trujillo, N., Ramirez, A., Florez, E., Liz, D.: Manual para realización de historia clínica odontológica del escolar. Universidad Cooperativa de Colombia (2007), https://books.google.com.co/books?id=89OZXQM067MC
14. Serna, A., Marcotegui, B.: Detection, segmentation and classification of 3d urban objects using mathematical morphology and supervised learning. ISPRS Journal of Photogrammetry and Remote Sensing **93** (07 2014). https://doi.org/10.1016/j.isprsjprs.2014.03.015
15. Smith, J.B., Vann, D.R., Dodson, P.: Dental morphology and variation in theropod dinosaurs: implications for the taxonomic identification of isolated teeth. The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology **285**(2), 699–736 (2005)
16. Spreitzer, G., Tunnicliffe, J., Friedrich, H.: Large wood (lw) 3d accumulation mapping and assessment using structure from motion photogrammetry in the laboratory. Journal of Hydrology **581**, 124430 (2020). https://doi.org/https://doi.org/10.1016/j.jhydrol.2019.124430, http://www.sciencedirect.com/science/article/pii/S0022169419311655
17. White, S.C., Pharoah, M.J.: The evolution and application of dental maxillofacial imaging modalities. Dental Clinics of North America **52**(4), 689 – 705 (2008). https://doi.org/https://doi.org/10.1016/j.cden.2008.05.006, http://www.sciencedirect.com/science/article/pii/S0011853208000487, contemporary Dental and Maxillofacial Imaging
18. Wu, K., Chen, L., Li, J., Zhou, Y.: Tooth segmentation on dental meshes using morphologic skeleton. Computers & Graphics **38**, 199–211 (2014)

Sistema de Detección y Reconocimiento Facial Mediante Algoritmos de Visión Computacional

Emiliano Gomez Blotta, Rosana Giménez, Miguel Méndez Garabetti y Pablo Sández

Laboratorio de Investigación en Ciencia y Tecnología (LabICyT-UDA)
Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua
Catamarca 147, Mendoza, Argentina
emi.gblotta@gmail.com

Resumen: Debido al contexto actual por la pandemia del COVID-19, se han adaptado algunas tecnologías para que las personas tengan el menor contacto posible entre ellas. Es por eso que en este informe se detalla un sistema de reconocimiento facial adaptado a aquellas empresas que cuentan con un sistema de confirmación asistencial al momento de que sus empleados ingresan o egresan del establecimiento. En esta etapa, solamente se abarcará un estudio preliminar sobre el reconocimiento facial que posteriormente se utilizará en un sistema que cumpla con el objetivo planteado.

Keywords: Reconocimiento facial · Pandemia · Vision computacional.

1 Introduccion

Debido a la pandemia por el COVID-19 que esta atravesando el mundo, se han adaptado nuevas formas de interacción entre las personas, sobre todo en el ámbito laboral, donde muchas empresas poseen dispositivos de ingreso y egreso de empleados a través de la huella digital, por ejemplo, comparando la huella dactilar contra una base de datos donde los datos se encuentran almacenados y permite con una precisión muy alta corroborar a quien corresponde cada huella. Es por eso que en este artículo presentamos una propuesta mediante la utilización de otra tecnología con una precisión similar a través del uso del reconocimiento facial, evitando así el contacto entre personas por medio del panel de detección de huellas.

Dicha propuesta hace uso de un software de reconocimiento facial utilizando vision artificial y redes neuronales artificiales que detectan el rostro de la persona y comparan el resultado con los datos almacenados en una base de datos donde ya se encuentra registrada la persona con su respectiva imagen.

El trabajo se organiza de la siguiente manera. La sección 2 describe brevemente el reconocimiento facial, distinguiendo entre reconocimiento facial y detección facial. La sección 3 detalla la librería de software Open CV. La sección 4 detalla la librera Face Recognition. La sección 5 muestra el funcionamiento del programa con los resultados obtenidos. Finalmente, la sección 6 describe la conclusión final del artículo.

2 Reconocimiento y detección facial

Antes de hablar del funcionamiento de los algoritmos, hay que diferenciar entre reconocimiento facial y detección facial. El reconocimiento facial es una técnica utilizada por computadoras que sirven para identificar automáticamente el rostro de una persona en una imagen digital como vemos en la figura 2. Esto se logra mediante el análisis de las características faciales de un sujeto, que se extraen de una imagen o de una fuente de video, y luego se comparan dichas características con una base de datos. Detección facial, en cambio, es reconocer el rostro de una persona como vemos en la figura 4.

3 Librería de visión artificial OpenCV

OpenCV es una librería para reconocimiento facial desarrollada en el lenguaje C, fue pensado para reconocer objetos en concreto. Utiliza el algoritmo de Viola-Jones[2], usado ampliamente para la detección de rostros en imágenes y video. Este algoritmo tiene un coste computacional bajo, y está compuesto por dos partes principales: un clasificador en cascada y un entrenador de clasificadores.

3.1 Clasificador en cascada

El clasificador en cascada o clasificar de Haar es un método eficaz de detección de objetos. Se trata de los elementos básicos para la detección, en el que el clasificador se entrena mediante muchas imágenes, tantas positivas (imágenes con rostro), como negativas (imágenes sin rostro). En las imágenes se buscan características muy simples, que consisten en la diferencia entre intensidades luminosas, ya que cada característica es un valor individual obtenido restando la suma de pixeles bajo el rectángulo blanco de la suma de pixeles bajo el rectángulo negro, como en la siguiente imagen:

En el trabajo de Viola-Jones[2], las características se definen sobre una ventana de búsqueda básica de 24x24 píxeles, lo que da lugar a más de 160000 características posibles. Para cada cálculo de características, necesitamos encontrar la suma de pixeles bajo rectángulos blancos y negros. Como el número de características es muy grande, se introducen imágenes integrales, que simplifican el cálculo de suma de pixeles a una operación que solo involucra cuatro pixeles, de esta manera se agiliza el proceso. A pesar de reducir las características, tenemos el problema de que muchas de ellas son irrelevantes, es por eso que el algoritmo usa Adaboost, que es un meta-algoritmo de machine learning, que mejora el rendimiento de los algoritmos de búsqueda, en este caso el de reconocimiento facial.

En este caso, ya contamos con diversos clasificadores en formato XML que la misma librería nos provee[7], en este artículo se utilizará el clasificador facial frontal, de esta manera, el algoritmo devolverá las posiciones de las caras detectadas.

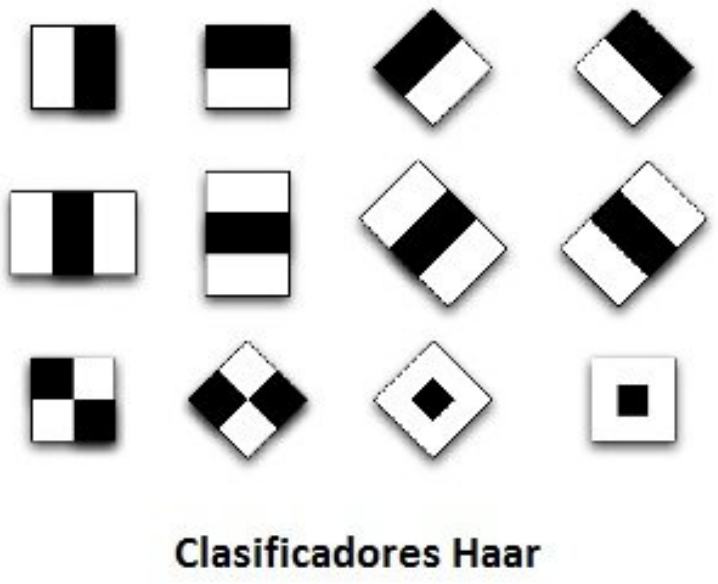


Fig. 1. Clasificador de Haar

4 Librería de reconocimiento Face recognition

Face Recognition es una librería de reconocimiento facial, desarrollada para funcionar con el lenguaje de programación Python, creada con Deep Learning, un proceso de Machine Learning, que mediante redes neuronales artificiales y compuesta por un numero de niveles jerárquicos, va ajustando sus pesos mediante un jerárquico, donde en el nivel inicial, la red aprende algo simple y luego pasa este aprendizaje al siguiente nivel. El siguiente nivel recibe esta información sencilla, la combina, compone una información un poco más compleja, y se lo pasa al siguiente nivel, y así sucesivamente hasta llegar a una solución lo más acertada posible.

5 Experimentación

Cabe recordar que solamente se mostrará el estudio preliminar en reconocimiento facial para despues aplicar dicho estudio al sistema mencionado en el objetivo. Para demostrar el funcionamiento del programa se utilizaron imágenes de muestra propias, las cuales el programa debería detectar los rostros para demostrar su correcto funcionamiento.

Como vemos en la imagen, el software tuvo un acierto del 75 por ciento teniendo en cuenta que detectó 3 de 4 rostros disponibles.

Fig. 2. Detección facial en una imagen fija

El siguiente paso fue demostrar el funcionamiento mediante la cámara web de la computadora, la cual es un acercamiento al objetivo del trabajo, detectar un rostro en tiempo real a través de una cámara.

El próximo paso es identificar a las personas que aparecen en las imágenes, para así acercarse más al objetivo final de este artículo, que es la detección de personas en tiempo real. Si el programa reconoce un rostro, aparecerá un rectángulo verde con el nombre de dicha persona, caso contrario aparecerá un rectángulo rojo. Para que el software de reconocimiento funcione de manera óptima, es necesario tener las imágenes de las personas a identificar etiquetadas y almacenadas en una base de datos.

El resultado es un poco diferente en cuanto a la imagen original, donde se puede observar que la imagen tiene cierto matiz azulado. Esto se debe a que OpenCV y matplotlib tienen diferentes órdenes de colores primarios. OpenCV lee imágenes en forma de BGR, matplotlib lo hace en orden RGB. Como podemos ver, el programa junto con las librerías trabaja de forma óptima, dando resultados 100 por ciento positivos, en este caso, identificando a todas las personas que se encuentran en la imagen.

6 Conclusiones

En este trabajo se demostró el funcionamiento de los algoritmos para detección y reconocimiento facial OpenCV y Face Recognition que pueden ser implementados tanto en imágenes como en video digital. Los resultados obtenidos nos demuestran que este software funciona casi de manera optima y aplicado a otros

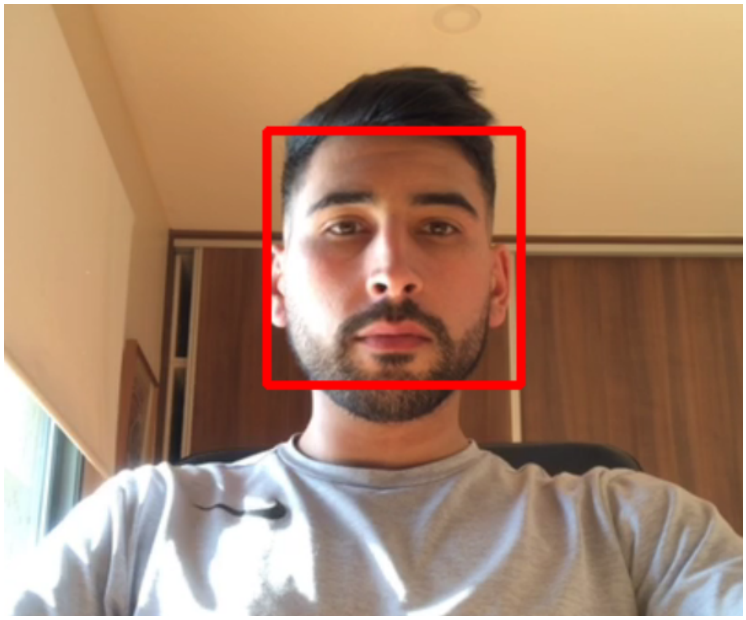


Fig. 3. Detección facial en video digital

Fig. 4. Reconocimiento facial de la figura 2

sistemas puede ser de gran ayuda y con un coste económico bajo. En este escenario solo se hizo pruebas con imágenes donde evidencia que funciona bien. Tambien se hicieron pruebas sobre un video en tiempo real y también se obtuvieron respuestas positivas. Por este motivo, se planea realizar como trabajo futuro un programa de detección y reconocimiento fácil en el ámbito laboral, con el objetivo de minimizar el contacto entre personas y agilizar los ingresos y egresos de empleados.

Referencias

1. Stan Z. Li.: Handbook of Face Recognition. (2011)
2. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. ACCEPTED CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, Location (2001)
3. Ray: El algoritmo de Viola y Jones (Diario de un programador). Blog de Programacion Extrema. (2018)
4. ROBOLOGS Homepage, http://https://robologs.net. Last accessed 5 Set 2020
5. INESEM Revista digital, https://revistadigital.inesem.es/informatica-y-tics/opencv/. Last accessed 5 Set 2020
6. Unipython Blog, https://unipython.com/deteccion-rostros-caras-ojos-haar-cascad/. Last accessed 5 Set 2020
7. OpenCV Homepage, https://opencv.org/

Biometría como Comprobación de Identidades

Facundo Collado, Giménez Rosana, Sandez Pablo Javier y Mendez Garabetti Miguel.

Laboratorio de Investigación en Ciencia y Tecnología (LabICyT-UDA).
Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua,
Catamarca 147, Mendoza, Argentina
colladofacu540@gmail.com

Abstract. La presente investigación aborda la biometría informática, conceptos básicos, algunas técncias y su aplicación en la seguridad. La biometría actualmente es una de las ciencias más importantes y provee una solución y confiabilidad óptima a la hora de preservar el acceso a distintos tipos de información, tanto lugares físicos y también sistemas. En este trabajo se dará a conocer las tencologías biométricas y dará pie a la futura implementación de una aplicación de comprobación de identidad mediante reconocimiento facial o de voz que pueda ser accesible a más personas y ámbitos.

Keywords: Biometría · Seguridad · Face ID.

1 Introducción

En la actualidad la preservación de los datos es una tareas más importantes en el mundo de la informática, ya que estos son fundamentales en cualquier institución, empresa e incluso personales. Estos datos tienen que tener su privacidad y quien pueda acceder a ellos para ya sea su manipulación o uso de los mismos es fundamental. Existen distintos metodos de asegurar la privacidad y el uso adecuado de la información, dentro de los cuales aparece la biometría que esta tiene errores de seguridad casi nulos, asegurando que el acceso y seguridad de los recursos esté bien resguardada [6].

Suele suceder que una persona hackee o tenga previo conocimiento de una contraseña, falsifique una firma, robe o imite tarjetas de acceso, etc. con el objetivo de acceder a zonas o sistemas restringidos [4].

Los sistemas biométricos son una solución ante la necesidad de poder asegurar de manera confiable y segura la información, estos se basan en obtener mediante patrones características físicas de una persona, tales como su rostro, voz, huella dactilar, ojos, etc.

La detección de rostros es una de las más importantes y más usada, gracias a su comodidad de instalación (cámara) y es muy aceptado por el público en general. Esta llega a ser muy segura y confiable al momento de identificar distintas personas dando una gran confiabilidad.

A lo largo de este trabajo se abordará la creación de una aplicación que sea posible de identificar distintos rostros o voces, como comprobación de identidad.

2 La Biometria

La biometría es un método de reconocimiento de personas basado en sus características fisiológicas (iris y retina, huella, rostro, geometría de la mano e imagen vascular) o de comportamiento (voz, firma, gesto y movimiento corporal) . Estos rasgos o cualidades son físicos inalterables y presentes en todas las personas, los cuales se van a ver traducidos como patrones tales como. Es un proceso similar al que realizamos los humanos reconociendo a otros mediante su aspecto físico, voz, forma de caminar, etc [1]. Cualquier característica fisiológica o de conducta de una persona puede ser una biometría, siempre que tenga las siguientes propiedades: **universalidad**: cada persona debe tener la característica, **unicidad**: indica que dos personas no deben ser iguales en términos de la caracterísitica, la **permanencia**: significa que la característica debe ser invariable al tiempo, **capacidad de recolección**: indica que la característica puede ser medida cuantitativamente [2], **rendimiento o actuación**: es decir un elevado nivel de exactitud, **aceptación** por parte del usuario y **resistencia al fraude** o usurpación [3].

3 Características y funcionamiento

Modo de Operación

Para que un sistema biométrico pueda ser usado se debe poder registrar la identidad de los individuos mediante la captura de sus parámetros biométricos o patrones. A este proceso se le denomina proceso de registro, se compone de tres fases distintas, como podemos observar en la Figura 1:

Fig. 1. Proceso de registro.**Captura** de los patrones. **Procesamiento** se crea una plantilla con las características personas de los patrones capturados. **Inscripción** de la plantilla procesada y guardando los datos. Una vez que finaliza la inscripción, el sistema ya está disponible para poder autentificar a las personas mediante el uso de la plantilla [1].

Una vez la plantilla se ha registrado con los datos biométricos de los individuos, mediante un **proceso de autenticación** se comparará con la información previamente guardada.

Verificación: (¿quién es?) Se compara la entrada con todos los patrones guardados de los distintos usuarios del sistema.
Autenticación: (¿quién dice ser?) Se compara la entrada con el patrón del usuario quien dice ser [3].

Característica	*Aceptación del usuario*	*Facilidad de uso*	*Coste*	*Utilidad*		*Estabilidad*	*Intrusismo*	*Fiabilidad*
				Identificación	*Verificación*			
ADN	Baja	Baja	Alto	✓	✓	Alta	Muy alto	Alta
Dinámica de escritura	Alta	Alta	Bajo	✗	✓	Baja	No	Baja
Firma	Media	Alta	Bajo	✗	✓	Media	No	Baja
Geometría de la mano	Media	Alta	Alto	✗	✓	Media	No	Media
Huella dactilar	Media	Alta	Bajo	✓	✓	Alta	Bajo	Alta
Iris	Media	Media	Alto	✓	✓	Alta	No	Alta
Reconocimiento facial	Media	Media	Bajo	✗	✓	Media	No	Media
Retina	Media	Baja	Alto	✓	✓	Alta	Alto	Alta
Voz	Alta	Alta	Bajo	✗	✓	Media	No	Baja

Característica	*Nivel de seguridad*	*Ratio de error*	*Precisión*	*Errores*	*Falso positivo*	*Falso negativo*
ADN	Alto	Sin datos	4	No conocidos.	5	5
Dinámica de escritura	Medio	Sin datos	1	Lesiones de mano, cansancio.	3	1
Firma	Medio	1/50	2	Cambios de escritura.	2	1
Geometría de la mano	Medio	1/500	3	Edad, lesiones varias.	4	2
Huella dactilar	Alto	1/500+	4	Sequedad, suciedad, edad.	5	5
Iris	Alto	1/131000	4	Iluminación inadecuada.	4	4
Reconocimiento facial	Medio	Sin datos	3	Pelo, gafas, edad, iluminación.	3	1
Retina	Alto	$1/10^6$	4	Gafas, lentillas.	5	5
Voz	Medio	1/50	2	Ruidos, ronquera, resfriados…	2	1

Fig. 2. Tabla de comparativa de ejemplos biométricos y sus características[3].

Por otro lado, los **sistemas de identificación biométrica** se dividen en:

- **Dinámicos**: Utilizan tecnologías de comportamiento que comparan acciones o movimientos.
- **Estáticos**: Utilizan tecnologías fisiológicas que miden y comparan rasgos físicos.
- **Multimodales**: Combinan técnicas estáticas y dinámicas.

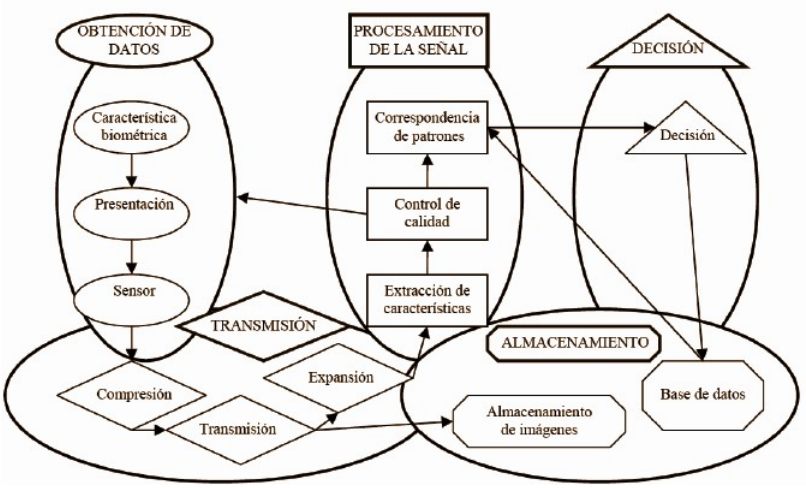


Fig. 3. Bloques y etapas de un sistema biométrico [3].

4 Tecnologías biométricas

4.1 Huella dactilar

Las huellas dactilares son crestas gráficas en forma de flujo presentes en los dedos humanos [2]. En los sistemas biométricos es la técnica más utilizada para la autenticación, usándose desde finales del siglo IXX en el campo forense [4].Es de las más utilizadas debido a que las huellas dactilares son únicas e inalterables. Además, sus ventajas son la alta tasa de precisión y facilidad de uso [1].

Generalmente, la imagen de una huella dactilar se captura de una de dos formas: escaneando una impresión entintada de un dedo o usando un escáner de huellas dactilares de escaneo en vivo (Figura 5).

Existen dos tipos de técnicas de búsqueda de coincidencias entre muestras de huella dactilar:
Basadas en minucias: Esta técnnica se basa en la autenticación de las minucias[Figura 6], es decir, en determinadas formas existentes en la huella dactilar. Así el modelo o plantilla correspondiente a cada usuario es un esquema en el que las minucias son detectadas, su posicion y las distancias que separan unas de otras.
Basadas en correlación: Mediante esta técnica se analiza el patrón

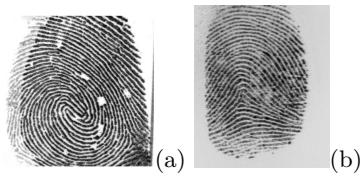

(a) (b)

Fig. 4. Una huella puede ser capturada por la presión en tinta(a) o imaginando directamente un dedo usando la frustrada tecnología de reflexión interna total(b). A esta última se le llama huella digital de escaneo en vivo [1].

global seguido por la huella dactilar, es decir, el esquema general del conjunto de la huella en lugar de las minucias [1]. **Ventajas**:

- Técnica muy desarrollada
- Gran aceptación (para determinadas aplicaciones)
- Sistemas de captura no invasivos y de bajo coste con posibilidad de incorporar fácilmente un sistema de detección de sujeto vivo
- Facilidad de integración en diferentes entornos
- Unicidad y estabilidad de la huella

Desventajas:

- Implicaciones policiales y judiciales
- Necesidad de elevada calidad de la imagen digital
- Necesidad de contacto físico con la superficie del sensor[3]

Fig. 5. Minucias de la huella dactilar.

4.2 Reconocimiento de voz

La voz es una característica única de un individuo. Sin embargo, existen personas con voces relativamente parecidas, por eso no se espera que sea suficientemente única para permitir la identificación de un usuario en una gran base de datos [2].

Para las aplicaciones de reconocimiento de voz se utilizan redes neuronales, estos algoritmos deben saber medir y estimar la similitud entre las muestras para poder devolver un resultado o lista de posibles candidatos. Muchas veces se complica debido a el ruido de fondo, calidad del microfono, canal de comunicación, etc.por lo que se debe tener en cuenta un margen de error [1].

Antes de extraer las características, la amplitud de la señal entrante puede ser normalizada y descomponerse en varias frecuencias de paso de banda canales. Las características extraídas de cada banda pueden estar en el dominio del tiempo o de frecuencia. Una de las características más utilizadas es la característica cepstral, que es un logaritmo de la transformación de Fourier de la señal de voz en cada banda [2].
Un sistema de reconocimiento de voz tiene dos etapas muy importantes una vez la señal es ingresada:

Detección de la duración de la muestra: Cuando la señal ingresa debe verificarse el tamaño de la misma. Se identificará el comienzo y final para eliminar espacios de silencios. Para así, enfocarse solamente en la voz.

Extracción de la información: La señal resultante se descompone en cuadros más pequeños, así se puede eliminar el ruido que tenga la señal, luego se le aplica un transformación de Fourier y la cual es pasada por un proceso de clasificación de patrones [4].
Finalmente esta información es comparada en una base de datos para comprobar la similitud entre las dos muestras. Luego se le entregará un resultado de la comprobación.

4.3 Reconocimiento facial

El reconocimiento facial es una técnica biométrica la cual se reconoce a un individuo mediante una imagen o fotografía. Este campo

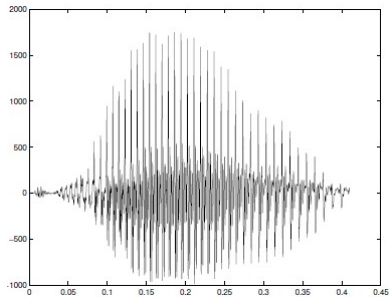


Fig. 6. Representación de una señal de voz en X e Y representando el tiempo y la amplitud de la señal.

en los últimos años se ha convertido en una de las áreas con más desarrollo, debido a la gran evolución en las tecnologías de video y captura de fotografía digital (el gran aumento de calidad de imagen en megapíxeles). Además, es de las más cómodas de utilizar debido a que solo se necesita una cámara que capture la imagen y es fácil de usar.
Dentro de los aspectos claves para poder comparar las muestras se encuentran mediciones como la distancia entre los ojos, longitud de la nariz o el ángulo de la mandíbula [1].
La detección consiste que dentro de la imagen, poder localizar la zona donde se encuentra el rostro y así poder realizar una extracción únicamente del rostro procesando solamente este [5].

5 Planteamiento del problema

Como objetivo del desarrollo propio en el que me voy a adentrar, será hacer un software de reconocimiento facial o de voz el cúal permita tener una fácil adaptación hacia los usuarios.
Como primera instancia el programa será solamente por computadora, accediendo a sus funciones mediante webcam y microfono integrado.
Luego como prototipo operativo de hardware se hará mediante la implementación de este software en modulos con camaras integradas, tomese como ejemplo “Esp32 Cam Modulo Wifi Bt Arduino Ov2640 2mp”.
Finalmente se quiere lograr que esté sistema requiera de recursos de hardware accesibles para poder obtener un mayor rango de aplicaciones, sin precios descomunales para facilitar el desarrollo y acceso.

6 Desarrollo

El desarrollo se llevará a cabo en el lenguaje de progamación Python junto a la libreria OpenCV.

7 Conclusiones

Actualmente los dispositivos biometricos son los métodos más eficaces para el control de seguridad y es de las grandes potencias a ser de las más utilizadas a futuro, reemplazando las antiguas y que no presentan un nivel de fiablidad del todo. Por el momento muchas de ellas están en desarrollo, algunas tienen un gran desempeño hasta la fecha, pero aún no se llega al pináculo de su potencial.
Las tecnologías biometrias facilitan el reconocimiento e identificación de personas, por ende es necesario el uso de estas para así poder aumentar el avance de la industria biométrica.

Referencias

1. Instituto Nacional de Ciberseguridad: Tecnologías biométricas aplicadas a la ciberseguridad,(2016)
2. Anil Jain, Michigan State University:Introduction to biometrics.
3. Carmen Sánchez Ávila: Aplicaciones de la Biometría a la Seguridad.
4. Nicolas Lopez Perez: Tecnicas de biometria basadas en patrones faciales del ser humano, (2012)
5. Darío Eduardo Villalón De La Vega: Diseño e implementación de una plataforma de software parareconocimiento facial en video, (2012)
6. Luzmila Elisa Pró Concepción: Tecnologías Biométricas aplicadas a la seguridad en las organizaciones, (2014)

Análisis Comparativo de Herramientas de Data Warehouse orientadas a Negocios Inteligentes

Rocha Saturnino, Rosana Giménez, Pablo Sández y Méndez Garabetti Miguel

Universidad del Aconcagua, Catamarca 147, Mendoza, Argentina
saturnino.rocha@gmail.com

Abstract. Para hacer una administración de la información en las empresas el Data Warehouse es una herramienta importante para los mercados nuevos en el mundo, estas herramientas permiten hacer análisis de la información facilitando la comprensión de sus negocios, mercados y a optimizar tomar decisiones. Para llevar a cabo dicho análisis es necesario establecer procesos de inteligencia empresarial. Existen gran variedad de herramientas que permiten realizar dichos procesos. Pero tienen diferencias en sus soluciones ofrecidas, dado que no lo cubren en su totalidad. El objetivo de esta investigación es estudiar las distintas plataformas existentes de Data Warehouse orientadas a Negocios inteligentes Pymes y trataremos de realizar la implementación de una aplicación OpenSource o Privativa.

Keywords: Negocios Inteligentes · Plataformas BI· Data Warehouse

1 Introdución

En un ambiente con una economía globalizada y conectada, la información va aumentando ampliamente, dentro de la empresa y de las fuentes externas. Los Sistemas de Información (IS) juegan un rol muy importante en la economía, permitiendo a las organizaciones y empresas a realizar tareas desde simples hasta complejas, almacenar la información, tener una comunicación con todos los niveles de la organización, transformar la información en conocimiento, automatizar rutinas y la interpretación de los datos y situaciones, entre otras.

La Inteligencia de Negocios (BI) y Data Warehouse (DW), como Sistemas de Información, tiene una serie de ventajas y beneficios para una organización lo que más se destaca es el manejar grandes cantidades de información y obtener conocimiento de ellas lo que va permitir un mejor desempeño dentro de la empresa, mejorar el manejo operativo de la empresa, tomar mejores decisiones y la optimización de tareas.

La Inteligencia de Negocio se puede definir como la obtención, administración y creación de reportes de los datos orientados a la toma de decisiones, las técnicas analíticas y los procesos informáticos que se usan para el análisis de la misma.

Existen varios programas BI que permiten mejorar los procesos y tareas aplicando diferentes métodos de soluciones, es decir, que hay aplicaciones BI que no cubren en su totalidad las necesidades o al ambiente que están dirigidos

(empresarial o académico). Cada vez existen más fuentes de datos y la presencia de procesos de inteligencia empresarial orientado al Big Data, hace importante conocer el estado y la posición de las herramientas en el mercado para facilitar una elección que se aproxime a la correcta.

La investigación tiene como objetivo hacer un análisis comparativo de las funcionalidades de las plataformas de BI OpenSource y Privativas, donde se puede llegar a conocer y elegir el cual puede ser óptimo para el uso en las Pequeñas y medianas empresas (Pymes) ya que se observa un gran desconocimiento sobre estas herramientas y la utilidad que tiene en el desarrollo empresarial, es importante el conocimiento y los beneficios que se pueden obtener, con el conocimiento de las distintas herramientas de BI y la elección optima de aplicación de BI.

Muchas Pymes no lo implementan o no cuenta un sistema BI en sus empresas, no solo por no disponer recursos tecnológicos o respaldo económico, hay otras opciones que lo limitan: la falta de información, bajos recursos financieros o poca motivación de automatizar tareas o procesos. En un mercado actual la falta de interés de BI, lleva a las empresas la perdida de ventajas competitivas. En la siguiente sección 2 analizaremos los métodos usados para la evaluación, comparación y análisis de las funcionalidades de BI existentes. El resto del trabajo se presenta de la siguiente forma. La sección 3 se presenta las características de 3 plataformas de BI de código abierto y 3 privados. La sección 4 presenta un análisis comparativo de las utilidades y funcionalidades donde se seleccionan un Privativo y otro OpenSource de plataforma de BI. Finalmente, en la sección 5 presentamos nuestras conclusiones.

2 Metodos de Evaluación

En este estudio se tomaron en cuenta los métodos utilizados por Gartner para evaluar las plataformas de BI. Estos se dividen en categorías de Integración, Entrega y el análisis de información.

La categoría Integración incluye la infraestructura de BI, metadatos de gestión, herramientas de desarrollo y criterios de colaboración.

La categoría Entrega de información contiene informes, cuadros de mando, consultas ad hoc, integración de Microsoft Office, búsqueda BI basado y móvil. La categoría de análisis incluye procesamiento analítico en línea (OLAP), visualización interactiva, modelado predictivo y minería de datos, cuadros de mando (KPI - indicadores clave de rendimiento) y modelado prescriptivo, simulación y optimización.

La categoría de análisis incluye OLAP, visualización interactiva, modelado predictivo y minería de datos, cuadros de mando y modelado prescriptivo, simulación y optimización. Desde el punto de vista de Gartner, los criterios anteriores son funcionalidades que una vez conectados a las plataformas de BI, permiten a las organizaciones tener más claros la clasificación y medición con el fin de dar apoyo a la toma de decisiones. En consecuencia, las organizaciones pueden mejorar sus niveles de rendimiento, eficiencia y rentabilidad, desarrollando objetivos y punto de vista diferentes de clientes, empleados y productos.

Otros dos nuevos métodos de Gartner: Cloud y tecnologías colaborativas. El concepto de nube (Cloud) se refiere al uso de la potencia de grandes niveles de procesamiento y almacenamiento existente en Internet. Por tanto, aplicaciones, datos y servicios están disponibles en Internet y ya no se limitan a computadoras y servidores propios. A pesar de las diversas ventajas (los bajos costos de equipo y operaciones, etc.). Las tecnologías colaborativas están siendo una herramienta importante para tomar decisiones en las organizaciones, ya que permiten la integración de datos de redes sociales, estas representan una fuente de información importante sobre comportamiento, tendencias, expectativas y necesidades. Estos datos una vez procesados puede ser determinantes en la planificación de estrategias y su posterior administración.

Hoy en día las plataformas de BI están en constante renovación, integrando nuevos criterios y tendencias de mercado. La capacidad de acceder a BI en dispositivos móviles (Android, iPhone, Windows Mobile). Esta funcionalidad que en los años 2007 no formaba parte de los criterios de la mayoría de los sistemas BI, después de varios años, se vio disponible en todas las plataformas consultadas.

Volviendo a los métodos mencionados y expuestos por Gartner seleccionamos: informes, cuadros de mando, consultas ad-hoc, Integración de Microsoft Office, BI móvil, OLAP, visualización interactiva, modelado de datos y cuadros de mando, indicador clave o medidor de desempeño (KPI). Las selecciones de estos métodos se basaron en la posibilidad medir las funcionalidades. Los otros métodos indican algunas restricciones comerciales de acceso y verificación.

El método de informe se base en la capacidad de generar informes interactivos con o sin parámetros, con distribución escalable y posibilidad de programación. Los cuadros de mando consisten en la publicación de informes en Desktop, web o en dispositivos móviles con paneles de control didáctico e interactivo que permiten la evaluación y comparación de información para cumplir con uno o más objetivos. Permiten una descripción general y un seguimiento más fácil del negocio. El uso de estos paneles es común para crear estrategias y manejo de datos en tiempo real.

Las consultas Ad-hoc permiten a los usuarios hacer sus propias preguntas sobre los datos, sin depender de TI para crear un informe. En particular, las herramientas deben tener una capa clara y robusta para permitir a los usuarios navegar por las fuentes de datos disponibles.

El criterio de integración de Microsoft Office se basa en que Microsoft Office, en particular Excel que es usado mucho en la creación de informes y análisis. En este contexto es necesario garantizar que la herramienta permita la integración con Microsoft Office.

La BI móvil permite a las organizaciones desarrollar y mostrar contenido en dispositivos móviles, donde se van toman las ventajas que ofrecen los dispositivos móviles como pantalla táctil, cámara, ubicación y búsquedas. Independientemente de la ubicación de los distintos actores (clientes, empleados o proveedores), se va obteniendo una información actualizada. Desde el punto de vista comercial en las organizaciones el saber la ubicación, gustos o compras

que hacen los clientes a través de los dispositivos móviles, es una estrategia de mercado On-Line.

El procesamiento analítico (OLAP) permite a los usuarios analizar datos con rápidas respuestas de sus consultas y cálculos, lo que permite un estilo de análisis conocido como "separar y cortar". Esta capacidad podría abarcar una variedad de arquitecturas de datos y almacenamiento. Los diferentes valores de datos comerciales (cifras de ventas, presupuestos, cantidades vendidas, etc.) se pueden obtener con mayor o menor detalle dependiendo según las necesidades del usuario.

El método de visualización interactiva permite la exploración de datos mediante la manipulación de imágenes de gráficos, con el color, el brillo, tamaño, forma y movimiento de los objetos visuales que representan aspectos del conjunto de datos que se analiza. Esto incluye una variedad de visualización que van más allá de las de los gráficos circulares, de barras y de líneas, mapas geográficos, diagramas de dispersión y otros con fines visuales especiales. Estas herramientas permiten a los usuarios analizar datos interactuando directamente con una representación visual de los datos.

El modelado predictivo y la minería de datos permite a las organizaciones categorizar las variables, para estimar las distintas variables e identificar patrones fijando modelos predictivos de eventos y escenarios que utilizan algoritmos matemáticos.

Los KPI consisten en un elemento indicador de desempeño. Dependiendo de los objetivos estratégicos de la organización, se hace esencial para seleccionar y priorizar el KPI. Esto es como medición del progreso del negocio, medir niveles de eficiencia y satisfacción, con el fin de potenciar objetivos de negocios.

Estos métodos ayudan a las organizaciones en la toma de decisiones y mejoras en el desempeño. Con los métodos citados anteriormente se intentará hacer pruebas en una plataforma de BI de código abierto (Pentaho) y en una comercial (Microsoft Power BI) que se describirán en la siguiente sección.

3 Plataformas de BI de código abierto y comercial, presentación, descripción y analisis

Hay dos modelos de soporte para código abierto. Uno es de código abierto basado en la comunidad, llamado código abierto y gratuito o FOSS. El otro es software comercial de código abierto, generalmente abreviado como COSS.

Los códigos abiertos comerciales se derivan del origen en el shareware y Mundo FOSS. En los primeros días, el código abierto a menudo fue diseñado y construido para uso personal por desarrolladores individuales en lugar de como parte de la infraestructura de TI de una organización.

Un proveedor comercial de código abierto es similar a un proveedor de software tradicional, pero con la diferencia es que el código fuente no es secreto. Esto permite una interacción más profunda entre clientes y desarrolladores, haciendo que el modelo de código abierto esté más centrado en la comunidad OpenSource.

A diferencia de la mayoría de los proyectos de software libre. Los proveedores de COSS utilizan elementos del modelo propietario, como proporcionar contratos de soporte o vender componentes de código no abierto que se pueden comprar además de, o en lugar de, la versión gratuita del software. Las dos ediciones diferentes de software que se ofrecen (comunitario y comercial) pueden causar confusión.

Las ventajas de usar software de código abierto, con la caracteristica principal el ser gratuito y sin costo alguno o el bajo costo para algunos. sus caracteristicas que destacan al software de código abierto con respectos a los comerciales son:

1. La libertad de ejecutar el programa, con cualquier propósito.
2. La libertad de estudiar cómo funciona el programa y editarlo como uno lo desee.
3. La libertad de redistribuir copias.
4. La libertad de distribuir copias modificadas de nuestras versiones.

Las herramientas de código abierto se distribuyen en varias licencias. La licencia va más allá de lo gratis que es el software, define las reglas de modificación del código fuente y su distribución, entre otras normas. Las licencias permiten el acceso sin restricciones y sin costo al desarrolador del software. Estas herramientas de código abierto permiten en la comunidad adaptar, reconfigurar, mejorar y reutilización codigo sin restricciones el uso del software.

Por el lado de las plataformas comerciales requiere el pago de licencias costosas por usuario o servidor donde se instale el servicio que ocaciona una limitación y restricción de acceso a programas BI comerciales, asociado por sus altos costos de licencias. Solo organizaciones con grandes respaldos economicos puedrian adquirirlos. Por eso, en este estudio nuestro enfoque no está solo en Plataformas de BI de código abierto, sino también al privativo. Donde se estudiara los siguientes programas de BI de código abierto (Vanilla, Pentaho y SpagoBI) y como comerciales o privativas (Microsof Power BI, Tableau y Qlik).

3.1 Vanilla

La empresa francesa BPM-Conseil fundada en 2004, son los Líder francés de Open Source BI y responsable del desarrollo de plataforma de vainilla. Vanilla es 100 porciento de código abierto y tiene un Licencia MPL (Licencia pública de Mozilla). Está escrito en Java y el portal web en tecnología GWT (Google Web Toolkit).

La plataforma Vanilla viene con todas las herramientas necesarias para administrar los datos desde el software aplicativo hasta la visualización del usuario final. El último La versión es Vanilla 6.6 que se compone de varios componentes: Portal web, Web de análisis gratuito, FreeWebReport, FreeDashboard, Arquitecto, BiGateway, BiWorkflow, BiProcessManager, FreeMetadata, complementos de Vanilla para Birt, FreeDashboard, FreeAnalysis Schema Designer, servicios empresariales y Norparena.

Vanilla ahora ofrece soporte para la última versión de la base de datos como MySql, MariaDb, Oracle, PostgreSql y Microsoft SqlServer2008 / SqlServer2012.

Las diferentes características de la plataforma, como integraciones con ETL, Key Performance Base de datos de indicadores, modo Saas, que permite alojar diferentes proyectos, entre otros.

3.2 SpagoBI

SpagoBI es una aplicación desarrollada por SpagoWorld, una organización fundada en 2006 con el apoyo de un código abierto comunidad.

La herramienta SpagoBI tiene una LGPL GNU (Lesser General Public Licencia), que permite a los usuarios utilizar todas las funcionalidades.

La suite SpagoBI es completamente de código abierto, se publica solo su versión estable. Es un alternativa al modelo de "licencia dual" o "codigo abierto", también incluye una versión paralela de software propietario como versión comercial. Esta desarrollado en Java y siempre esta disponible su versión gratuita.

La herramienta SpagoBI contiene una variedad de funcionalidades como la creación y exportación de informes, creación de gráficos, ETL, OLAP, Data Minería, KPI, servidor de aplicaciones, servicio de inicio de sesión, tablero, tarea,horario y servidor web.

Todas sus módulos están conectados con el núcleo de la plataforma, así como su capacidad de desarrollo. Es una plataforma de BI completa y robusta, segura y escalable, tiene soluciones únicas, no solo de informes y herramientas de gráficos, también cuenta con funcionalidades: Inteligencia de localizacion [Geo], construir libremente su propia consulta y generar la plantilla [QbE], KPI, paneles interactivos y en tiempo real.

3.3 Pentaho

La plataforma Pentaho sigue el modelo de desarrollo código abierto comercial. Su desarrollo está controlada por Pentaho Corporation y se distribuye en dos ediciones: Community Edition y Enterprise Edition.

La Community Edition es la versión de código abierto, mientras que la Enterprise Edition es un producto comercial que incluye no solo todas las funciones de Community Edition, pero también algunas más herramientas avanzadas para ayudar al desarrollo de Modelos BI e informes y soporte por contrato.

La plataforma Pentaho BI está desarrollada en lenguaje Java, que incluye las siguientes funcionalidades: ETL, OLAP, Data Mining y modelado predictivo, informes y cuadros de mando. Como se puede ver en la Figura 2, la plataforma consta de lo siguiente componentes: Pentaho Data Integration (a través del código abierto para la herramienta ETL Kettle), Pentaho Analysis Services (también a través de un herramienta de código abierto - Mondrian OLAP server), Pentaho Reporting, Minería de datos de Pentaho (utilizando el software Weka), paneles y Pentaho para Apache Hadoop.

La plataforma de BI de Pentaho proporciona la arquitectura y infraestructura necesaria para crear soluciones a problemas de inteligencia empresarial. Esta herramienta proporciona servicios que incluyen motores de autenticación, registro, auditoría, servicios web, entre otros servicios.

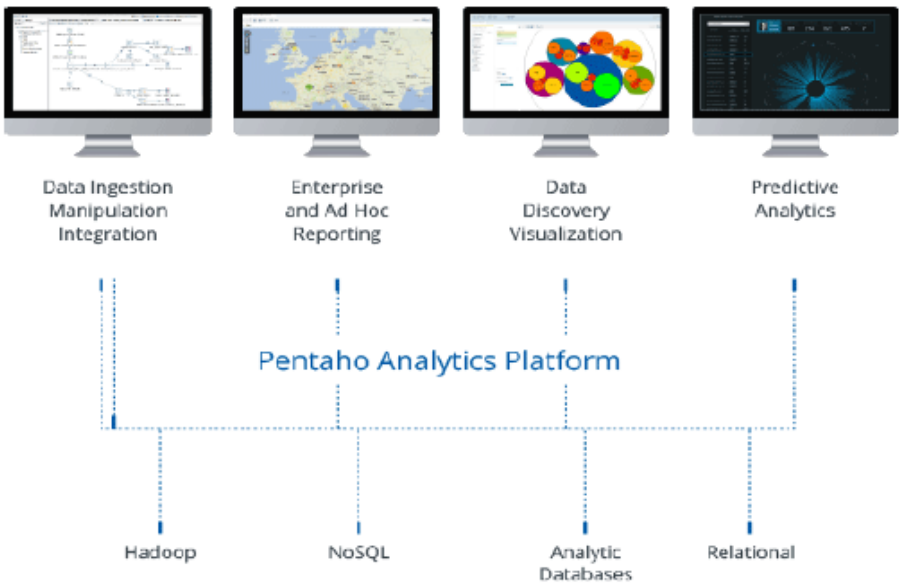


Figura 1 - Arquitectura de Pentaho Open BI Suite [Fuente: http://www.bizcubed.com.au/pentaho-pillars].

3.4 Microsoft Power BI

Es un software gratuita que es proporcionado por microsoft, power BI realiza la transformacion de datos en objetos visuales.

Power BI es un conjunto de aplicaciones de analisis de negocios que permite analizar datos y compartir informacion. Controla el estado de la organización creando paneles e informes interactivos.

1) PowerBI Desktop: el análisis visual con una intuitiva creación de informes. Arrastra y coloca el contenido exactamente donde deseas en un lienzo flexible y fluido.

2) PowerBI Mobile te permite conectar a los datos desde cualquier lugar y en cualquier momento con la aplicación para Windows, iOS y Android.

Excel PowerPivot.

PowerPivot es un complemento de Excel que puedes usar para realizar un análisis de datos eficaz y crear modelos de datos sofisticados.

3) Combina grandes volúmenes de datos de orígenes diferentes.

4) Realiza análisis de la información rápidamente.

5) Comparte puntos de vista con facilidad.

Excel PowerView.

Explorar rápidamente los datos de forma visual e interactiva. PowerView es una herramienta que permite crear nuevos tipos de informes y visualizaciones.

1) Orientada a la explotación de datos de forma interactiva.

2) Permite mostrar gráficos, tablas y mapas en una ventana. 3) Vincula todos los componentes de los tableros.

4) Se alimenta con los datos de PowerPivot o modelos tabulares.

5) Ideal para usuarios de negocios como analistas, tomadores de decisiones y personas que trabajan con información y datos.

Excel PowerQuery.

PowerQuery te permite conectar con cualquier origen de datos.

1) Transforma, combina y acomoda los datos según tu necesidad.

2) Expórtalos a una hoja Excel o a un modelo de datos de PowerPivot.

3) El proceso de conexión y transformación de los datos queda grabado para poder ser reutilizado.

4) Te permite buscar y conectar con fuentes públicas de datos.

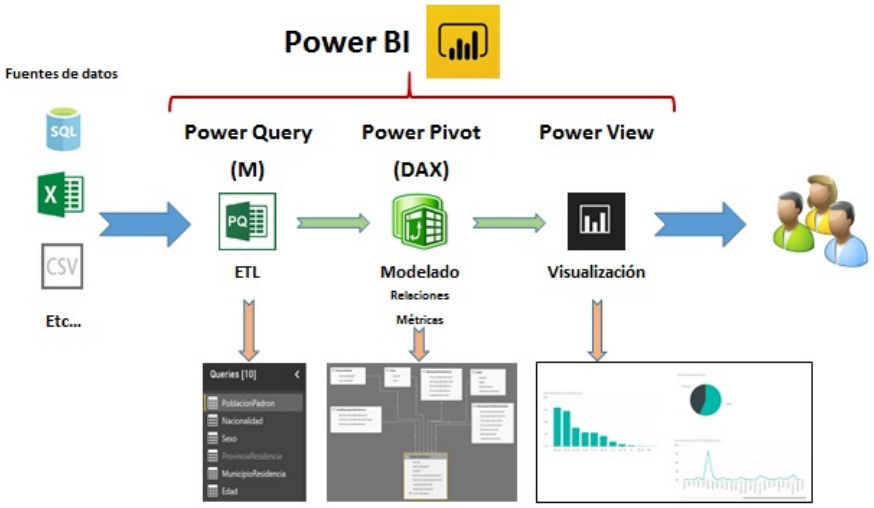


Fig. 2. Modelo conceptual power BI.

3.5 Qlik

La compañia Qlik tiene como herramienta para Inteligencia de Negocio a QlikView que simplifica al usuario en la toma de decisiones. Esta herramienta permite crear e implementar aplicaciones de analisis guiadas e interactivas y con un panel de control flexible. Los usuarios finales obtienen una mejor busqueda interactiva de datos que diferencia a Qlik algunos beneficios se presentan a continuacion:

1) Controlar: Cree descubrimientos de datos impulsados por la empresa con rutas de analisis guiadas que son altamente personalizables y rigurosamente gobernadas.

2) Seguro: Control de arriba hacia abajo de las aplicaciones de analisis, permisos y administracion de datos.

3) Flexibilidad: Cree las herramientas exactas que necesita para su organizacion y equipos.

4) Personalizable: Desarrolle aplicaciones personalizadas con scripts QlikView y amplıe el desarrollo con QlikView Workbench.

5) Combinable: Integre QlikView en sus aplicaciones comerciales y software de administracion de sistemas con API.

6) Integracion de datos: Unifique las fuentes de datos para una vision completa de la informacion, los datos y las aplicaciones administrados centralmente hacen que sea mas facil descubrir las estadısticas.

7) Informes avanzados: Crear y distribuir informes, integración de documentos Microsoft Office con Qlik NPrinting, QlikView y Qlik Sense, QlikView permite el almaceniemto de los documentos localmente en la computadora escritorio en la en la red, en QlikView Server permite que los informes puedan ser visto en diferentes formatos como PDF, Excel, Word, PowerPoint y HTML.

3.6 Tableau

Tableau ofrece una experiencia de exploración intuitiva, interactiva y basada en lo visual que permite a los usuarios comerciales acceder, preparar, analizar y presentar hallazgos en sus datos sin habilidades técnicas o codificación. La oferta de Tableau se implementa principalmente en las instalaciones, ya sea como una aplicación de escritorio independiente o integrada con un servidor para compartir contenido; Tableau Online es la oferta SaaS basada en la nube. En 2018, Tableau introdujo una nueva función de Visor de menor precio y ahora lidera con precios de usuario basados en suscripciones. Tableau Prep se lanzó para mejorar la preparación de datos y la creación de perfiles en Tableau Desktop, y las capacidades de programación basadas en servidor más sólidas se encuentran en las pruebas beta. Tableau también adquirió Empirical Systems en 2018 para ampliar sus capacidades analíticas aumentadas.

Exploración visual fácil y manipulación de datos: Tableau permite a los usuarios ingerir rápidamente datos de una amplia gama de fuentes de datos, combinarlos y visualizar los resultados utilizando las prácticas en la percepción visual. Los datos se pueden manipular mientras se visualizan, como cuando se crean grupos, ubicaciones y nuevas jerarquías, todo con un alto grado de facilidad de uso.

4 Análisis comparativo de las funcionalidades de cada plataforma de BI

Basado en las especificaciones técnicas se eligieron dos plataformas una de Open-Source(Pentaho) y otra privativa comercial (Microsoft Power BI) de las plataformas de BI estudiadas y se analizó según los métodos definidos en sección 2, crear un cuadro comparativo sobre sus caractiristicas.

La tabla está creada como una ayuda para la incursión y selección adecuada de la plataforma de BI para una organización, según las necesidades de negocios de la organización. También se intenta evaluar la usabilidad de las herramientas

y funcionalidades de la herramienta de BI seleccionada, según el contexto organizacional, el rendimiento de las herramientas indica lo que es capaz de hacer la aplicación. Por lo tanto, si es viable implementar tomando en cuenta los costos y la plataforma correcta según el tipo de negocio de la empresa.

Algunas plataformas de código abierto tienen versiones comerciales con un gran número de funcionalidades y soporte, lo que informa las versiones de comunidad que tienen una cantidad reducida de funcionalidades con respecto a las versiones puramente comerciales.

FUNCIONALIDADES	PLATAFORMAS BI	
	PENTAHO	MS POWER BI
USABILIDAD	*	*
REPORTES	*	*
PANEL DE CONTROL	*	*
CONSULTAS AD-HOC	*	*
INTEGRACION MS OFFICE	*	*
BI MOBIL	*	*
OLAP	*	*
VISUALIZACION INTERACTIVA	*	*
PREDICCION MODELADO/MINERIA DE DATOS	*	*
KPI	*	*
TECONOLOGIAS COLABORATIVA	*	*
INTERFACES AMIGABLES	*	*
ASISTENTES DE CONTENIDOS		*
BIBLIOGRAFIA Y TUTORIALES	*	*
SOPORTE PRIVADO Y COMUNIDAD GNU	*	*
CLOUD	*	*

Fig. 3. Tabla comparativa BI

5 Conclusiones

La inteligencia de negocios se ha convertido en parte importante dentro de la administración y marketing de las empresas, una herramienta de este tipo puede llevar a las empresas al éxito o no dependiendo del grado de importacia que se le asigne al BI, puede llevarlo al fracaso. El análisis de las herramientas que se ha hecho, permite observar que en general todas las soluciones tienen unas funciones básicas que todo negocio necesita, sin embargo, cada empresa busca diferenciar su producto por medio del fortalecimiento de alguna de estas funciones, con el fin de posicionar su producto en lo mas alto de su mercado. No se tiene información

sobre los costos asociados a todas las herramientas, dado que no todas las empresas están dispuestas a proporcionar esta información en su sitio de internet, por lo que es necesario contactar directamente a las empresas proveedoras y esta información tiene un tiempo de espera que no permitió tener esta información a la fecha. La comparación realizada, da como resultado a Microsoft Power BI como objeto de estudio con el fin de evaluar los elementos más importantes que han de tenerse en cuenta para la implementación de las herramientas en las Pymes. Por medio del estudio hecho a la herramienta Pentaho y Microsoft Power BI, se logró obtener que los costos de aprendizaje son un elemento importante en la adquisición e implementación de herramientas de inteligencia de negocios en las empresas, en especial las pequeñas y medianas. Adicionalmente, podemos decir que la herramienta Microsoft Power BI es una herramienta que les permitira un análisis de la información sin mayor problemas de llevar acabo ya que uso es muy asociado a la herramienta Excel de MS office, sin la necesidad de ser un experto del tema, su fácil uso, operatividad y adaptación a usuarios finales hace que esta herramienta sea para las Pymes una opción importanre que los lleve a incursionar en el mundo competitivo del internet de las cosas y automatizar sus procesos y sin grandes costos de implementación y compras de licencias.

References

1. Angie Alexandra Gutierrez Camelo, Miguel Angel Devia Llanos, Giovanny Mauricio Tarazona Bermúdez F. - Research inteligencia de negocios: estudio de caso sector tecnológico colombiano (2016) https://revistas.udistrital.edu.co/index.php/REDES/article/view/10728/12133
2. ÖyküIşıkaMary C.Jonesb, Anna Sidorovab - Articulo Business intelligence success: The roles of BI capabilities and decision environments (2013) https://www.sciencedirect.com/science/article/abs/pii/S0378720612000833
3. Temitope Adeoye Oketunji, Raufu Olalekan Omodara: Tesis - Estudio Design of Data Warehouse and Business Intelligence System (2011) https://www.diva-portal.org/smash/get/diva2:831050/FULLTEXT01.pdf
4. Sinnexus - Articulo Qué es Business Intelligence? (2011) https://www.diva-portal.org/smash/get/diva2:831050/FULLTEXT01.pdf
5. Carlos Fernandez - Blogs Comparativa de costes de adquisición y mantenimiento de plataformas de BI (2012) https://www.dataprix.com/files/bi-comparison-of-open-source-and-traditional-vendor.pdf
6. Castánder Martín Iñigo, Gerrikagoitia Arren, Jon Kepa, Alzua-Sorzabal, Aurkenez - Articulo Estudio comparativo de las herramientas de Business Intelligence: Empoderando el criterio de selección a las Pymes (2013) https://www.researchgate.net/publication/259442931
7. Hugh J. Watson, Barbara H. Wixoms - Revista The Current State of Business Intelligence (2007) https://www.researchgate.net/publication/2961945
8. Itrends - ArticuloEvolución de la analítica y la inteligencia de negocio (2020) https://www.ittrends.es/actualidad/2020/02/evolucion-de-la-analitica-y-la-inteligencia-de-negocio

9. José Antonio, García-Calvo Fernández - Articulo Business Intelligence: Cuadrante mágico de Gartner 2020 (2020) https://www.inforges.es/post/cuadrante-magico-de-gartner-2020-business-intelligence
10. Kruzylko Ostojic, Claudia Yanina - Tesis Aplicación de técnicas y estrategias de Inteligencia de Negocio (BI) para analizar/integrar información de los alumnos de la Facultad de Informática de la UNLP (2012) http://sedici.unlp.edu.ar/handle/10915/63168
11. Juan Manuel Alberto Perusquia, Velasco EduardoAhumada Tello -Articulo Inteligencia de negocios: estrategia para el desarrollo de competitividad en empresas de base tecnológica
(2016) https://www.sciencedirect.com/science/article/pii/S0186104215000807
12. Clarysabel Tovar-Tesis Investigación sobre la aplicación de BusinessIntelligence en la gestión de lasPymes de Argentina
(2017) https://www.palermo.edu/economicas/cbrs/pdf/pbr15/PBR-15-05-Tovar.pdf
13. Nuno Leite Isabel Pedrosa and Jorge Bernardinos - Articulo Open Source Business Intelligence Platforms' Assessment using OSSpal Methodology
(2018) https://www.scitepress.org/Link.aspx?doi=10.5220

POSTERS
RESÚMENES

Implementación de Business Intelligence para la Gestión de Información en Organismos Públicos

Ezequiel Moyano[1], Martín Villarreal[1], Abraham Jurado[1]

[1] Instituto de Desarrollo Económico e Innovación, UNTDF, Yrigoyen 879, (9410) Ushuaia, Argentina. www.untdf.edu.ar
{emoyano, mvillarreal}@untdf.edu.ar, vakito85@gmail.com

Resumen: La toman decisiones es cada ves más importante, contar con información precisa y concreta es esencial.
Los organismos públicos ofrecen diversos servicios a su comunidad para satisfacer sus demandas y necesidades. Los grandes volúmenes de datos no alcanzan para generar información, y menos conocimiento; contar con información correcta, consistente y clara es vital para la gestión.
El IPV es el organismo encargado de definir y ejecutar la política habitacional de la provincia de TDF, sus sistemas no permiten obtener un conocimiento de la realidad, lo que afecta la toma de decisiones para satisfacer las demandas planteadas. La implementación de un sistema Business Intelligence permitirá generar información precisa y confiable a sus autoridades.
El presente trabajo centra su estudio en la implementación de la técnica mas adecuada de Business Intelligence en el ámbito del IPV, para mejorar la calidad de servicios en la toma de decisiones.

Palabras clave: Business Intelligence, Data Werehouse, Organismo Públicos, IPV.

Contexto

La línea de investigación se desarrolla en el Instituto de Desarrollo Económico e Innovación (IDEI) de la UNTDF, por docentes-investigadores que llevan adelante proyectos en la temática. Actualmente ésta línea se desarrolló en el marco de un proyecto de investigación, y derivó en la tesis de grado de un Alumno de la UNTDF.

1 Introducción

1.1 Business Intelligence

El objetivo de Business Intelligence (BI) es apoyar de forma sostenible y continuada a las organizaciones para mejorar su competitividad, facilitando la información necesaria para la toma de decisiones[1], Conesa define "*conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información en una organización"* [2].

Se puede definir BI como un conjunto de herramientas integradas para recopilar, almacenar, analizar y administrar datos para respaldar la toma de decisiones en todos los niveles de una organización.

La BI está compuesta de factores claves que permiten determinar la relación de los procesos de negocio[3]: los datos, la información y el conocimiento.

Los beneficios más importantes y claves en el uso de BI [4] son:

- Integrar datos de diferentes fuentes y acceder a la información en un formato único.
- Una visión única, histórica, persistente y de calidad de toda la información.
- Aporta la información basada en tiempo y hechos reales.
- Las herramientas de BI permiten una fácil y rápida interacción con los usuarios.
- Permite que la empresa tenga un continuo seguimiento de sus procesos.

Las funciones clave en un sistema de BI se centran en un data warehouse: recopilación y adquisición de datos, limpieza e integración, almacenamiento, análisis y entrega de conocimientos [5]. Para lo cual las soluciones de BI integran cinco capas, como se expone en la siguiente imagen figura1, cada capa es vital para garantizar la calidad de datos y el flujo de información en un sistema de BI[6]. 1-Capa de orígenes de datos, 2-Capa de extracción, transformación y carga (ETL), 3-Capa de Data Warehouse o Data Marts, 4-Capa de Análisis y 5-Capa de Presentación[6].

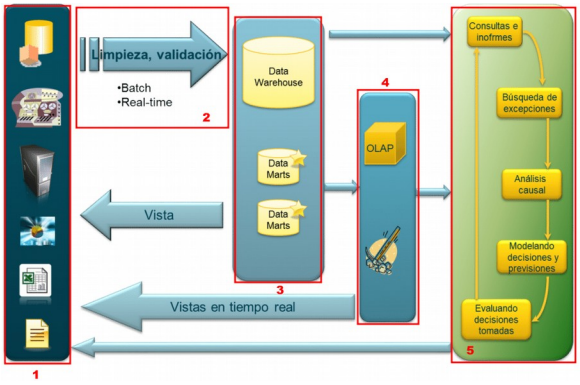


Figura 1: Arquitectura de un sistema de Business Intelligence

Un Data Warehousing es el proceso de extraer y filtrar datos de las operaciones comunes de la organización, procedentes de los distintos sistemas de información, para transformarlos, integrarlos y almacenarlos en un almacén de datos con el fin de acceder a ellos para dar soporte en el proceso de toma de decisiones[4].

En su definición más simple un Data Warehouse (DW) es una copia de los datos transaccionales crudos, en datos estructurados específicamente para la consulta y análisis[2]. El esquema de solución de BI, un Data Warehouse es la pieza principal como fuente de información para las herramientas de análisis[4].

Un Data Mart (DM) es un subconjunto de un DW que suele estar orientado a un área específica de la organización. La diferencia entre un DM y un DW es su alcance. El Data Mart cubre las necesidades de un determinado departamento dentro de la organización, en cambio el ámbito del Data Warehouse es toda la organización[7].

1.2 Organismo Público: el Instituto Provincial de Vivienda IPV.

El Instituto Provincial de Vivienda y Habitat, es un organismo público con 42 años de servicio en Tierra del Fuego. Su misión principal es promover el acceso a una solución habitacional a los sectores de la población que no tienen capacidad de ingresar al mercado inmobiliario, mediante la definición de políticas habitacionales y la ejecución de programas que tiendan a satisfacer la demanda habitacional de TDF.

Disponer de información clara y precisa es importante para quienes toman decisiones a nivel estratégico que permita obtener conocimiento de las realidades sociales acercando a los directores a decisiones adecuadas. El Área Informática del IPV ha desarrollado sistemas administrativos que no cubren las necesidades de análisis de información específica en tiempo real que los directores necesitan.

Los análisis implican un tipo de exploración profunda no se resuelve con consultas SQL, sino del análisis de grandes cantidades de datos para extraer información útil como por ejemplo: tendencias, patrones de comportamiento, elaboración de informes complejos, etc., que el IPV no posee y de ahí la necesidad de implementar BI.

El presente trabajo pretende implementar una solución de BI mediante la construcción de un Data Warehouse en donde se almacenen datos recopilados para su posterior análisis y presentación como información útil y valiosa. Este modelo de solución estará basado en herramientas de Software Libre y utilizando la propia infraestructura tecnológica del IPV.

2 Líneas de Investigación y Desarrollo

La línea de investigación general se enfoca en el estudio y desarrollo de las metodologías más usadas de diseño y construcción de un DW. Posteriormente se comentan y comparan las características de las herramientas Open Source orientadas a BI, y se determina la más adecuada para la implementación de la solución BI.

Las metodologías más comunes de DW son Bill Inmon[8], Ralph Kimball[9] y la metodología Hefesto[10] (del argentino Dario Bernabeu), utilizada en muchos proyectos de BI en Latinoamérica, con grandes resultados.

La metodología Inmon toma los datos como punto de partida y los requerimientos al final del ciclo de desarrollo utilizando un bucle de retroalimentación. Se vale de un enfoque top-down en el que el modelo de datos se desarrolla primero y se utiliza para la construcción del DW. Esta metodología es generalmente recomendado para proyectos grandes donde no se tiene una visión completa de la organización.

La metodología Kimball se basa en los requerimientos, el modelado de datos usa un enfoque buttom-up, comienza con un único modelo dimensional de negocios que evoluciona hacia el modelo de datos corporativo más grande. Esta metodología es recomendada en proyectos chicos o medianos para un desarrollo rápido e incremental.

La metodología Hefesto puede adaptarse tanto a proyectos grandes como medianos, se construye un DW a partir de los requerimientos. Sigue un proceso simple e intuitivo de fácil comprensión a medida que se avanza en el desarrollo.

Existen diversas herramientas Open Source orientadas a BI, entre las más recomendadas se estudiaron en profundidad: Jaspersoft Community Edition, Microsoft Power BI Desktop, Tableau Desktop y Pentaho Community Edition[11].

3 Resultados

Se identificaron las necesidades de información de las áreas del IPV y se implementó una solución BI que posibilitó disponer de información histórica para su análisis y procesamiento como base para la toma de decisiones. La solución BI permitió comprender las necesidades del organismo como mejorar las prácticas y procesos organizacionales.

Para la construcción de los DW, la metodología Hefesto es la que mejor se adaptó a las necesidades del IPV, ya que especifica puntualmente los pasos a seguir en cada fase, se adapta fácilmente a cualquier etapa de desarrollo aplicando un método híbrido, la documentación es clara y permite identificar los requerimientos rápidamente, su flexibilidad lo hace independiente de la plataforma de desarrollo.

Se determinó que la Suite Pentaho era la mejor opción disponible para llevar adelante la implementación de una solución BI ya que incluye los componentes esenciales para el caso de estudio propuesto: integración, gestión y análisis de la información, análisis multidimensional OLAP, presentación de informes, capacidades ETL y creación de cuadros de mando para el usuario. La ventaja de no tener costo de licencia lo hace ideal para el desarrollo en instituciones públicas y el multiplataforma.

4 Formación de Recursos Humanos

El equipo de trabajo está conformado por docentes investigadores de la UNTDF. Se encuentran en desarrollo dos tesis de Magister en Ing. de Software, UNLP, Ezequiel Moyano y Martín Villarreal; y una Tesis de Grado (entregada) de Abraham Jurado.

5 Bibliografía

1. Schiff M. Data Warehousing: Remembering the basics transforming data intelligence, 2010.
2. Marqués M.P., Business Intelligence Técnicas, herramientas y aplicaciones, 2015.
3. Muñoz-Hernández H., Osorio-Mass R., Zúñiga-Pérez,Clío L., Inteligencia de los negocios. Clave del Éxito en la era de la información, 2016.
4. Conesa Caralt J., Curto Díaz J., Introducción al Business Intelligence, 2010.
5. Cano J. L., Business Intelligence: Competir con Información, 2007.
6. Ramos S., Arquitecturas Business Intelligence: símil con un restaurante. SolidQ Blogs, 2017,
7. Curto J., Data Warehousing, Data Warehouse y Datamart, 2007. Disponible en
8. Inmon, W. H., Building the data warehouse. John wiley & sons, 2005
9. Dertiano, V., El enfoque de Ralph Kimball. BI Geek Blog, October 2019.
10. Bernabeu R. D., García Mattío M., Hefesto: Metodología para la construcción de un data warehouse,Versión 3, 2018, https://sourceforge.net/projects/bihefesto/
11. Pentaho Mondrian Documentation, 2019, Mondrian.pentaho.com.

Una Aproximación al Uso de Rúbricas en la Evaluación de la Ingeniería de Software

Ezequiel Moyano[1], Matías Moncho[1]

[1] Instituto de Desarrollo Económico e Innovación, UNTDF, Yrigoyen 879, (9410) Ushuaia, Argentina. www.untdf.edu.ar
{emoyano, mmoncho}@untdf.edu.ar

Resumen: El presente trabajo tiene como objetivo sobre los dispositivos evaluación en la asignatura de Ingeniería de Software I de la UNDTF, incorporando una nueva forma de evaluar los contenidos y, el impacto del uso de rúbricas a los efectos de determinar el conocimiento y avance de los estudiantes. La evaluación formativa tiene como fundamento considerar la evaluación como un trabajo cotidiano en el aula, para orientar el proceso de enseñanza - aprendizaje y la toma de decisiones en beneficio a los estudiantes; que debe contemplarse como un instrumento de mejora de la enseñanza.
Las rúbricas se presentan ante los alumnos y docentes, como un documento o guía en los cuales puedan conocer los objetivos, expectativas y logros que deben alcanzar dentro de las competencias establecidas en la currícula.

Palabras clave: Enseñanza, Evaluación Formativa, Rúbrica, Ingeniería de Software.

Contexto

El trabajo surge de una experiencia en la cátedra de Ingeniería de Software I como una preocupación de sus docentes en el marco de los dispositivos de evaluación utilizados y como mejorar los procesos de enseñanza-aprendizaje.

1 Introducción

La evaluación se constituye en uno de los elementos más importantes de todo el proceso de enseñanza-aprendizaje, proceso sistemático y continuo mediante el cual se determina el grado en que se están logrando los objetivos de aprendizaje. La misma no debe pensarse solo como verificación de los aprendizajes sino como un proceso dinámico, continuo, presente durante todo el proceso educativo.

Evaluar es visto habitualmente como sinónimo de calificar, de enjuiciamiento objetivo y preciso[1]; donde su la función esencial es medir la capacidad y el desempeño de los estudiantes. Esta apreciación hace ver la evaluación como un instrumento que afecta decisivamente lo que se pretende medir, abriendo el camino a un replanteamiento para ser considerada como un instrumento de intervención y no de simple constatación[1].

La evaluación debe convertirse en un instrumento de aprendizaje[2], netamente formativa. Sustituir los juicios terminales sobre los logros y capacidades de los estudiantes sino un instrumento de mejora de la enseñanza[3].

El espacio curricular sobre el que se va a desarrollar el presente trabajo es Ingeniería de Software I constituye en el primer contacto del alumno con los conceptos vinculados al proceso de desarrollo de software y a los diversos métodos y herramientas de la ingeniería de software.

El proceso de desarrollo de software se compone de etapas que no se pueden entender, analizar ni desarrollar por separado, por otro lado no se puede concebir que el proceso de desarrollo esté a cargo de una sola persona, hoy la construcción de sistemas de software involucra una gran cantidad de personas que trabajan cooperativa y colaborativamente y con estas premisas se trabaja con los estudiantes.

Los estudiantes eran evaluados a través de dos parciales prácticos, que presentaba algunos inconvenientes: los parciales eran individuales y los ejercicios presentaban situaciones disimiles y una etapa diferente para cada una.

Estas formas de evaluar se contraponen con la dinámica de la cursada y generaba preocupaciones ya que en esta forma de evaluar es fuente de angustia y de estrés[4]. Por otro lado conllevan varios factores subjetivos a la hora de examinar y calificar a los alumnos, el *efecto de halo*, *hipótesis de la personalidad implícita*, *tendencia a la categorización*, *primacía de la primera impresión, influencia del aspecto físico, etc.*[5].

Razón por la cual el cuerpo docente decidió reflexionar sobre la forma de evaluar a los estudiantes, sobre todo por el concepto general que se le intenta dar al espacio curricular. La evaluación no se puede ver como algo solitario y por fuera del proceso de enseñanza-aprendizaje, por cual se propone trabajar sobre el modelo constructivista que propone un proceso de enseñanza dinámico, participativo e interactivo orientada a evaluar los procesos de construcción personal del conocimiento[6].

Apuntar a una enseñanza formativa y en dispositivos que permitan la práctica de evaluaciones formativas es en camino a recorrer, y donde el uso de la rúbricas tiene un impacto positivo como guías de estos dispositivos.

2 Desarrollo e Implementación

Se trabajó desde un modelo constructivista considerando los contenidos de la asignatura y de acuerdo a las competencias establecidas; se elaboró un dispositivo evaluativo desde la perspectiva de la evaluación formativa, y la rúbrica como herramienta principal para la evaluación continua de los aprendizajes.

Se eliminaron las instancias de parciales tradicionales. Se establecieron a través de la plataforma virtual auto-evaluaciones a efectos que el estudiante conozca el grado de conocimiento alcanzado.

Se diseñó como dispositivo evaluación un único Trabajo Practico Integrador (TPI) que cumpla y aborde los contenido que se desarrollan en el trascurso de la cursada. El TPI representa una problemática concreta y abarcativa que se realiza durante todo el cuatrimestre sobre la cual se desarrollará el proceso de desarrollo de software; la cátedra utiliza metodología de desarrollo ágil, incremental e iterativo, junto con el uso de herramientas digitales provee.

Básicamente la nueva propuesta consiste en definir 7 (siete) etapas bien determinadas, para lo cual: Cada etapa representa una etapa del proceso, los alumnos trabajan en grupos de 3 estudiantes, los docentes exponen y entregan la rúbrica con los objetivos a alcanzar, las fechas de entrega de cada etapa para su presentación, los grupos deben realizar una defensa oral para explicar los criterios, estrategias, complicaciones, etc. de su desarrollo. Para Regularizar la materia de exige haber aprobado al menos 5 etapas del TPI.

En el diseño de la rúbrica se consideraron algunos aspectos, para la primera experiencia (2018) se definió: utilizar una rúbrica genérica para todas las etapas del proceso de desarrollo, elaborada por el cuerpo docente. Se determinaron las dimensiones y los criterios de los objetivos de aprendizaje y se establecieron la cantidad y niveles de calidad.

Después de la primera experiencia, surgieron algunos interrogantes en el cuerpo docentes y de consultas realizadas a los alumnos; para el año 2019 se definieron algunos cambios importantes respecto al año anterior; el dispositivo de evaluación y la metodología fue muy bien recibida por lo que se mantuvo, pero se modificó la estructura de la rúbrica que se puede observar en la figura 1. Entre los cambios más significativos se puede mencionar: Se dio participación a los alumnos en la conformación de la misma. se modificaron los nombres de los niveles de calidad, se redefinieron los criterios de forma más conceptual.

Criterios	**Ponderación**	**Niveles de Calidad**			
	100%	**Muy Bien**	**Bien**	**Regular**	**Insuficiente**
Compresión del problema de la etapa	15	Comprende la etapa del proceso desarrollo	Es capaz de comprender la etapa en el proceso desarrollo con alguna ayuda	La comprensión de la etapa es pobre, necesita contar con asistencia.	Necesita asistencia constante para comprender la etapa.
Planificación del trabajo	25	Cumple totalmente lo planificado.	Demoras pequeñas según la planificación.	Necesita ajustar fechas en la planificación.	La planificación no se cumple.
Uso de herramientas conceptuales	30	Selecciona y utiliza correctamente los diferentes diagramas.	Presenta algún problema para seleccionar y utilizar los diagramas	Presenta serios problemas para seleccionar y usar los diagramas.	No selecciona ni sabe utilizar los diagramas.
Utilización de TICs	12	Implementa todo el trabajo con TICs.	Uso de TICs en gran medida	Muy poco uso de herramientas TICs	Desarrollo exclusivo en papel
Resultados y entregables	18	Resultados satisfacen Completamente.	Los resultados son satisfactorios	Resultados son confusos.	Resultados no aceptables.

Figura 1: Rúbrica para la evaluación de cada etapa del proceso de desarrollo

3 Resultados

Del análisis comparativo respecto a los dispositivos de evaluación utilizados en los últimos años en la cátedra se pueden obtener varios resultados entre los parciales prácticos (2016/7) y con el TPI (2018/9) y el uso de rúbricas; Para el análisis se utilizaron encuestas anónimas que se realizan cada fin de cursada, con el objetivo de conocer la opinión general de los estudiantes y que sirva de partida de la reflexión para el cuerpo docente. El primer análisis que se realizó fue comparar el porcentaje de estudiantes que han regularizado la materia, en este caso el parámetro utilizado no refleja una gran variación con un promedio general de 70%.

Interesante es la relación es ante diferentes criterios consultados sobre aspectos de la evaluación, donde se refleja una mayor recepción respecto al dispositivo utilizado en los últimos años (Trabajo Practico Integrador + rúbricas) respecto a los parciales tradicionales, figura 2; y el relevamiento realizado sobre el uso de la rúbrica, para lo cual se consideró las respuestas de 25 alumnos en total, como se ve en la figura 3.

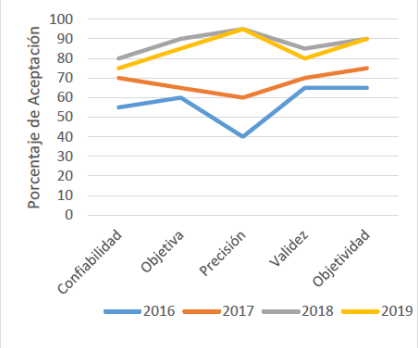


Figura 2: Porcentaje de aceptación de la evaluación.

Figura 3: Relevamiento del uso de la Rúbrica.

Se pueden obtener algunas conclusiones: Las producciones sobre el TPI resultaron más productivas y eficientes, los grupos obtuvieron resultados mas significativos, favoreció una mejor comprensión de avances de los alumnos en el proceso de enseñanza, feedback continúo con los alumnos, fuerte retroalimentación entre ellos beneficiando el trabajo cooperativo y colaborativo, los alumnos pudieron identificar claramente sus puntos fuertes y debilidades.

La experiencia ha resultado gratamente satisfactoria y se seguirá implementado en los próximos años con las adecuaciones necesarias producto de las experiencias.

Como consideraciones a futuro a tener en cuanta es que la elaboración de la rúbrica requiere tiempo y esfuerzo, se aconseja que el grupo docente trabaje con bastante anticipación, es fundamental presentar y explicitar la rúbrica la primera clase y propiciar la participación de los estudiantes en su confección final.

5 Bibliografía

1. Alonso Sánchez, M.; Gil Pérez, D., Evaluar No es Calificar la Evaluación y la Calificación en una Enseñanza Constructivista. Investigación en la escuela, ISSN 0213-7771. 1996
2. Bordas M. Estrategias de evaluación de los Aprendizajes centrados en el proceso, Revista Española de Pedagogía, 2001.
3. Romero M., Importancia de la evaluación y algunos instrumentos para evaluar, Bogotá, 2014.
4. Astolfi, J. P. El "error": Un Medio para Pensar, Madrid, 1999.
5. Camilloni A. Las Apreciaciones Personales del Profesor. Mimeo, 1987.
6. Saturnino de la Torre, Aprender De Los Errores El tratamiento didáctico de los errores como estrategias innovadoras. Ed. Magisterio Río de la Plata, 2004.
http://rubistar.4teachers.org/

Diseño de Estrategias Didácticas para la Enseñanza de Ingeniaría de Requerimientos

Ezequiel Moyano[1], Martín Villarreal[1], Matías Moncho[1], Marcela Jerez[1], Lucas Romano[1]

[1] Instituto de Desarrollo Económico e Innovación, UNTDF, Yrigoyen 879, (9410) Ushuaia, Argentina. www.untdf.edu.ar
{emoyano, mvillarreal, mmoncho,mjerez,lromano}@untdf.edu.ar

Resumen: Los procesos de enseñanza-aprendizaje se caracterizan por el papel determinante del docente en identificar, planificar e instrumentar estrategias en la formación de los estudiantes. Las estrategias de enseñanza y de aprendizaje se encuentran involucradas de forma que el aprendizaje no sea solo la adquisición de conocimientos. A su vez la didáctica exige la utilización de estrategias y métodos centrados en el sujeto que aprende.
Enseñar ingeniería de requerimientos es fundamental debido a que es la primera, y más importante, etapa de todo paradigma de ingeniería de software. La especificación de requerimientos representa la comprensión de las necesidades, deseos y expectativas de los stakeholders.
Las estrategias tradicionales (narrativas descriptivas de un contexto sin "interacción directa con un usuario") no son suficientes, resulta fundamental incorporar estrategias didácticas adecuadas a las particularidades y dificultades planteadas, que permitan mejorar las prácticas y los aprendizajes de los estudiantes en los procesos que intervienen en la ingeniería de requerimientos.

Palabras clave: Estrategias Didácticas, Educación, Ingeniería de Software, Ingeniería de Requerimientos

Contexto

La línea de investigación se desarrolla en el Instituto de Desarrollo Económico e Innovación (IDEI) de la UNTDF, por docentes-investigadores que llevan adelante proyectos en la temática. Actualmente ésta línea se presenta como un proyecto de investigación y se espera derive en una tesis de postgrado.

1 Introducción

El proceso de enseñanza-aprendizaje se caracteriza, por reconocer el papel determinante del docente en la identificación, planificación e instrumentación de estrategias adecuadas en la formación de los estudiantes. Las estrategias de enseñanza y de aprendizaje se encuentran involucradas de forma que el aprendizaje no sea solo

la adquisición de conocimientos, sino también la búsqueda de "*...los medios que conducen a la solución de los problemas...*"[1], donde el estudiante de forma consciente elige, planifica, coordina, aplica métodos y procedimientos para conseguir un objetivo o aprendizaje.

El término estrategia se utiliza cada vez con mayor frecuencia y presupone la planificación de acciones a corto, mediano y largo plazo; permitiendo lograr la racionalidad de tiempo, recursos y esfuerzos. Son indiscutibles las ventajas que su adecuada utilización puede ofrecer en los procesos educativos[2].

La didáctica exige la utilización de estrategias y métodos adecuados, en la actualidad se reconoce la necesidad de una didáctica centrada en el sujeto que aprende, lo cual exige enfocar la enseñanza como un proceso de orientación del aprendizaje, para que los estudiantes no sólo se apropien de los conocimientos, sino que desarrollen habilidades, formen valores y adquieran estrategias que le permitan actuar de forma independiente. La expresión *estrategia didáctica* presupone "...enfocar cómo enseña el docente y cómo aprende el alumno a través de un proceso..."[3].

Entre los métodos de enseñanza-aprendizaje que en los últimos tiempos están siendo reconocidos en la didáctica y, que los docentes deben considerar, se encuentran: el aprendizaje basado en problemas, el aprendizaje basado en proyectos, el método de casos, las simulaciones, etc., los cuales están relacionados con la problemática a tratar.

En la enseñanza específica de la Ingeniería de Software, los futuros profesionales deben tener habilidades y aptitudes determinadas (como abstracción, trabajo colaborativo y cooperativo, negociación, comunicación, toma de decisiones, comprensión del problema, entre otras) que poco se orientan en la enseñanza tradicional[4]. En la actualidad la enseñanza no se basa sólo en los conceptos impartidos por el docente, sino que se inclina hacia el estudiante como eje central de la clase; los objetivos generales del aprendizaje distinguen tres elementos[5]:

- Conceptual: incremento del conocimiento teórico del saber de un área.
- Procedimental: la ampliación de conocimiento práctico del saber de un área.
- Integrador: Relacionado con el crecimiento de conocimiento sobre las destrezas, aptitudes y actitudes propias del ejercicio de una profesión.

Enseñar ingeniería de requerimientos es fundamental ya a que es la primera, y más importante, etapa de todo paradigma de ingeniería de software, según Brooks *"La parte más difícil de construir un sistema de software es decidir qué construir".* Con este proceso se obtiene una Especificación de Requerimientos (ERS) que se desarrolla sobre la base de un modelo conceptual y a representar abstracciones, supuestos y restricciones sobre el dominio de aplicación[6]. La especificación de requerimientos representa las necesidades, deseos y expectativas de los diferentes stakeholders.

Para esto se necesitan técnicas que se incorporen en el proceso de enseñanza en el área de ingeniería de software, más concretamente en el área de la ingeniería de requerimientos, con el propósito de crear estrategias de enseñanza que permitan abordar su complejidad inherente, la cual es una actividad de transferencia dificultosa cuando la fuente del conocimiento es un ser humano, y se requieren habilidades especiales (tales como el tiempo del usuario o dificultades psicológicas de la interacción)[7].

La necesidad mejorar las prácticas de Ingeniería de Requerimientos en los jóvenes profesionales ha sido mencionado en la literatura y se ha desarrollado como tema de estudio en las carreras de grado en los últimos años.

La dificultad de la obtención de los requerimientos radica fundamentalmente en la incertidumbre derivada del proceso de transferencia de información desde una o varias fuentes, no siempre consistentes. Esta dificultad inherente a la elicitación se ve incrementada cuando se considera el problema de la enseñanza en un curso universitario de las técnicas de elicitación de requerimientos (grupos focales, JAD, prototipos, entrevistas, escenarios, lluvia de ideas, etc).

Existen varios problemas para una buena enseñanza en las aulas de estas técnicas, el primero de ellos es la dificultad de ejecutar una práctica real: las técnicas son meramente descriptas y, en el mejor de los casos, practicadas en un caso simplificado.

Otro problema importante es la subestimación por parte de personas de formación tecnológica de las técnicas "blandas" (metodologías que tratan con las interacciones humanas y los procesos sociales) que requiere la elicitación de requerimientos.

Para la enseñanza de estos conceptos se percibe que las estrategias tradicionales (simples narrativas descriptivas de un contexto sin "interacción directa con un usuarios") no son suficientes. Resulta fundamental incorporar estrategias didácticas adecuadas a las particularidades y dificultades planteadas como casos pilotos que permitan mejorar las prácticas y los aprendizajes de los estudiantes en los procesos que intervienen en la ingeniería de requerimientos.

Las estrategias didácticas permiten en la enseñanza de las técnicas de elicitación y especificación de requerimientos, minimizar los problemas planteados en los párrafos anteriores.

2 Líneas de Investigación y Desarrollo

Dentro de la línea de investigación general que se enfoca el proyecto se presente abarcar el siguiente objetivo general:

"Diseñar estrategias que permitan promover nuevas propuestas didácticas en la enseñanza de las técnicas de elicitación, especificación y validación de la ingeniería de requerimientos."

Para cumplir este objetivo general, la investigación plantea los siguientes objetivos específicos:

1.- Sintetizar el estado del arte respecto a la situación actual de la enseñanza de ingeniería de requerimientos en otras universidades.

2.- Diseñar diversas estrategias didácticas en el marco y contexto de la asignatura de Ingeniería de software de la carrera de Licenciatura en Sistemas de la UNTDF.

3.- Definir la/s estrategias que mejor se adecúan a las posibilidades aúlicas y los tiempos de la asignatura.

4.- Implementar la/s estrategias seleccionadas en el curso de Ingeniería de Requerimientos en dos ciclos lectivos (2021 y 2022)

5.- Comparar las experiencias y producciones obtenidas respecto a los años previos.

3 Hipótesis General del trabajo

Como hipótesis general de la investigación se plantea que las estrategias didácticas tradicionales de enseñanza en la ingeniería de requerimientos (simples narrativas descriptivas de un contexto particular), dificultan al alumno comprender cabalmente los problemas que presentan las interacciones con los usuarios reales. Poder implementar estrategias adecuadas con prácticas más cercanas a la realidad, permitirá mejorar la enseñanza de los docentes, y sobre todo, el aprendizaje de los alumnos en los procesos involucrados en la ingeniería de requerimientos.

4 Resultados Esperados

Si bien el proyecto de investigación se está iniciando, el mismo es incipiente y los resultados del proyecto permitirán aportar nuevas estrategias a la enseñanza de la ingeniería de requerimientos, principalmente poder ejecutar prácticas reales, es decir, con usuarios con problemas reales. Poder incorporar estrategias didácticas adecuadas que permitan a los alumnos tener contacto con posibles usuarios, a través de diversos recursos tecnológicos, que permitan mejorar las prácticas y los aprendizajes de los estudiantes en los procesos que intervienen en la ingeniería de requerimientos, como el uso de de videos-conferencia, grabaciones, wikis, foros, etc, permitirá lograr el objetivo propuesto.

5 Formación de Recursos Humanos

El equipo de trabajo está conformado por docentes investigadores de la UNTDF. Se encuentran en desarrollo dos tesis de Magister en Ing. de Software, UNLP, Ezequiel Moyano, Martín Villarreal y dos tesis de Magister en Tecnología Informática aplicada a la Educación, UNLP, Jerez Marcela y Lucas Romano.

6 Bibliografía

1. Burón, J. Aprender a aprender: Introducción a la metacognición.: Ed. Mensajero; 1994, p.94.
2. Montes de Oca, N. Estrategias docentes y métodos de enseñanza-aprendizaje en la Educ. Superior; 2011.
3. Fernández, A.; Ginoris, F. Didáctica y optimización del proceso de enseñanza-aprendizaje. La Habana: Instituto Pedagógico Latinoamericano y Caribeño (IPLAC); 1998.
4. Zapata-Jaramillo, C.M; Awad-Aubad G., "Requirements Game: Teaching Software Project Management," CLEI Electronic Journal, vol.10 (1), 2007.
5. Guitart M.; Rodríguez, E.; J. Cabot; "Elección del modelo de evaluación: caso práctico para asignaturas de ingeniería del software," Jornadas Enseñanza Univ. Informática, 2006.
6. Loucopoulos P., Karakostas V., *Systems Requirements Engineering*. McGraw-Hill, 1995.
7. Oliveros A.; Aprendizaje de la elicitación y especificación de requerimientos, 2014.

Técnicas de ciencias de datos para el análisis de la teoría moderna de optimización de carteras de inversión en el marco del mercado argentino de valores

Rodrigo Gonzalez[1,2], Carlos A. Catania[1,3], and Santiago Eguren[4]

[1] Facultad de Informática y Diseño, Universidad Champagnat, Belgrano 721, M5501, Mendoza
{gonzalezrodrigo,cataniacarlos}@uch.edu.ar
[2] GridTICs, Universidad Tecnológica Nacional, Rodriguez 273, M5502AJE, Mendoza
[3] LABSIN, Facultad de Ingeniería, Universidad Nacional de Cuyo, Padre Jorge Contreras 1300, M5502JMA, Mendoza
[4] Facultad de Ciencias Económicas, Universidad Nacional de Cuyo, Padre Jorge Contreras 1300, M5502JMA, Mendoza

Resumen La teoría moderna de configuración de carteras de inversión (*Modern Portfolio Theory*, MPT) propone diseñar una cartera que proporcione la máxima rentabilidad asumiendo una cantidad de riesgo determinada. En general, la MPT ha sido fuertemente analizada en mercados de países industrializados. El mercado argentino de valores (MAV) es un mercado cuya volatilidad está en sintonía con la de los mercados internacionales, pero que también responde a circunstancias propias. Además, a menudo la inflación no es tenida en cuenta cuando se analiza el desempeño de una cartera de inversión en mercados bursátiles de países industrializados. Contrariamente, esta variable debe tenerse en cuenta cuando se desea formar una cartera de inversión con activos del MAV. La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de un sistema a partir de sus datos. Además, provee herramientas para la visualización conveniente de todos los aspectos relevantes sobre el sistema que se trata de modelar. Por tanto, este proyecto propone determinar diferentes carteras de inversión óptimas con activos argentinos tomando los últimos 10 años para distintos niveles de riesgo, según la teoría moderna de configuración de carteras de inversión, con técnicas del análisis cuantitativo (*quantitative analysis*) y de ciencia de los datos.

Keywords: Modern Portfolio Theory · Análisis cuantitativo · Ciencia de datos · Merval.

1. Introducción

La teoría moderna de configuración de carteras de inversión (*Modern Portfolio Theory*, MPT) argumenta que es posible diseñar una cartera óptima que

proporcione a un inversor la máxima rentabilidad asumiendo una cantidad de riesgo determinada. La MPT fue desarrollada por el economista Harry Markowitz [6]. Esta teoría destaca la importancia de las carteras de inversión, el riesgo, la diversificación y las conexiones entre los diferentes tipos de activos. Los activos enfrentan tanto riesgos sistemáticos (riesgos de mercado) como riesgos no sistemáticos, propios de cada activo. La diversificación adecuada de una cartera no puede prevenir el riesgo sistemático, pero puede amortiguar, si no eliminar, el riesgo no sistemático. En general, la MPT y sus teorías derivadas [7,11,12,4] han sido fuertemente analizadas en mercados de países industrializados [5,10]. El mercado argentino de valores (MAV) es un mercado cuya volatilidad está en sintonía con la de los mercados internacionales, pero que también responde a circunstancias propias, como la gran caída que sufrió el índice Merval en agosto de 2019. Por otra parte, en mercados de países industrializados generalmente los bonos de un país son considerados instrumentos de baja volatilidad. En el caso del MAV, los bonos del estado argentino han mostrado en los últimos años una volatilidad alta cuando se los compara con instrumentos similares de otros países [8]. Además, en el contexto del MAV es necesario tener presente los efectos de la inflación en una cartera de inversión [13,3,1]. En general, la inflación no es tenida en cuenta cuando se analiza el desempeño de una cartera de inversión en mercados bursátiles de países industrializados. Por lo expuesto, el MAV presenta ciertas características que lo hacen particular. El MAV puede considerarse formado por dos grandes mercados financieros, ByMA [2] y ROFEX [9], donde el primero se especializa en acciones y bonos, y el segundo en contratos de futuros y opciones. Así, el MAV está compuesto por acciones nacionales y del mercado de los EE.UU. (CEDEAR), bonos nacionales y provinciales, bonos corporativos (ON, obligaciones negociables), contratos de futuros y opciones (*puts* y *calls*) sobre acciones, bonos, commodities (cereales, carne vacuna, petróleo, metales), cotización del dólar (dólar futuro) e índices [9]. Si se considera la cotización de todos estos activos en los últimos 10 años, se estima que el total de instrumentos de inversión puede llegar a 500. Por tanto, para hacer un análisis exhaustivo para la confección de una cartera de activos óptima en el MAV tomando los cotizaciones de los últimos 10 años, es necesario utilizar herramientas que permitan acceder, analizar, almacenar y visualizar una gran cantidad de información. Respecto a los datos requeridos, estos no están concentrados en una sola fuente, por lo que es necesario recurrir a diferentes APIs para acceder a datos de los mercados ByMA y ROFEX. La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados. Además emplea técnicas y teorías extraídas de muchos campos dentro del contexto de las matemáticas, la estadística, la ciencia de la información y la informática. Es una continuación de algunos campos de análisis de datos como la estadística, minería de datos y aprendizaje automático. Se la define como ün concepto para unificar estadísticas, análisis de datos, aprendizaje automático, y sus métodos relacionados, a efectos de comprender y analizar los fenómenos reales". La ciencia de datos es una herramienta muy utilizada en

finanzas. Permite manejar la gran cantidad de datos que proveen los mercados financieros diariamente, y analizar las intrincadas relaciones que pueden existir entre diferentes activos. Además, provee herramientas para la visualización conveniente de todos los datos relevantes sobre el fenómeno financiero que se desea modelar. Por lo expuesto, este proyecto propone analizar el comportamiento del MAV en los últimos 10 años con técnicas del análisis cuantitativo (*quantitative analysis*) y la ciencia de los datos, para determinar diferentes carteras de inversión óptimas con activos argentinos para distintos niveles de riesgo, según la teoría moderna de configuración de carteras de inversión.

2. Objetivos de la investigación

La hipótesis central de este proyecto se basa en que debido a las particularidades del mercado argentino de valores, aplicar la teoría moderna de creación de carteras de inversión puede arrojar resultados muy diferentes cuando se los compara con los tipos de carteras que se obtienen al analizar mercados de valores de países industrializados.

2.1. Objetivo principal

Analizar el comportamiento de la teoría moderna de carteras de inversión en el mercado argentino de valores en los últimos 10 años con técnicas de ciencia de datos, para de esta forma poder conformar diferentes carteras de inversión óptimas para diferentes niveles de riesgo.

2.2. Objetivos específicos

Este proyecto propone 3 objetivos específicos:

1. Entender los alcances y limitaciones de la teoría moderna de carteras de inversión (MPT).
2. Crear un entorno computacional para adquisición, análisis, almacenamiento y visualización de datos de mercados de valores.
3. Determinar diferentes carteras de inversión óptimas según diferentes niveles de riesgo para el mercado argentino de valores entre los años 2010 y 2020.

3. Resultados esperados

Al término del tiempo de duración de este proyecto, se espera obtener los siguientes resultados:

1. Sólida comprensión de los beneficios y limitaciones de la teoría moderna de creación de carteras de inversión.
2. Creación de un entorno computacional para el análisis y la visualización de carteras de inversión con información de los mercados de valores argentino y estadounidense.

3. Mejor entendimiento de las particularidades del mercado argentino de valores, respecto al mercado de valores de los EE.UU.
4. Obtención de diferentes tipos de carteras de inversión para el mercado argentino, según el nivel de riesgo que desee tomar un inversor en particular.

Los resultados encontrados durante la ejecución de este proyecto de investigación serán difundidos en el ámbito de la Universidad Champagnat. Se planea dictar una serie de seminarios en el marco de la Maestría de Negocios y Tecnología (MBT-UCH) relacionados con programación en R y Python para Finanzas y análisis de carteras de inversión en los mercados argentino y estadounidense. Además, los resultados obtenidos serán publicados tanto en conferencias internacionales de alta calidad con referato, como en revistas especializadas de alto impacto, parte del catálogo de Scopus y Web of Science.

Referencias

1. Brito, Mateo y Goldrossen, Ronen. Instrumentos de cobertura contra la inflación argentina. Tesis de grado. Departamento de Economía. Universidad de San Andrés. Diciembre de 2018.
2. Bolsa y Mercados Argentinos S.A. CABA, Argentina. Sitio web: https://www.byma.com.ar/.
3. Dapena, José P. La evolución del índice Merval (precio de acciones) y la inflación. Newsletter CEA, Dpto. Finanzas. Universidad del Cema. Marzo de 2013.
4. Fama, Eugene and French, Kenneth. "Common risk factors in the returns on stocks andbonds" Journal of Financial Economics Volume 33, Issue 1, February 1993, Pages 3-56.
5. Konno, H. and H. Yamazaki (1991). "Mean-Absolute Deviation Portfolio Optimization Model and Its Application to Tokyo Stock Market."Management Science. Vol. 37, No. 5, May 1991, pp. 519–531.
6. Markowitz, H. (1952) "Portfolio Selection."Journal of Finance. Vol. 7, No. 1, March 1952, pp. 77–91.
7. Markowitz, H. M. (1959). Portfolio Selection: Efficient Diversification of Investments. John Wiley & Sons, Inc., 1959.
8. Rodriguez, J.N., Bueno, J.G., Barrios, G.R., y Díaz, S. (2017). Diversificación internacional de portafolio en los mercados accionarios de Argentina, Brasil, Chile, Colombia, México y Perú. En el Congreso Internacional en Administración de Negocios Internacionales (CIANI 2017). Septiembre de 2017, pp. 355-375.
9. ROFEX S.A. Ciudad de Rosario, Santa Fe, Argentina. Sitio web: https://www.rofex.com.ar/.
10. Schanzenbach, M. M., and Sitkoff, R. H. (2017). The prudent investor rule and market risk: an empirical analysis. Journal of Empirical Legal Studies, 14(1), 129-168.
11. Sharpe, William F. (1964). "Asset Prices: A Theory of Market Equilibrium under-Conditions of Risk". The Journal of Finance, Vol. 19, No. 3 pp. 425-442.2.
12. Sharpe, William F. (1994). "The Sharpe Ratio". The Journal of Portfolio Management.21 (1): 49–58.3.
13. Spotorno, Fausto y Aranda, Alejandro Javier. Portafolio de cobertura contra la inflación en Argentina. Tesis de Maestría. Universidad del Cema. Diciembre de 2008.

Inteligencia artificial en diagnósticos médicos

Emmanuel Ricardo Orozco Gazzo

Laboratorio de Investigación en Ciencia y Tecnología (LabICyT-UDA),
Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua,
Catamarca 147, Mendoza, Argentina
emmanuelo483@gmail.com
https://www.uda.edu.ar/index.php.

Resumen. El presente trabajo muestra el desarrollo de una aplicación que permita clasificar las imágenes radiográficas de tórax de pacientes de diferentes patologías respiratorias. Este sistema tiene dos partes fundamentales: La primera es el módulo en el que se hace la detección de las imágenes torácicas para analizar. Para esta se aplica un algoritmo de detección de bordes para obtener solo el contorno del objeto a analizar, la red neuronal se encarga de ser el detector de patrones, por medio de un entrenamiento previo con imágenes anteriormente cargadas en una base de datos. Después de que la red neuronal ha sido entrenada y trabaja correctamente, La segunda parte: Diferencia entre radiografías de neumonías y normales.

Keywords: Inteligencia Artificial· aprendizaje profundo· radiografías· neumonía

1 Introducción

Más de 1 millón de personas enferman y son hospitalizadas con neumonía y alrededor de 50.000 mueren a causa de la enfermedad cada año solo en los E.E.U. U [1]. Las radiografías de tórax son actualmente el mejor método disponible para diagnosticar la neumonía y otras enfermedades respiratorias [2]. Desempeñando un papel crucial en la atención clínica [3] y estudios epidemiológicos [4]. Sin embargo, la detección de neumonía en las radiografías de tórax es una tarea desafiante que depende de la disponibilidad de radiólogos expertos, situación que origina un gran costo en recursos humanos para la atención de dichos pacientes por lo que se tratara de disminuir dichos costos con la utilización de una inteligencia artificial (IA).

La radiografía de tórax (CXR) se está volviendo crucial para seguir la evolución clínica de los pacientes. El elevado número de casos de covid19 [5] detectados en todo el mundo aumenta la necesidad de una herramienta automática para detectar los signos pulmonares (neumonía o infiltración) en la CXR de pacientes, lo que a su vez reduciría la carga de trabajo del especialista y les permitiría enfocar sus esfuerzos en los pacientes más críticos [6]. LA IA será un aliado que les liberará también de algunas tareas monótonas, como el análisis de las imágenes médicas, el problema al que se enfrentan los médicos es el gran volumen de datos que se debe revisar. Necesitando mejorar la forma de manejo de la información,

sería una herramienta para ofrecer diagnósticos en países en vías de desarrollo o zonas remotas donde no hay suficiente personal médico tan cualificado.

2 Descripción del problema

En este trabajo se analiza si la IA puede discriminar registros y determinar que pacientes tendrían riesgos elevado de sufrir infecciones agregadas a COVID-19 y cuáles no. Teniendo en cuenta parámetros vitales y registros electromagnéticos (RX de tórax) [6] de pacientes con síntomas compatibles a COVID-19, si aplicamos a la IA para entender el desarrollo del organismo humano y mejorar la discriminación de dichos pacientes para una mejor detención de la taza de morbimortalidad. Este estudio va a depender de la disponibilidad de las imágenes por lo tanto primero se trabajará con casos confirmados de neumonías y otras enfermedades respiratorias. Pretendemos en este trabajo, realizar un modelo que pueda detectar automáticamente neumonía a partir de radiografías de tórax a un nivel que supera la práctica. Utilizaremos un motor de IA para realizar un análisis comparativo de imágenes múltiples, de registros electrónicos de salud (RX), discriminando los pacientes con riesgos elevados a tener infecciones agregadas a covid19, de acuerdo a la información obtenida en las diferentes imágenes. Para tareas en las que interviene la imagen médica, se ha demostrado que el paradigma del aprendizaje profundo (DL) es un poderoso instrumento [7]. Se ha demostrado que el rendimiento de los modelos basados en esta tecnología puede mejorarse mediante el uso de algunos pasos de procesamiento de imágenes [8] [9] [10] [11] con el fin de pre- entrenar modelos y adquirir experiencia en el objetivo de detectar casos de covid19, utilizaremos modelos DL para clasificar a los pacientes neumonía (P), normales (N). Estos modelos se actualizarán para adaptarse a la tarea original (detección de COVID-19), cuando los CXR COVID-19 estén disponibles.

3 Inteligencia Artificial

A medida que la ciencia y la tecnología han ido avanzado, el objetivo se ha ido perfeccionando. Uno de los retos más importantes a los que se enfrenta el ser humano de nuestra generación es el de la construcción de sistemas inteligentes. Aquí sistema puede ser entendidos como cualquier dispositivo físico o lógico capaz de realizar la tarea requerida. Este es precisamente el objetivo de la disciplina científica conocida con el nombre de inteligencia artificial [12] Se consideran inteligentes porque dotan a las computadoras con cierta capacidad de imitar las habilidades cognitivas del ser humano. La parte más importante de una red de neuronas artificiales es el aprendizaje. El esquema de aprendizaje de una red es lo que determina el tipo de problemas que será capaz de resolver [12]. Los métodos basados en la inteligencia artificial (IA) han surgido como herramientas poderosas para transformar la atención médica. Aunque los clasificadores de aprendizaje automático (CAA) ya han demostrado un alto rendimiento en los diagnósticos basados en imágenes, el análisis de datos de registros electrónicos de

salud (RES) diversos y masivos sigue siendo un desafío. Aquí, mostramos que los CAA pueden consultar RES de una manera similar al razonamiento hipotético-deductivo utilizado por los médicos y descubrir asociaciones que los métodos estadísticos anteriores no han encontrado [13]. Utilizando modelo aplica un sistema automatizado de procesamiento del lenguaje natural que utiliza técnicas de aprendizaje profundo para extraer información clínicamente relevante de los RES.

1. La inteligencia artificial es capaz de identificar patrones ocultos en los datos que pueden ayudar a un médico a prescribir tratamientos más personalizados

2. Puede encontrar información que indique si una persona es apta o no para un ensayo clínico.

3. Reduce el tiempo que un médico dedica a leer literatura médica para dejarle más tiempo para tratar a sus pacientes. En este estudio, se utiliza una estrategia de colaboración humano-máquina creando un programa para la detección de casos de Neumonía usando Imágenes RX. Usando el lenguaje de programación Python con la plataforma de tensorflow y el api kerass se importan las librerías necesarias para entrenar el modelo, procesar imágenes, crear las capas convoluciones y las redes neuronales secuenciales. Se buscan rutas de imágenes de entrenamiento y validación, colocando la altura de las imágenes y la longitud a 224,224 se re escalan las imágenes para que varíen del 0 al 1 los pixeles y se colocan varios filtros en el conjunto de entrenamiento y en el conjunto de validación. Solo se re escala, clasificamos las imágenes de entrenamiento y validación en el modo categorical. Cargamos el modelo pre entrenado, cargamos sus capas en nuestro modelo secuencial y procedemos a quitar la última capa. Añadimos dos capas una sigmoidal y otra softmax a través de esto queremos lograr que se entrenen solo las últimas dos capas, para lo cual se compila nuestro modelo usando Adam. Una vez entrenado este modelo es guardado. Utilizando una función creada que importa el modelo guardado anteriormente, esta puede diferenciar las imágenes discriminándolas entre neumonía y normales, seguimos trabajando en el entrenamiento de dicho modelo para lograr resultados favorables ya que hasta el momento no se cumplen las expectativas esperadas.

4 Datos

Usamos el conjunto de datos Chest X-ray [14] que contiene más de 17.000 imágenes de rayos X de vista frontal dichas imágenes se dividen 2 etiquetas las de patologías torácicas y las normales. Utilizando métodos de extracción automática en los informes de radiología. Etiquetamos las imágenes que tienen neumonía fig [1] como una de las patologías anotadas como ejemplos positivos y etiquetamos todas las demás imágenes normales fig [2] como ejemplos negativos para la tarea de detección de neumonía. Dividimos aleatoriamente todo el conjunto de datos en un 80 de entrenamiento y un 20 de validación.

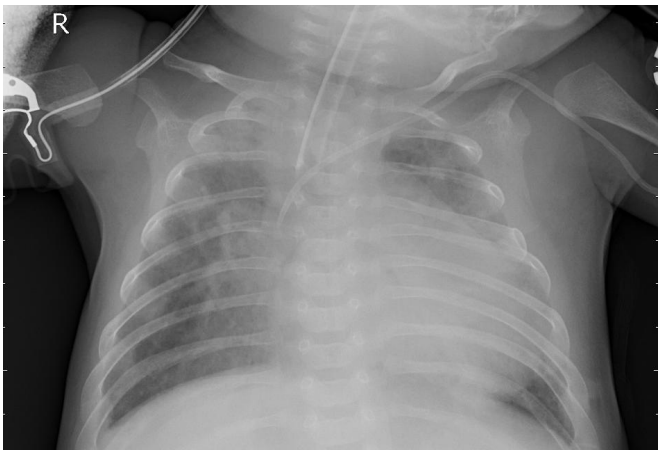


Fig. 1. Una imagen del conjunto que tiene la etiqueta neumonía

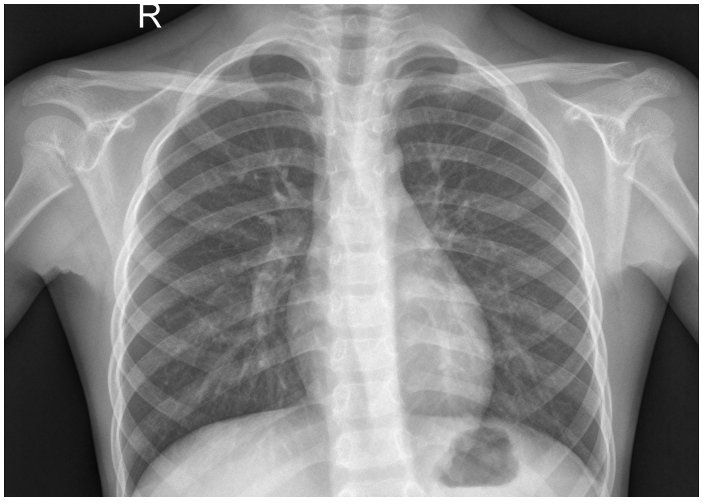


Fig. 2. Una imagen del conjunto que tiene la etiqueta normal

5 Conclusión

El presente trabajo muestra el desarrollo de una aplicación que permitiría clasificar y discriminar las imágenes radiográficas de tórax de pacientes de diferentes patologías respiratorias, para ello se aplica un algoritmo y entrenando previamente a la IA con imágenes anteriormente cargadas, para diferencia entre radiografías de neumonía y normales. Podemos concluir que esta aplicación tendría un enorme potencial ya que permitiría el correcto aprovechamiento, de recursos humanos en salud, siendo de gran beneficio a los profesionales de salud, ya que disminuiría el tiempo de trabajo de los médicos originado veneficios económicos.

Referencias

1. Centros para el Control y la Prevención de Enfermedades, 2017
2. (OMS, 2001)
3. (Franquet , 2001)
4. Aydogdu, M, Ozyilmaz, E, Aksoy, Handan, Gursel, G y Ekim, Numan. Predicción de la mortalidad en la neumonía adquirida en la comunidad que requiere ventilación mecánica; valores de neumonía y puntuaciones de gravedad de la unidad de cuidados intensivos. Tuberk Toraks, 58 (1): 25– 34 de 2010.
5. Organizacion mundial de la salud Manejo clínico de la infección respiratoria aguda grave presuntamente causada por el nuevo coronavirus (2019-nCoV) (28/01/2020) Ultimo acceso 18 Sept 2020. https://www.paho.org/hq/index.php?option=com$_docmanview = downloadalias = 51723 - manejo - clinico - de - la - infeccion - respiratoria - aguda - grave - presuntamente - causada - por - el - nuevo - coronavirus - 2019 - ncovcategory_slug = materiales - cientificos - tecnicos - 7992Itemid = 270lang = esInstitutoTecnolCOVID - 19 - Detecciónt empranaenimágenesderayosXdetórax - 9deabrilde2020$
6. Fang, Y. et al. Sensibilidad de la TC de tórax para covid-19: Comparación con RT-PCR. Radiología 200432 (2020).
7. D. Rav'i', C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo y G.-Z. Yang, "Profundo aprendizaje para la informática de la salud" IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, págs. 4 a 21, 2016. 1
8. RM Sundhari, "Mejora de nódulos y red neuronal basada en ecualización de histograma mejorada detección basada en radiografías de tórax", Revista de Inteligencia Ambiental y Computación Humanizada, págs. 1–9, 2020. 1
9. Y. Gordienko, P. Gang, J. Hui, W. Zeng, Y. Kochura, O. Alienin, O. Rokovyi y S. Stirenko, "Aprendizaje profundo con segmentación pulmonar y técnicas de exclusión de sombras óseas para el análisis de rayos X de tórax del cáncer de pulmón", en Conferencia internacional sobre aplicaciones de la informática, la ingeniería y la educación, págs. 638–647, Springer, 2018. 1
10. AK Bhandari, "Un esquema de modificación de histograma basado en la ley logarítmica para la imagen de naturalidad mejora de contraste," Revista de Inteligencia Ambiental y Computación Humanizada, vol. 11, no. 4, págs. 1605–1627, 2020. 1
11. R. Sivaramakrishnan, S. Antani, S. Candemir, Z. Xue, J. Abuya, M. Kohli, P. Alderson y G. Thoma, "Comparación de modelos de aprendizaje profundo para el cribado de población mediante radiografía de tórax", en Medical Imagin 2018: Diagnóstico asistido por computadora, vol. 10575, pág. 105751E, Sociedad Internacional de Óptica y Fotónica, 2018. 1

12. Autor,P. Isasi: Redes de neuronas artificiales: un enfoque practico. Libro Pearson Educacion, 2004
13. Dávila Newman, Gladys El razonamiento inductivo y deductivo dentro del proceso investigativo en ciencias experimentales y sociales Laurus, vol. 12, núm. Ext, 2006, pp. 180-205 Universidad Pedagógica Experimental Libertador Caracas, Venezuela Ultimo acceso 18 Sept 2020. https://www.redalyc.org/pdf/761/76109911.pdf
14. Stefan Jaeger, 1 Sema Candemir, 1 Sameer Antani, 1 Yì-Xiáng J. Wáng, 2 Pu-Xuan Lu, 3 y George Thoma 2014 Dec; Ultimo acceso 18 Sept 2020. https://www.kaggle.com/kmader/pulmonary-chest-xray-abnormalities

Integración de una BCI para rehabilitación de problemas cognitivos en pacientes neurológicos infantiles implementando VRPN como protocolo de comunicaciones

Javier J. Rosenstein, Rodrigo Gonzalez, Nicolás Ignacio Zárate Alvarez, Matías Campos

Instituto de Investigaciones, Facultad de Informática y Diseño, Universidad Champagnat, Belgrano 721, 5501 Godoy Cruz, Mendoza, Mendoza, Argentina.

rosensteinjavier@uch.edu.ar, gonzalezrodrigo@uch.edu.ar, nicolaszarate23@gmail.com, mgcampos9785@gmail.com

RESUMEN

En el desarrollo de sistemas de realidad virtual uno de los inconvenientes que se encuentran es la comunicación entre las aplicaciones y los dispositivos de adquisición de ondas cerebrales. Ya sea por no disponer de un método de acceso en forma directa o por necesitar independencia entre ambos, es decir que las aplicaciones corran en una plataforma y los dispositivos en otras. Para lograr esta independencia y a su vez permitir la integración de todo el sistema de realidad virtual, es necesario la implementación de algún protocolo de comunicaciones que permita esta vinculación heterogénea en tiempo real. Los dispositivos generalmente están asociados a funciones o características de los individuos que los utilizan y se necesita integrar los movimientos que estos representan a la aplicación de realidad virtual correspondiente. El presente trabajo trata del análisis e implementación del protocolo de comunicaciones VRPN (Virtual Reality Protocol Network) entre las partes de un entorno multimedia donde interactúan la adquisición de movimientos del usuario y la representación visual en un escenario virtual que permita la retroalimentación al usuario en tiempo real logrando una experiencia interactiva e inmersiva.
Palabras clave: VRPN, EOG, EEG, BCI, Serious Games, Mirror Neurons, Neurofeedback

CONTEXTO

El presente proyecto forma parte del trabajo de tesis de Maestría en Teleinformática de la Universidad de Mendoza (Ciudad, Mendoza), correspondiente al tesista Javier Rosenstein, el mismo se desarrolla en el Instituto de Investigaciones de la Facultad de Informática y Diseño de la Universidad Champagnat (Godoy Cruz, Mendoza), en el marco de la Licenciatura en Sistemas de Información; en cooperación con el Laboratorio de I+D+i en Neurotecnologías de la empresa Neuromed Argentina S.A. (Godoy Cruz, Mendoza).

Este trabajo es parte del proyecto de investigación que dio inicio en Diciembre de 2018 denominado “Diseño y desarrollo de un prototipo de Serious Game destinado a la rehabilitación de problemas neurológicos implementando VRPN para la comunicación de la BCI”.

1. INTRODUCCIÓN

El objetivo principal de la presente línea de investigación consiste en el diseño y desarrollo de una BCI (Brain Computer Interface) [1] [2], que permita interactuar entre las señales generadas por un paciente neurológico [3] y una interfaz de realidad virtual. Este sistema debe permitir lograr el principio de neurofeedback [4], o retroalimentación hacia el paciente. De este modo se podrán mejorar sus capacidades cognitivas correspondientes. Esto tiene aplicación directa en los tratamientos de rehabilitación en pacientes de patologías neurológicas y cognitivas, principalmente en niños, este enfoque es importante ya que está demostrado que es la mayor causa de discapacidad en niños, según estadísticas como las publicadas por el Indec de su último informe del 2018 (ver figura1), en argentina dentro de las principales discapacidades, la mental-cognitiva es la mayor de las discapacidades entre las edades de 6 a 14 años.

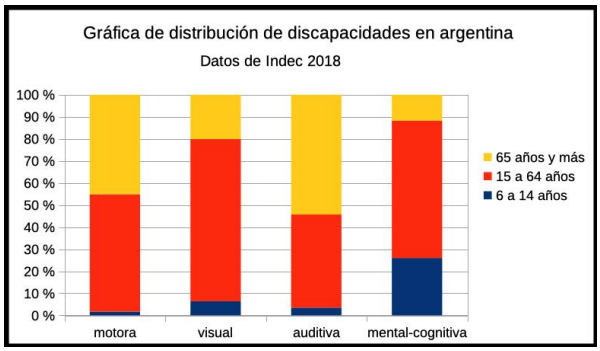


Figura 1: Discapacidades según Indec en argentina

Incluso, estudios indican la posibilidad de tratar patologías psiquiátricas como la depresión [5] [6].

La implementación del trabajo se organiza de acuerdo a las siguientes etapas:

a) Adquisición de señales mediante técnicas de Electroencefalografía (EEG) [7] y adquisición de movimientos oculares mediante las técnicas de electrooculografía (EOG) [8] [9] [10] [11] [12].
b) Análisis de estas señales en tiempo real para poder identificar la voluntad de movimiento del individuo, así como la dirección del movimiento.
c) Transformar la voluntad de movimiento en comandos del protocolo VRPN (Virtual Reality Protocol Network) que permitan transmitir la información al componente software / hardware que la requiera.
d) Como continuación al resultado obtenido en el punto anterior, poder comandar una interfaz gráfica de aprendizaje o interfaz Cerebro / Computadora (BCI – Brain Computer Interface).

Finalmente se presenta la BCI como un sistema de adquisición de datos, procesamiento del protocolo serie a VRPN y luego la representación en nuestro modelo de prototipo de Serious Game [14] [15]. Una vez adquiridas las señales EEG/EOG, estas se analizan y codifican con las bibliotecas desarrolladas como parte de este proyecto. Esto permite su comunicación mediante VRPN con las interfaces virtuales que interpretan este protocolo.

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

Este trabajo está compuesto por tres etapas o fases de trabajo. La primera de ellas corresponde a la implementación de VRPN entre un sistema simulado de captura de datos provenientes de un paciente, y una interfaz virtual básica. La segunda etapa del proyecto pretende avanzar sobre la captura de datos reales para que, luego de procesados, se transfieran a través de VRPN hacia la interfaz virtual de neurofeedback. Una breve descripción de cada una de las etapas se describe a continuación:

1. La primer parte consiste en la implementación de un simulador de señales EEG y EOG necesarias para el análisis e interpretación de la voluntad del usuario de la BCI. Estas señales una vez procesadas se deben codificar en comandos de VRPN para poder ser transmitidas hacia una interfaz virtual. Esta debe poder interpretar las señales transmitidas y representar la voluntad inicial del usuario correspondiente. De este modo se cumplen los objetivos de captura, análisis, procesamiento, transmisión, recepción y representación, lo cual produce el efecto de neurofeedback deseado sobre una interfaz virtual de capacitación a nivel prototipo.
2. La segunda etapa consiste en el desarrollo y mejora del proyecto mediante la adquisición real de señales EEG y EOG por electrodos ubicados superficialmente sobre la cabeza del paciente. Esta modificación al sistema requiere de un diseño e implementación electrónica así como del desarrollo del firmware que permita adquirir, analizar y procesar estas señales, que luego serán transmitidas a través del protocolo VRPN hacia una interfaz virtual. Así, el sistema diseñado se incorporará en forma transparente al proyecto implementado en la primer etapa.
3. La etapa final de este proyecto consiste en el diseño y desarrollo del escenario virtual de rehabilitación cognitiva para pacientes neurológicos según las indicaciones concretas por parte del especialista en neurología. Se partirá de un relevamiento de las técnicas de aprendizaje que se requieren implementar y los resultados que se pretenden obtener, indicados por el neurólogo o experto afín.

3. RESULTADOS ESPERADOS

Los resultados esperados se pueden dividir en dos grandes grupos:

1. BCI e interfaz virtual de neurofeedback: se espera obtener como producto final un sistema de retroalimentación a un supuesto usuario simulado por la generación de señales que comandan el escenario de realidad virtual via comandos codificados en el protocolo VRPN que se transmiten a través de una red Ethernet.

2. Adquisición de datos reales de EEG y EOG: análisis y preprocesamiento; se identifican los comandos necesarios para su transmisión hacia el equipo generador de tramas VRPN para que finalmente se tenga el proyecto totalmente terminado y funcional que implemente el efecto de neurofeedback propuesto.

4. FORMACIÓN DE RECURSOS HUMANOS

La línea de I+D presentada está vinculada con el desarrollo de una tesis de postgrado, por parte del estudiante de maestría en teleinformática de la Universidad de Mendoza, Javier J. Rosenstein. Dicha tesis se centra en la implementación del protocolo VRPN demostrando su uso en un sistema BCI.

Además cuenta con la dirección del Lic. Javier J. Rosenstein (UCH) y como investigador principal el Dr. Rodrigo Gonzalez (UCH). En lo que respecta a la formación de estudiantes de la licenciatura, esta línea de investigación cuenta con dos tesis de grado en curso, pertenecientes a los estudiantes Nicolás Ignacio Zárate Alvarez y Matías Campos, cuyos planes de tesis se encuentran específicamente dentro del marco de este proyecto. Ambos cursan la licenciatura en Sistemas de Información en la Universidad Champagnat.

Una vez que el proyecto se encuentre implementado desde su primer etapa será utilizado como recurso para el dictado de talleres de comunicaciones, redes, programación de microcontroladores, Programación en C/C++ y Diseño y programación de interfaces virtuales de capacitación/rehabilitación en general, tanto para estudiantes de la universidad, como así también para aquellos interesados externos a la UCH.

5. BIBLIOGRAFÍA

1. Y. Wang, X. Gao, B. Hong, and S. Gao, “Practical designs of brain-computer interfaces based on the modulation of EEG rhythms,” in Brain-Computer Interfaces. Springer, 2009, pp. 137–154.
2. J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, “Brain-computer interfaces for communication and control,” Clinical neurophysiology, vol. 113, no. 6, pp. 767–791, 2002.
3. J. A. Pineda, “The functional significance of mu rhythms: translating “seeing” and “hearing” into “doing”,” Brain Research Reviews, vol. 50, no. 1, pp. 57–68, 2005.
4. S. Enriquez-Geppert, R. J. Huster, and C. S. Herrmann, “Boosting brain functions: Improving executive functions with behavioral training, neurostimulation, and neurofeedback,” International Journal of Psychophysiology, vol. 88, no. 1, pp. 1–16, 2013.
5. R. Ramirez, M. Palencia-Lefler, S. Giraldo, and Z. Vamvakousis, “Musical neurofeedback for treating depression in elderly people.” Frontiers in neuroscience, vol. 9, pp. 354–354, 2014.
6. W. Rief, “Getting started with neurofeedback,” 2006.
7. J. D. Kropotov, Quantitative EEG, event-related potentials and neurotherapy. Academic Press, 2010.
8. A. Bulling, J. A. Ward, H. Gellersen, and G. Troster, “Eye movement analysis for activity recognition using electrooculography,” IEEE Transactions on pattern analysis and machine intelligence, vol. 33, no. 4, pp. 741–753, 2011.
9. H. Singh and J. Singh, “A review on electrooculography,” International Journal of Advanced Engineering Technology, vol. 3, no. 4, pp. 115–122, 2012.
10. D. P. Bautista, I. A. Badillo, D. De la Rosa Mejía, and A. H. H. Jiménez, “Interfaz humano-computadora basada en señales de electrooculografía para personas con discapacidad motriz,” ReCIBE, vol. 3, no. 2, 2016.
11. S. Yathunanthan, L. Chandrasena, A. Umakanthan, V. Vasuki, and S. Munasinghe, “Controlling a wheelchair by use of EOG signal,” in 2008 4th International Conference on Information and Automation for Sustainability, IEEE, 2008, pp. 283–288.
12. V. C. C. Roza, “Interface para tecnologia assistiva baseada em eletrooculografia,” 2014.

13. A. C. Gaviria, I. C. Miller, S. O. Medina, and D. R. Gonzales, “Implementación de una interfaz hombre-computador basada en registros EOG mediante circuitos de señal mixta PSoC,” in V Latin American Congress on Biomedical Engineering CLAIB 2011 May 16-21, 2011, Habana, Cuba. Springer, 2013, pp. 1194–1197.
14. P. Rego, P. M. Moreira, and L. P. Reis, “Serious games for rehabilitation: A survey and a classification towards a taxonomy,” in 5th Iberian Conference on Information Systems and Technologies. IEEE, 2010, pp. 1–6.
15. J. S. Breuer and G. Bente, “Why so serious? On the relation of serious games and learning,” Eludamos. Journal for Computer Game Culture, vol. 4, no. 1, pp. 7–24, 2010.
16. Instituto Nacional de Estadísticas y Censos, INDEC, Argentina, 2018, ISBN:978-950-896-519-6, https://www.indec.gob.ar/ftp/cuadros/poblacion/estudio_discapacidad_07_18.pdf

Redes bayesianas para el diagnóstico de COVID-19

Maschke Julián

Laboratorio de Investigación en Ciencia y Tecnología (LabICyT-UDA)
Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua.
julianmachke@gmail.com

Resumen La Inteligencia Artificial (IA) es una disciplina que intenta comprender y construir entidades inteligentes. El término se aplica cuando una máquina imita funciones cognitivas del ser humano relacionadas con otras mentes humanas con el fin de realizar una tarea. Existen diversas técnicas dentro de la IA las cuales hacen referencia a diferentes campos de investigación y de desarrollo de aplicaciones. Dentro de estas técnicas se encuentran las redes bayesianas o redes de creencia, que son un modelo probabilístico multivariado que relaciona un conjunto de variables aleatorias usando un grafo dirigido para indicar una influencia casual de manera explícita. Debido a la nueva pandemia causada por el COVID-19 el uso de las tecnologías ha tomado un papel vital e indispensable. Por este motivo, su correcta utilización puede ser de gran ayuda para combatir esta pandemia. En ese sentido, este proyecto busca diagnosticar casos de COVID-19 mediante Inteligencia Artificial y uso de redes bayesianas, ya que se trata de una herramienta muy útil para calcular probabilidades en casos de nuevas evidencias, lo que permitirá establecer patrones y estudiarlos.

Palabras clave Inteligencia artificial · COVID-19 · Redes Bayesianas.

1 Introducción

En medicina, cuando hablamos de diagnóstico o propedéutica clínica nos referimos al procedimiento por el cual se identifica una entidad nosológica, enfermedad, síndrome o cualquier estado de salud o enfermedad[1]. Actualmente, es de conocimiento público que la medicina utiliza indispensables adelantos relacionados con el uso intenso de alta tecnología. Una de estas herramientas corresponde a la aplicación de técnicas de Inteligencia Artificial [2], las que son sumamente útiles cuando aprovechan los datos disponibles y la experiencia clínica [3].ebido a esta relación y sinergia existentes entre la medicina y la alta tecnología, mediante la ejecución del presente proyecto se intentará realizar un aporte al área de diagnóstico de la enfermedad COVID-19 [4] a través el uso de la IA.

Una enfermedad compleja es aquella patología que puede ser causada por diversos factores y su interacción con el entorno, el contacto físico o con el medio ambiente, el consumo de alimentos, el uso de productos específicos, entre otros [5].

Prevenir el avance de estas enfermedades podría salvar muchas vidas. Para lograrlo, la población debe tener precaución y seguir protocolos de autocuidado para prevenir la propagación de la enfermedad. Por su parte, el personal de salud y el equipo médico debe contar con herramientas que faciliten la tarea de diagnosticar y atender a los pacientes de manera segura y eficaz. Debido a que el COVID-19 es considerada como enfermedad compleja, tanto por la inexistencia de vacunas o tratamientos como por el desconocimiento de su comportamiento, será de gran ayuda generar una técnica o herramienta capaz de realizar diagnósticos de esta enfermedad y evitar su propagación masiva.

La utilización de Inteligencia Artificial puede ser un método eficaz para prevenir contagios y evitar la rápida propagación del virus, conteniéndolo. A la hora utilizar la IA para diagnosticar enfermedades, estudios demuestran que las técnicas más apropiadas son[3]:

1) Las redes neuronales artificiales[6].

2) El razonamiento basado en casos[6].

3) Las redes bayesianas[8], [9].

Al momento de comparar estas técnicas, las redes bayesianas tienen ventajas versus las redes neuronales y el razonamiento basado en casos que las distinguen para ejecutar diagnóstico de enfermedades[3]. Según González Benítez et. al [3] algunas de estas ventajas son:

1) Poseen una interpretación sencilla, apoyada en métodos probabilísticos

2) Son robustas,debido a que pequeñas alteraciones en el modelo no afecta la estructura de la red.

3) Se pueden aportarr los conocimientos o creencias previas de un experto a través de la distribución a priori.

4) Están especialmente diseñadas para descubrir las relaciones entre variables que son causales y no pueden ser observadas de otro modo,lo que permite realizar predicciones más precisas.

5) No se afectan con el sobreajuste de los datos.

Teniendo en cuenta todos los puntos mencionados anteriormente, el objetivo de este proyecto será el de utilizar la técnica más apropiada para el análisis de enfermedades con el fin de diagnosticarlas correctamente, en este caso determinar los contagios por COVID-19. Debido a la ventaja que tienen las redes bayesianas, se usará este modelo para la resolución del problema.

2 Metodología

Para el desarrollo de este proyecto será necesario el uso de un software que nos permita crear y utilizar redes bayesianas junto con el conjunto de datos médicos que se utilizarán. Existen diversos softwares que permiten llevar a cabo la tarea de crear y utilizar redes bayesianas. Algunos de ellos han sido creados específicamente con el fin de manejar este modelo (como Elvira ó Ergo) mientras que otros han sido generados como extensiones de programas estadísticos ya existentes (como TETRAD ó Matlab) [8].
En este proyecto se utilizará el programa Netica [Norsys Software Corp.], debido a que esta especializado en redes bayesianas ademas de que ofrece un entorno de trabajo amigable. Este software es de uso libre y gratuito y puede ser obtenido desde la página de Norsys software corp [10]. En un principio los datos necesarios para el desarrollo de este proyecto serán extraídos desde la página oficial del Ministerio de Salud de la República Argentina. Dicho conjunto de datos tiene la característica de abarcar múltiples casos de diversas enfermedades haciendo énfasis en los síntomas.
En base a los síntomas se construirá un software en el cual el usuario pueda ingresar una cantidad indefinida de síntomas. Teniendo en cuenta los síntomas introducidos el sistema deducirá mediante el uso de redes bayesianas cual es el porcentaje de que la sintomología ingresada concuerde con un diagnóstico positivo de COVID-19. En el caso de que los síntomas se alejen de un posible caso de COVID-19, el sistema contará tambien con la capacidad de informar con que porcentaje puede tratarse de alguna otra enfermedad. El sistema podrá predecir alrededor de 10 enfermedades comunes en Argentina en base a la sintomología ingresada.

3 Conculsiones

La ejecución de este proyecto, permitirá el correcto aprovechamiento de los datos e información disponibles de este virus y la experiencia clínica adquirida durante la observación del comportamiento del mismo, con el fin de diagnosticar esta enfermedad para evitar su propagación y de esta manera aportar a la sociedad con el control del contagio y cuidado de la población.
El uso de redes bayesianas permitirá un correcto manejo de la data disponible debido a sus ventajas sobre otras tecnologías por la interpretación sencilla de los datos mediante métodos probabilísticos [8].

Referencias

1. Diagnóstico médico (en línea). Wikipedia, La enciclopedia libre. Ültimo acceso Sept 2020. https://es.wikipedia.org/wiki/diagnostico medico.
2. Julio Cesar Ponce Gallegos, Aurora Torres Soto, Fátima Sayuri Quezada Aguilera, Antonio Silva Sprock, Ember Ubeimar Martínez Flor, Ana Casali, Eliana Scheihing,

Yván Jesús Túpac Valdivia, Ma. Dolores Torres Soto, Francisco Javier Ornelas Zapata, José Alberto Hernández A, Crizpín Zavala D, Nodari Vakhnia, Oswaldo Pedreño.: Inteligencia Artificial. Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn), (2014)

3. González Benítez Neilys, Estrada Sentí Vivian, Febles Estrada Aylin.: Estudio y selección de las técnicas de Inteligencia Artificial para el diagnóstico de enfermedades. Rev Ciencias Médicas **22**(3), 131-141 (2018).
4. Preguntas frecuentes sobre el Nuevo coronavirus COVID-19. Argentina.gob.ar. Último acceso 15 Sept 2020. https://www.argentina.gob.ar/coronavirus/preguntasfrecuentes .
5. J.M. Soria, J.F. Dilmé, J. Martínez-Gonzalez, M. Camacho, C. Rodriguez, J.M. Romero, S. Bellmunt, J.R. Escudero, y L. Villa.: Métodos de Estudio de las Enfermedades Complejas: Aneurismas de la Aorta Abdominal. Angiología **62**(2),58-64, 2010
6. Pedro Isasi Viñuela, Inés. M. Galván León.: Redes de Neuronas Artificiales, un enfoque práctico. PEARSON EDUCATION, Madrid (2004)
7. Ochoa CA, Mena LJ, Felix VG, García V.: Diagnóstico para Síndrome de Asperger utilizando una aplicación inteligente a partir del razonamiento basado en casos. Revista Electrónica de Divulgación de la Investigación **10** (2015)
8. Jorge López Puga.: Cómo Construir y Validar Redes Bayesianas con Netica.: Revista Electrónica de Metodología Aplicada. Universidad de Almería. **17** (1), 1-17 2012 .: Disponible en: https://dialnet.unirioja.es/descarga/articulo/4107401.pdf .
9. Cheon S.-P., Kim S., Lee S.-Y., Lee C.-B.: Las redes bayesianas basaron la predicción de eventos raros con datos de sensores. Sistemas basados en el conocimiento **22**(5), 336–343 (2009).
10. Norsys software corp. Último acceso 18 Sept 2020. https://www.norsys.com/download.html .

Aplicación que te acerca al Reino Fungi junto al uso de la Realidad Aumentada

Giuliana Pessina

Laboratorio de Investigación en Ciencia y Tecnología (LabICyT-UDA),
Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua,
Catamarca 147, Mendoza, Argentina
Email: giulib.-@hotmail.com

Abstract. El Reino Fungi, el cual aún esta desbordado de misterios, cada dia nos llama más la atención, ya sea desde su biología hasta sus aspectos físicos. Para quitarnos la mayoria de dudas y curiosidades contamos con la tecnología y el desarrollo de una aplicación la cual nos ayudará a entender más a estos organismos, y como extra tendremos ayuda también de la Realidad Aumentada.

Keywords: Reino Fungi· Realidad Aumentada· Aprendizaje· Conocimiento

1 Introducción

Cuando hablamos del Reino Fungi, simplemente nos referimos al reino de los hongos. Durante mucho tiempo estos organismos estuvieron englobados en el reino Plantae, pero los científicos se dieron cuenta de algunos aspectos un tanto diferentes con respecto a lo que conocían sobre los seres vivos. Este reino es muy diverso, abarca las conocidas setas, las levaduras y los mohos, entre otros. Algunos hongos son muy grandes pero otros son tan pequeños que sólo pueden verse bajo la lente de un microscopio.Los hongos existen a lo largo y ancho del mundo y en distintos hábitats, apareciendo en distintas formas y presentaciones: la idea tradicional que tenemos de ellos es la seta, con capuchón y cuerpo blanco alargado, pero esa es apenas una de las numerosas especies conocidas dentro del reino fungi.En cuanto a su importancia, destaca la ecológica. La gran mayoría de los hongos se nutren de sustancia orgánica muerta, desempeñando un papel fundamental en la naturaleza al eliminar la hojarasca y demás materia orgánica muerta de la superficie del suelo, que descomponen en materia mineral incorporándola al suelo y manteniendo así la fertilidad del mismo para las plantas verdes, base de la cadena alimenticia del ecosistema.
Como se puede observar en la Fig. 1. una imagen de ejemplo de una muy pequeña porción del Reino Fungi.

Fig. 1. Reino Fungi.

Su importancia en la industria también es muy grande: utilización de levaduras, que se utilizan en muchos productos (bebidas, quesos, etcétera). Está también la industria medicinal. Los hongos son probióticos,lo que significa que ayudan al organismo a combatir las enfermedades, restaurando el bienestar y el equilibrio natural, haciendo que nuestro sistema inmunológico funcione correctamente para eliminar a los agentes externos que podrían desequilibrar nuestra salud.

Por otra parte, está la importancia alimentaria. La recolección de setas silvestres es una afición destacada para algunas personas, y una profesión que da sustento a muchas personas del mundo rural en varias comunidades autónomas.

Luego de lo explicado, podemos decir que los hongos poseen mucha importancia en muchas áreas, tanto para el ser humano como para el medio ambiente en general. Todas estas importancias que posee el Reino Fungi ha llevado a capturar mucho el interes del ser humano por todo el reino, por lo tanto surgen nuevas dudas, y curiosidad por tener mas cerca organismos de este tipo [1][2].

¿Qué mejor manera para acercarnos más a estos organismos que aprender sobre ellos? Gracias a la curiosidad del ser humano, podemos dar inicio a una aplicacion la cuál como objetivo principal tendra el enseñarnos a los usuarios sobre estos organismos tan misteriosos. Al comienzo nos plasmará una guía de cultivo de hongos en casa. Ésta guía nos demostrará cómo podemos cultivar y

tener hongos en nuestra propia casa, también mencionará todos los elementos que necesitaremos, tiempos estimados de crecimiento de hongos y cada una de las etapas de cultivo explicadas paso por paso. Además de ser una guía virtual, también contamos con el beneficio de experimentar y contemplar el desarrollo de un hongo "virtual" el cual cuidaremos en la aplicación y lo ayudaremos a crecer. Y finalmente contaremos con una pequeña experiencia de Realidad Aumentada la cual nos plasmara mediante la cámara de nuestro smartphone el hongo virtual ya desarrollado que tanto cuidamos y ayudamos en todo su desarrollo mediante la aplicación.

La Realidad Aumentada amplía las imágenes de la realidad, a partir de su captura por la cámara de un equipo informático o dispositivo móvil avanzado que añade elementos virtuales para la creación de una realidad mixta a la que se le han sumado datos informáticos[3].

2 Conclusiones

Se puede decir que esta aplicación podría sernos de gran ayuda en un futuno no muy lejano, para experimentos o investigaciones mas simples a las cuales nosotros podamos ser los protagonistas y adémas ya de por si en el presente se le podrían dar muchos usos, como en el area de la educacion, adquisición de conocimientos, etc.

Referencias

1. BioEnciclopedia **,** https://www.bioenciclopedia.com/reino-fungi/, 13 de septiembre del año 2020.
2. Hipertextos del área de la biologia **,** http://www.biologia.edu.ar/fungi/fungi.htm
3. Realidad Aumentada, una evolución de las aplicaciones de los dispositivos mobiles, Javier Fombona Cadavieco, María Ángeles Pascual Sevillano, María Filomena Madeira Ferreira Amador.

Sistema de Detección de Noticias Falsas

Augusto Lombino

Laboratorio de Investigación en Ciencia y Tecnología (LabICyT-UDA),
Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua,
Catamarca 147, Mendoza, Argentina
juanaugustolombino@protonmail.ch

Abstract. El objetivo del siguiente estudio es hacer uso de una red neuronal, alimentando a la misma con un dataset de noticias, tanto falsas como reales, y lograr que la misma sea capaz de categorizar las noticias en falsas o reales.

Keywords: Fake-news · Neural Networks · Redes Neuronales.

1 Introducción

Las redes sociales dominan nuestras vidas digitales, y a pesar de que nunca se cuestionó su impacto, no fue hasta el 2016, con las elecciones presidenciales de EEUU y con el proceso del Brexit en el Reino Unido, que logramos observar el impacto que poseen y la rapidez con la que se difunden las noticias falsas, también llamadas "fake news"[1]. Las investigaciones realizadas indican que se generaron un total de 115 noticias falsas favorables al actual presidente de Estados Unidos que se compartieron en Facebook un total de 30 millones de veces, frente a las 41 fake news en beneficio de Hillary Clinton compartidas un total de 7,6 millones de veces [1]. Según un estudio realizado por Pew Research Center, en el 2016, un 62% de los adultos estadounidenses consumen información a través de las redes sociales. El estudio también califica a Facebook como la red social en la que un 44% de sus usuarios la utilizan para obtener información, seguida por Youtube y Twitter [2]. Esto también está respaldado por un estudio realizado por Digital News Report cuyo grupo de investigación, un 51% mencionó usar las redes sociales como fuente de información, con Facebook nuevamente obteniendo el premio por la red más usada para este propósito. La mayoría de estos artículos, incluso pueden surgir de sitios web cuya ideología política no representa el contenido que producen, tal es el caso de la ciudad de Veles en Macedonia, que se caracteriza por albergar más de 100 sitios web políticos de EEUU, siendo la mayoría pro-Trump [3]. A pesar de esto los adolescentes que se encuentran a cargo, muestran poco interés por Estados Unidos y su situación política. Su objetivo es crear contenido controversial que pueda ser compartido hasta lograr su viralización, y generar ingresos a través de las publicidades que se pueden insertar dentro de las noticias [4].

2 Objetivos

Es necesario, entonces, disponer de herramientas que nos faciliten la detección de este tipo de noticias para poder evitar la desinformacióny su divulgación en las redes sociales. Por esto, a través del uso de redes neuronales, se busca entrenar a una red con un dataset lo suficientemente amplio y con la información necesaria para estar preparado para determinar, con un porcentaje, que tan posible es que una noticia sea falsa. Entre los datos que se busca analizar, se encuentran, la fecha de publicación, ya que como mencionamos previamente, la mayoría de las fake news se producen por asuntos políticos, y en época de elecciones, es más probable encontrarlas. El título y contenido de la noticia, y el sitio web del que proviene.

3 Desafíos

- Los Datasets actuales, están limitados en el sentido de que es difícil encontrar fake news que no tengan alguna motivación política. Últimamente están surgiendo más fake news con respecto a los temas ambientales o científicos, pero no son suficientes. Además, no es fácil encontrar Datasets que no sean en inglés.
- Establecer una arquitectura adecuada para la red neuronal. La misma es escencial para lograr el funcionamiento esperado de la red y poder obtener resultados deseados.
- Establecer claramente qué es lo que diferencia a una fake news, el lenguaje que utilizan, como el manejo de la subjetividad u objetividad, dramatización o sensacionalismo. Se busca saber también, qué tanto de un artículo es falso, es decir si utilizan noticias reales y las mezclan con detalles falsos o si todo el artículo es falso.

4 Herramientas en uso

Actualmente, se está haciendo uso de distintas herramientas para poder comparar y obtener los mejores resultados posibles. El software usado hasta el momento es "Rapidminer", ya que provee una interfáz gráfica para los procesos de minería de datos y permite seleccionar distintos tipos de modelo, ya sea para la detección de anomalías en un conjunto de datos, asociaciones y correlaciones, agrupamiento, regresión o clasificación, que es el modelo que será utilizado en este trabajo. Además se está haciendo uso de Datasets obtenidos de la página web "Kaggle", ya que permite observar si la información presentada en los mismos posee demasiados valores faltantes, son irrelevantes o si se encuentran repetidos, todos detalles importantes que afectarían el procesamiento de los mismos y por lo tanto los resultados finales. Uno de los datasets que se está usando, llamado "Getting Real about Fake News"[5] posee un total de 12999 noticias únicas y cada una posee nombre del autor, fecha de publicación, título, texto del artículo, lenguaje, país de publicación, sitio web de orígen y datos sobre la interacción que

posee la noticia en las redes sociales, como el número de me gusta o cantidad de comentarios. El segundo dataset que se usará, es el llamado "Fake and real news"[6], el mismo no posee tantos metadatos sobre las noticias, cada artículo posee un título, texto, fecha de publicación y la categoría de la noticia, es decir si corresponde a deporte o a un tema político. A pesar de esto, el dataset cuenta con un total de 17903 noticias falsas únicas y 21192 noticias verdades únicas, sumando un total de 39095 artículos, sin contar aquellos que poseen título o cuerpo repetidos.

5 Conclusiones

Hasta el momento sólo se ha realizado una prueba utilizando RapidMiner con los dataset de "Fake and real news", se unió los conjuntos de datos verdaderos, con los falsos y se utilizó un clasificador bayesiano ingenuo. Debido a que este clasificador simplifica el aprendizaje, al tomar cada atributo del dataset como un valor independiente, y toma mucho menor tiempo de procesamiento[7] como resultado se obtuvo una red capaz de clasificar noticias con un margen de 63% de error. Antes de probar otro modelo de Rapidminer, se buscará reproducir el mismo experimento con el otro dataset para observar la diferencia de los resultados. Tras compararlos, se buscará ajustar el algoritmo de clasificación para que esté lo más adaptado posible y finalmente se desearía emplear otro tipo de software, como Weka, Tensorflow o incluso a través de Python, para comparar los resultados con Rapidminer.

Referencias

1. Allcott, Hunt, and Matthew Gentzkow. 2017."Social Media and Fake News in the 2016 Election."Journal of Economic Perspectives, 31 (2): 211-36.
2. Fernandez, Nuria. (2017). Fake News: Una Oportunidad para la Alfabetización Mediática. Nueva Sociedad. 269.
3. Simon Oxenham (2019) "I was a Macedonian fake news writer" –BBC FUTURE.
4. Pangrazio, Luci. (2018). What's new about 'fake news'? Critical digital literacies in an era of fake news, post-truth and clickbait.
5. Meg Risdal "Getting Real about Fake News Text & metadata from fake & biased news sources around the web". https://www.kaggle.com/mrisdal/fake-news
6. Clément Bisaillon "Fake and real news dataset", formado por 17903 noticias únicas. https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset
7. Irina Rish. (January 2001). An Empirical Study of the Naïve Bayes Classifier
8. Newman, Nic & Levy, David & Nielsen, Rasmus. (2015). Reuters Institute Digital News Report 2015. SSRN Electronic Journal.

Aplicación de redes neuronales para la predicción de series temporales de consumo de productos

Di Giorgio Bruno

Laboratorio de Investigación en Ciencia y Tecnología (LabICyT-UDA),
Facultad de Ciencias Sociales y Administrativas, Universidad del Aconcagua,
Catamarca 147, Mendoza, Argentina
digiorgiobruno92@gmail.com

Resumen. El aprendizaje automático (Machine Learning) consiste en dotar a los ordenadores con la capacidad de aprender por sí mismos, sin necesidad de ser programados para esto. El termino se viene usando desde la década de los 50's pero ha tomado especial relevancia en los últimos años. Esto se debe al aumento de la capacidad de computación, así como de los grandes volúmenes de datos que las empresas deben manejar. Los algoritmos de Machine Learning se aprovechan de la llegada del Big Data creando sistemas que aprenden automáticamente. Dicho en otras palabras, sistemas que aprenden mediante ejemplos. La predicción de la demanda de un producto puede ser algo muy aleatorio para muchos negocios, sin embargo, con un buen histórico, podemos ajustarnos más gracias a esa experiencia. En este proyecto se trata de automatizar este proceso, analizando un conjunto de datos a partir de diferentes modelos de Machine Learning para la predicción.

Keywords: Redes neuronales · Series temporales · Predicción de consumo.

1 Introducción

Los avances en los dispositivos de hardware de los últimos tiempos trajeron aparejado un progreso importante en el tratamiento de datos, lo que ha permitido crear sistemas capaces de obtener conocimiento de manera automática [1]. En el presente proyecto se intenta realizar la predicción de la demanda de un producto haciendo uso de dos tipos de algoritmos de aprendizaje automático [1]. Esto con el fin de determinar cual de los algoritmos es el más indicado.

Tener conocimiento sobre la demanda futura es para las empresas e industrias información muy valiosa a la hora de administrar el stock y los recursos. Es posible predecir la demanda los productos analizando datos históricos [1].

El caso problema de este trabajo plantea realizar la predicción de la demanda de un producto específico de un mes a otro, haciendo uso de algoritmos de aprendizaje automático, más específicamente redes neuronales [1]. Los cálculos se realizarán usando dos tipos de redes neuronales, el perceptrón multicapa (MLP)

y redes neuronales recurrentes (RNN) [1], con el fin de contrastar los diferentes resultados y determinar cuál es la mejor opción.

En líneas generales, se analizará un conjunto de datos de series de tiempo que consta de datos de ventas diarias en un mes. Con la información extraída del conjunto de datos se intentará predecir como será la demanda de un producto específico para el mes siguiente. Dicho conjunto se extrajo de Kaggle, una comunidad que permite a los usuarios encontrar y publicar conjuntos de datos, explorar y crear modelos en un entorno de ciencia de datos basado en la web, etc. Cabe mencionar que los datos fueron proporcionados por una de las empresas de software más grandes de Rusia 1C Company [2]. Con este análisis se busca determinar cuál de los dos algoritmos es el más conveniente aplicar.

El método de análisis será comparar los resultados obtenidos de las dos redes, usando una función de error. Las redes neuronales serán entrenadas, usando aprendizaje supervisado, esto significa que los datos que se usarán para entrenar la red tendrán dos atributos, los datos del conjunto de entrada propiamente dicho, y cierta información relativa a la solución del problema. En este caso se dice que hay un profesor externo que observa la salida de la red, y compara la salida obtenida con la salida esperada, y a partir de esto realiza ajustes en los parámetros de la red para poder obtener la salida deseada [1].

Como función de error se ha decidido utilizar el error cuadrático medio o RMSE (Root Mean Squared Error) que es una de las medidas más utilizadas para medir el error entre el valor real y el predicho [1].

2 Redes neuronales aplicadas a la predicción de series temporales

Este apartado veremos una breve descripción redes neuronales y como estas nos pueden ayudar a realizar predicciones en series temporales.

2.1 Perceptrón multicapa

Una de las arquitecturas de redes neuronales más ampliamente utilizada es la que se conoce con el nombre de perceptrón multicapas, la cual se caracteriza por el hecho de que sus neuronas se agrupan en capas. Hay tres tipos de capas diferentes: la capa de entrada que sirven como entrada y propaga la señal proveniente del exterior al interior de la red, las capas ocultas que realizan el procesamiento no lineal de los patrones recibidos y la capa de salida que actúa como salida de la red proporcionando al exterior las respuestas de la red para cada uno de los patrones de entrada [1].

Con respecto a la arquitectura del MLP, podemos saber a priori cuantas entradas y salidas vamos a tener, esto dependerá de las variables que definen el problema, pero el número de capas ocultas y de neuronas en estas se determinará mediante prueba y error.

El MLP tiene un aprendizaje supervisado, lo que significa que se observa la salida de la red y la compara con la salida esperada, y a partir de esto se realizan

ajustes en los pesos de la red usando el algoritmo de backpropagation [3]. Este el proceso de aprendizaje o entrenamiento de la red, es iterativo y en cada iteración se van ingresando diferentes conjuntos de valores a la red de manera tal de poder ajustar la red y prepararla para la predicción. Para el presente trabajo se introducirá el dataset en un MLP y se medirá el error usando diferentes herramientas de medición de error, como cálculo del error cuadrático medio y error medio absoluto, que nos proporcionen una idea de la precisión del pronóstico.

2.2 Redes neuronales recurrentes

La gran diferencia entre los distintos tipos de redes neuronales radica en como es la arquitectura de su red, y esta relacionada a la cantidad de capas ocultas y la cantidad de neuronas en cada capa oculta [6].

Las redes neuronales presentan ventajas en las predicciones debido a su poder adaptativo no lineal [1]. Pero al momento de hablar sobre series de tiempo, las redes neuronales tradicionales como el MLP presentan falencias debido a que no son capaces de capturar la dependencia temporal que puedan tener las observaciones [6]. Es por eso, que nacen las redes neuronales recurrentes (RNN), pues utilizan eventos previos para poder inferir eventos futuros. En cuanto a la arquitectura de las RNN, presentan conexiones recurrentes o bucles que conectan las neuronas [1]. Las RNN tienen estructura con forma de cadena como módulos repetidos con la idea de usar estos módulos para almacenar información importante de los pasos de procesamiento anteriores. Gracias a esto las RNN han tenido éxito en el aprendizaje de secuencias [6]. No obstante, estas redes presentan una gran desventaja la cual es conocida en la literatura como el problema del gradiente de fuga [5]. Debido a lo anterior, se crearon en 1997 las RNN de tipo LSTM gracias a los científicos de la computación S. Hochreiter y J. Schmidhuber [7]. Este tipo de redes poseen bucles de retroalimentación compuestos por células de memoria que poseen un sistema de decisiones basado en compuertas, las cuales que se encargan de "recordar" patrones históricos importantes y "olvidar" patrones no relevantes. Gracias a estas células de memoria, una RNN de tipo LSTM es capaz de asimilar dependencias a corto y largo plazo y, además, atender eficientemente el problema del gradiente de fuga [6].

Al igual que con el MLP, se medirá la eficacia de las RNN tipo LSTM para realizar predicciones en series temporales.

2.3 Predicción de series temporales

El estudio de las series temporales tiene una gran importancia en el campo de la economía dada la abundancia de este tipo de observaciones; de hecho, las series temporales constituyen la mayor parte del material estadístico con el que trabajan los economistas [6]. Por definición, una serie temporal es una sucesión de observaciones de una variable realizadas a intervalos regulares de tiempo [4]. Para el caso de estudio se cuenta con una base de datos que podemos considerar una serie de datos de tipo flujo ya que está referida a un período determinado de tiempo.

El dataset a analizar consta de datos históricos de ventas diarios de varias tiendas en el período de tiempo comprendido entre enero de 2013 a octubre de 2015. La tarea a realizar en este trabajo es pronosticar la cantidad total de vendidos de un producto seleccionado. De esta forma acotamos el espacio de nuestro problema.

La metodología a utilizar en el presente trabajo esta basada en la utilizada en el trabajo de J. Colin Pérez [1] porque se observó que dio buenos resultados. Primero se realizará un procesamiento de datos para que el proceso de predicción sea más efectivo. Este consta limpieza de datos y eliminación de características redundantes o poco relevantes. Luego se aplicarán los dos modelos mencionados anteriormente, este proceso implica afinar cada modelo, ya que como se mencionó anteriormente estos procesos muchas veces implican prueba y error. Por últimos se recolectarán los datos y se sacarán conclusiones

3 Conclusiones

En el presente trabajo se parte de la hipótesis de que las redes neuronales recurrentes presentan mejores resultados que las redes neuronales convencionales, es por esto que se decidió realizar una comparativa para demostrar esto de forma empírica. Debido a que el presente trabajo está en fase de desarrollo no se pueden dar grandes conclusiones hasta el momento, pero se puede comentar que el enfoque planteado es prometedor ya que se ha visto que ofrece buenos resultados en trabajos similares.

Referencias

1. P. Isasi: Redes de neuronas artificiales: un enfoque práctico. Libro Pearson Educación, 2004
2. 1c Homepage, http://www.1c.com/. 28/09/2020
3. J.Colin Pérez, F.: Aplicación de machine learning para la predicción temporal en un caso de marketing. Trabajo de fin de grado (2018)
4. F. Parra: Estadística y Machine Learning con R. Manual, 2019
5. J. A. Puentes: Ethanol Fuel Demand Forecasting in Brazil Using a LSTM Recurrent Neural Network Approach. Artículo(2020)
6. S. E. Altamirano: Aplicación de redes neuronales recurrentes y modelos de series de tiempo bayesianos a la predicción de rentabilidad de fondos de pensiones. Tesís(2020)
7. S. Hochreiter: Long Short-Term Memory, Neural Computation,vol. 9, nº 8, pp. 1735-1780, (1997)

Lista de autores

——/ A /——

Aballay, Alicia 59
Aballay, Laura Nidia 155
Abraham, Jurado 283
Aceval, Cristian 83
Aguilera, Cynthia 245
Andrade Ramirez, Jaime 123
Atencio, Hernán 245

——/ B /——

Becerra, María del Carmen 59
Bittencourt, Luiz Fernando 71

——/ C /——

Caffaratti, Gabriel 209
Campos, Matías 307
Cappo, Cristian 83
Castrillón Cobo, Helder 107
Catania, Carlos 209
Catania, Carlos A. 297
Collado, Facundo 263

——/ D /——

Di Giorgio , Bruno 325

——/ E /——

Eguren, Santiago 297
Erazo, Leydi 107
Espinosa García, Alexander 123

——/ F /——

Farías, Andrés Francisco 13
Fröhlich, William da Rosa 47

——/ G /——

Gallego, Erika 107
García, Pablo Marcelo 13
Georgantas, Nikolaos 71
Giménez, Rosana 143, 221, 257, 263, 271
Gomez Blotta, Emiliano 257
Gomez, Martín 245
Gonzalez, Rodrigo 297, 307
Guerra, Jorge 209
Guevara-Pérez, S.V. 233
Gómez-Mendoza, J.B. 233

——/ H /——

Herrera, Myriam Beatriz 95

——/ J /——

Jerez, Marcela 293

——/ L /——

Lombino, Augusto 321
Lopez-Fenner, Julio 71
Lund, Maria Inés 155
Lund, María Inés 35

——/ M /——

Marchetta, Martin 209
Maschke, Julián 313
Medel, Diego 245
Medina Albarracín, Cristian 123
Migani, Silvina 35, 181
Molina, Ana Laura 169
Moncho, Matías 289, 293
Montejano, Germán Antonio 13
Moreira Costa, Fabio 71
Moyano, Ezequiel 283, 289, 293
Murazzo, María 169, 245
Méndez-Garabetti, Miguel 143, 221, 257, 263, 271

——/ O /——

Olivera, Ana C. 1
Olivera, Ana Carolina 25
Orozco Gazzo, Emmanuel Ricardo 301
Ortiz, Tomás 143

——/ P /——

Paez Chanaga, Anderson 131
Parra, Lorena 169
Pessina, Giuliana 317
Pinciroli, Fernando 197
Puebla, Adrián 221

——/ R /——

Ramirez Silva, Ana 131
Rigo, Sandro José 47
Riveros, Alejandro 35
Rocha, Saturnino 271
Rodrigue, Clarissa Almeida 47
Rodríguez, Nelson 169, 245
Rojas, Matias G. 1
Romagnano, María 59, 95
Romano, Lucas 293
Romero, Juan Manuel 209
Rosenstein, Javier J 307
Ruiz, María Eugenia 47
Ruiz, Susana B. 35

——/ S /——

Sanchez Galvis, Ivan 131
Sanchez, Federico 169
Sarmiento, Adriana 59
Scheffer, María del Valle 95
Sepúlveda, Samuel 71
Sández, Pablo 143, 221, 257, 263, 271

——/ T /——

Tamayo-Quintero, J.D. 233
Tinoco Barreto, Andrés 123

——/ V /——

Vargas, Lucas 245
Vera, Cristina 181
Vidal, Pablo J. 1
Vidal, Pablo Javier 25
Villarreal, Martín 283, 293

——/ Z /——

Zárate Alvarez, Nicolás 307

UNIVERSIDAD
CHAMPAGNAT

fcefn
Facultad de Ciencias Exactas, Físicas y Naturales
Universidad Nacional de San Juan

Universidad
CATÓLICA
de Pereira
VIGILADO MINEDUCACIÓN