NÚMERO ESPECIAL

REVISTA POLITÉCNICA

ESCUELA POLITÉCNICA NACIONAL

ISSN: 1390-0129
eISSN: 2477-8990
Volumen 50, Nro. 3, Diciembre 2022

REVISTA POLITÉCNICA

ESCUELA POLITÉCNICA NACIONAL

ISSN: 1390-0129
eISSN: 2477-8990
Volumen 50, Nro. 3, Diciembre 2022

TEMÁTICA Y ALCANCE

La Revista Politécnica es una publicación periódica trimestral editada por la Escuela Politécnica Nacional del Ecuador, creada en el año 1961, siendo la primera revista científica ecuatoriana, cuyo objetivo es contribuir al conocimiento científico y tecnológico, mediante la publicación de estudios científicos relacionados con las áreas de ciencias básicas (Física, Química, Biología y Matemática) e ingenierías (Química y Agroindustria, Civil y Ambiental, Eléctrica y Electrónica, Geología y Petróleos, Mecánica, y Sistemas). La Revista Politécnica está dirigida a profesionales e investigadores que trabajan en estos campos del conocimiento.

La Revista Politécnica está incluida en SCOPUS, Scientific Electronic Library Online (SciELO), Red de Revistas Científicas de América Latina y el Caribe, España y Portugal (Redalyc), Catálogo 2.0 de Latindex, Directory of Open Access Journals (DOAJ), Red Iberoamericana de Innovación y Conocimiento Científico (REDIB), Matriz de Información para el Análisis de Revistas (MIAR), Bielefeld Academic Search Engine (BASE), CiteFactor, Google Scholar, ResearchBib e ICI Journals Master List 2020.

Coordinador Técnico Operativo

Ing. Ricardo Villarroel
ricardo.villarroel@epn.edu.ec

MSc. Karina Játiva
karina.jativa@epn.edu.ec

Proofreader

MSc. María Eufemia Torres

Diseño de Portada

Cristian Basurto
Diseñador Gráfico

Editorial

Congreso internacional de investigación aplicada a Ciencia de Datos y II Congreso Nacional de R Users Group-Ecuador

El Big Data y la Ciencia de Datos son actualmente las áreas que han motivado, no solo el desarrollo y la aplicación de nuevos modelos matemáticos y estadísticos, sino también, la forma en que se hacen los negocios y se toman las decisiones. Al hablar de estas áreas, se debería considerar la relación inherente y complementaria que existe entre ellas. Si bien es cierto el Big Data proporciona herramientas y técnicas para poder administrar y procesar grandes cantidades de datos, no se enfoca en la interpretación y análisis de los resultados. Por otro lado, la Ciencia de Datos a través del uso de las técnicas avanzadas en estadísticas y matemáticas, proporciona resultados que generan perspectivas de negocio para la toma de decisiones.

En vista de los nuevos entornos donde las empresas deben tomar decisiones, la Facultad de Ciencias de la Escuela Politécnica Nacional, con la colaboración del Colegio de Científicos de Datos del Ecuador (CCDE), R-Users Group-Ecuador y la Sociedad Ecuatoriana de Estadística (SEE), organizó el Congreso de Investigación Aplicada a Ciencia de Datos y II Congreso Nacional de R Users Group, cuyo objetivo fue generar espacios para la interacción entre estudiantes, profesores, profesionales, instituciones públicas y privadas, para dar a conocer los principales aspectos metodológicos y prácticos de los métodos de matemática aplicada utilizados en los diferentes campos de las Ciencias de Datos.

El Congreso se llevó a cabo de forma virtual, del 24 al 28 de enero de 2022. Se trataron siete temáticas de aplicación, estas fueron: Economía y Finanzas, Ciencias Sociales, Educación, Ciberseguridad, Medio Ambiente, Industria; y Aplicación del software R en la Academia e Industria. Se contó con la participación de siete ponentes internacionales de diferentes países como España, Colombia, Chile y Argentina, así como dos ponentes nacionales. Adicionalmente, se presentaron dos talleres prácticos a través de metodologías implementadas con el lenguaje de programación estadístico R.

El congreso tuvo una buena acogida entre la comunidad académica y profesional a nivel regional. Se contó con la participación de un promedio diario de 100 personas en tiempo real (transmisión en redes sociales y en la sesión de zoom). Mientras que, en las vistas de las grabaciones se alcanzaron más de 5000 interacciones en la semana del evento.

Se contó con más de 30 ponencias y 15 pósters durante el congreso, los cuales fueron seleccionados por el comité científico a través de la evaluación de los resúmenes de los trabajos enviados. De estos trabajos se seleccionaron cinco con la finalidad de ser sometidos a revisión por pares y poder ser considerados para la edición especial de la Revista Politécnica de la Escuela Politécnica Nacional. Los trabajos sometidos, se presentan como artículos de alto nivel académico y con una aplicación que abarcan las temáticas del congreso.

El Comité Editorial de la Revista Politécnica desea agradecer una vez más a los autores que han presentado sus artículos y a los revisores que han velado por la calidad científica de estos documentos. Esperamos que la comunidad científica y la Escuela Politécnica Nacional disfrute de este volumen especial desarrollado a través del congreso.

Editorial

International Congress of applied research to Data Science and II National Congress of R Users Group-Ecuador

Big Data and Data Science are currently the areas that have motivated, not only the development and application of new mathematical and statistical models, but also the way business is done and decisions are made. When talking about these areas, the inherent and complementary relationship between them should be considered. While it is true that Big Data provides tools and techniques to be able to manage and process large amounts of data, it does not focus on the interpretation and analysis of the results. On the other hand, Data Science through the use of advanced techniques in statistics and mathematics, provides results that generate business perspectives for decision making.

Given the new environments where companies must make decisions, the Faculty of Sciences of the National Polytechnic School, with the collaboration of the College of Data Scientists of Ecuador (CCDE), R-Users Group-Ecuador and the Ecuadorian Society of Statistics (SEE), organized the Congress of Research Applied to Data Science and II National Congress of R Users Group, whose objective was to generate spaces for interaction between students, teachers, professionals, public and private institutions, to publicize the main methodological and practical aspects of the applied mathematics methods used in the different fields of Data Science.

The Congress was held virtually, from January 24 to 28, 2022. Seven application topics were discussed, these were: Economics and Finance, Social Sciences, Education, Cybersecurity, Environment; Industry and Application of R software in Academia and Industry. It was attended by seven international speakers from different countries such as Spain, Colombia, Chile and Argentina, as well as two national speakers. Additionally, two practical workshops were presented through methodologies implemented with the statistical programming language R.

The congress was well received by the academic and professional community at the regional level. A daily average of 100 people participated in real time (broadcast on social networks and in the zoom session). While in the hearings of the recordings more than 5000 interactions were reached in the week of the congress.

There were more than 30 papers and 15 posters presented during the event, which were selected by the scientific committee through the evaluation of the abstracts of the works submitted. Of these works, five were selected in order to be submitted to peer review in order to be considered for the specialedition of the Revista Politecnica of the National Polytechnic School. The submitted works are presented as articles of high academic level and with an application that covers the topics of the congress.

The Editorial Committee of the Revista Politecnica wishes to thank once again the authors who have presented their articles and the reviewers who have ensured the scientific quality of these documents. We hope that the scientific and polytechnic community will enjoy this special volume developed through the congress.

Contenido
Vol. 50, No. 3
Diciembre 2022

Automatic Detection of Fake News in Spanish: Ecuadorian Political Satire

Mafla, Nicolás [1,*]; **Flores, Miguel** [2]; **Castillo-Páez, Sergio** [3]; **Andrade, Roberto** [4]

[1]*Escuela Politécnica Nacional, Facultad de Ciencias, Quito, Ecuador*
[2]*Escuela Politécnica Nacional, Facultad de Ciencias, Departamento de Matemática, Grupo MODES, SIGTI, Quito, Ecuador*
[3]*Universidad de las Fuerzas Armadas ESPE, Departamento de Ciencias Exactas, Ecuador*
[4]*Escuela Politécnica Nacional, Facultad de Ingeniería en Sistemas, Quito, Ecuador*

Abstract: The circulation of fake news on internet, especially those of political satire through social media, has affected the majority of the Ecuadorian population. This work presents a methodology based on statistical learning that accurately and automatically detects fake news in Spanish using machine learning and natural language processing techniques. The document begins by presenting basic concepts related to fake news and works related to their automatic detection. The second section explains the news corpus creation process, text processing, numerical representation with TF-IDF and training of supervised classification algorithms with two different data sets. Results obtained from the training are analyzed in the third section, being the models with support vector machines the ones that offer the best predictions, improving approximately 15%, 6% and 3% to the performance of the models with naive bayes, random forests and boosting trees respectively. Finally, conclusions of the research and future work is presented in the fourth section.

Keywords: Fact-checking, machine learning, natural language processing, supervised classification

Detección Automática de Noticias Falsas en Español: Sátira Política Ecuatoriana

Resumen: La circulación de noticias falsas en internet, especialmente las de sátira política a través de redes sociales, ha afectado a la mayoría de la población ecuatoriana. Este trabajo presenta una metodología basada en el aprendizaje estadístico que detecta de forma precisa y automática noticias falsas en español utilizando técnicas de aprendizaje automático y procesamiento del lenguaje natural. El documento comienza presentando conceptos básicos relacionados con las noticias falsas y trabajos relacionados con su detección automática. La segunda sección explica el proceso de creación del corpus de noticias, procesamiento de los textos, representación numérica con TF-IDF y entrenamiento de algoritmos de clasificación supervisados con dos conjuntos de datos diferentes. Los resultados obtenidos del entrenamiento se analizan en la tercera sección, siendo los modelos con máquinas de soporte vectorial los que ofrecen mejores predicciones, mejorando aproximadamente un 15%, 6% y 3% al rendimiento de los modelos con naive bayes, random forests y árboles boosting respectivamente. Finalmente, las conclusiones de la investigación y el trabajo futuro se presentan en la cuarta sección.

Palabras claves: Fact-checking, machine learning, procesamiento del lenguaje natural, clasificación supervisada

1. INTRODUCTION

The time we currently spend browsing the internet and consuming content on social media occupies a large part of our day (Bergström and Jervelycke Belfrage, 2018), without a doubt they have become one of the most powerful communication tools used today since it allows access to consumption and disclosure information instantly available to anyone. This has led to the preference of digital media over traditional media, consequence of this free access and the easy generation of content, the information that circulates online is not always reliable (López-Buroll et al., 2018). The growth of social networks increases the spread of fake news on the internet, information distributed by these media is massive, fast and heterogeneous, which can cause serious impact on the entire society (Zhang and Ghorbani, 2020).

Fake news has had a negative impact on several sectors, among these, the most important communication, economy, health and politics; mentioning some cases we have: the 2016 US presidential elections (Allcott y Gentzkow, 2017), the threat to global public health caused by the massive infodemic originated around the pandemic produced by the covid-19 virus (Pulido et al., 2020) and a curious case is of a scientific publication in the American Journal of Biomedical Science & Research

*nicolas.mafla.checa@gmail.com
Recibido: 11/10/2021
Aceptado: 01/05/2022
Publicado: 23/12/2022
10.33333/rp.vol50n3.01

(Shelomi, 2020), noting that even the scientific community is not exempt from sharing false information that pretends to be true.

The problem of fake news that circulates on the internet and also on social networks affects us all directly or indirectly, it is for this reason that this research provides the theoretical bases for the creation of applications that help to fight against media misinformation, as mentioned above, many sectors are affected. Additionally, the development that is achieved will potentially serve for future research on this topic and will be an advance with respect to supervised classification models that involve natural language processing techniques in the Spanish language.

The objective of this research is to create a model that allows to accurately and automatically identify fake news in Spanish from Ecuadorian news and satire pages. For this, a text corpus was created extracting the news manually or using web-scraping techniques, which were later processed with NLP techniques to create the database that served to train the algorithms and compare their results.

1.1 Fake news

The term fake news is recent and has gained popularity in last years, but false information has been part of humanity for a long time, fake news is as old as the printed news that circulated since the invention of the printing press, or even older (Soll, 2016). There is no universal definition for fake news and it is not easy to formulate a generally accepted one for the term, since these tend to be diverse in terms of topics, styles and even platforms, which is why several authors have proposed their own definitions. Fake news is defined as manufactured information that imitates the content of the media in form but not in the organizational process or intention, in addition they lack the standards and editorial processes of the media that ensure the accuracy and credibility of the information (Lazer et al., 2018). They refer to all kinds of false stories that are published and distributed mainly on the internet, in order to deceive or deliberately entice readers for financial, political or other benefit (Zhang and Ghorbani, 2020). And lastly, fake news is news articles that are intentionally created and verifiable false that could mislead readers (Allcott y Gentzkow, 2017). Based on how the concepts presented were created, they share three characteristics in common, the first is related to the authenticity of the news information (containing some statement based on facts or not), the second refers to the intention to create fake news (with the aim of deceiving or entertaining the public), and finally, if they are really news (Zhou and Zafarani, 2020).

1.2 Knowledge-based detection

According to Zhang and Ghorbani (2020), and Zhou and Zafarani (2020), a process known as fact-checking is generally used to detect fake news from a knowledge approach, this method was initially developed by journalists, and is currently used by a large part of the media. Fact-checking process aims to verify the authenticity of the information, comparing the knowledge extracted from the content of the news to be verified with known facts. Both evaluation criteria and visual metrics are used to determine the level of veracity of the news in the fact-checking process.

Manual fact-checking is usually carried out by a select group of professionals called fact checkers of great credibility since this leads to highly accurate results. This process can require a lot of execution time with a high maintenance cost, it also presents difficulties when the information content to be verified is massive. Another way of conducting fact checking is data verification through collective sources, which is based on a large population of regular individuals acting as fact-checkers.

Compared to expert-based fact checking, it is relatively difficult to administer, less credible and accurate due to the political bias of its verifiers, but having better scalability. Manual fact-checking does not adapt to the volume of information that is created every day online, especially on social media. For a better scalability of fake news detection, automatic fact-checking techniques have been developed, which are largely based on extraction of information through natural language processing and classification using machine learning techniques (Bondielli and Marcelloni, 2019).

Given the diversity of ways of speaking Spanish, which largely depends on the geographical area where the speaker comes from, this research focused on creating a model capable of identifying fake news in Spanish from Ecuador. The model was trained from a set of publications identified as real and fake news, extracted from the main national newspapers and pages of political satire on Facebook, in addition to comparing different types of classification algorithms used in the detection of fake news in the English language (Bondielli and Marcelloni, 2019).

1.3 Related work

Singharia et al. (2017), present a three-level hierarchical attention artificial neural network (3HAN): words, sentences and headings for accurate detection of fake news. The model is based on the representation of a news article as a vector, which is used to classify an article by assigning a probability of being false. The data used for the training of the model belongs to the period of the USA presidential elections of 2016 and was extracted from the PolitiFact platform to create a set of fake news and from the list of popular verified sites in the USA provided by Forbes to create the set of real news. Ciampaglia et al. (2015), present a model based on networks or graphs, in which the verification of facts can be approximated quite well by finding the shortest path between nodes denoting statements under properly defined semantic proximity metrics on knowledge graphs. The data used was extracted from Wikipedia, which includes all the factual statements extracted from the Wikipedia information boxes, thus creating a knowledge graph with 3 million entities linked by approximately 23 million edges. Posadas-Durán et al. (2019), present a model to analyze and detect misleading information present in a large number of Spanish-language websites. For the training of the model, a set of news collected manually from different websites was used to create a corpus of news labeled as fake and real news. The training was carried out using supervised classification algorithms: vector support machines, logistic regression, random forests and gradient boosting, evaluating performance by removing stop words and

considering stop words. Extraction of information and its representation was carried out through linguistic characteristics obtained from three techniques: bag-of-words, n-grams and POS tags n-grams.

2. METHODOLOGY

In the absence of a set of news data classified as true and false, the first step was to create it from Ecuadorian information sources available on the internet and social media. News texts were cleaned and represented numerically with the TF-IDF technique to create the detection models. In order to compare the ability to detect fake news, four algorithms were compared: vector support machines, random forest, boosting trees and the naive bayes classifier. The entire process from extraction to modeling was carried out in Python, specifically for the training of the classification algorithms scikit-learn was used (Pedregosa et al., 2011). Methodology for creating the models from text processing to algorithm training was based on followed by Posadas-Durán et al. (2019).

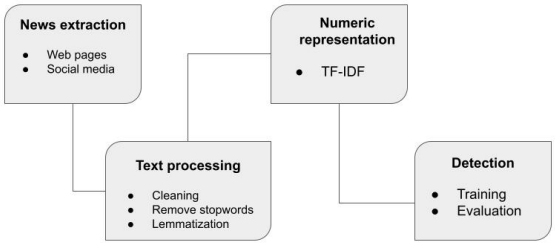

Figure 1. Graphic diagram of the problem modeling process

2.1 News extraction

The corpus consists of a collection of news from the Ecuadorian sphere extracted from different pages within the internet: websites of the main real national newspapers and pages of political satire on social networks, collected from September to December 2020 what is the period in which the information gathering process was carried out.

The publications made by the political satire pages on Facebook used in this work are characterized by being short messages, so the information obtained from the newspaper pages was not the complete story, it was the summary of the news provided by the same newspapers in order to maintain the pattern that the pages of satire follow. To label the collected news, the following was considered:

- The news was labeled as real if there was evidence that the publication came from a trusted page, such as it is for newspaper web pages.
- The news was labeled as false if there was no evidence about the veracity of the information presented and the credibility of the page from which this news came, in this case the pages of political satire.

2.2 Text processing

The components of interest in this work are the words of each news item, which is why elements that cannot be considered words were extracted from the corpus. The elements removed from the texts were:

- Links referring to external pages
- Tags to users´ accounts
- Hashtags
- Punctuation marks
- Numeric and special characters

Finally, all the words frequently used in languages that are not helpful by themselves in understanding the context of the news and in general of a text called stopwords were removed and all document in the corpus were transformed to lowercase.

Lemmatization is the process in which given all the different inflected forms of a word, its base form is found, this helps us to considerably reduce the number of words with similar meanings in a text. This task was carried out with the help of the text processing library spaCy, in it there is a great variety of pre-trained models based on neural networks in different languages including Spanish.

2.3 Numerical representation

To numerically represent the news texts, the TF-IDF technique was used, which is based on representing each of the texts as a vector. The term frequency (TF) measures the frequency with which a term appears in a given document, while the inverse document frequency (IDF) measures the importance of a term within the corpus, weighing less weight to the terms that are very common within the corpus, while it weighs more unusual. Multiplying both metrics gives the TF-IDF representation (Vajjala et al., 2020).

$$TF\text{-}IDF(p_i,d_j) = \frac{f(p_i,d_j)}{|d_j|} \cdot log\left(\frac{M}{|\{d \in D : p_i \in d\}|}\right) \quad (1)$$

For a corpus $D = \{d_1,...,d_M\}$ and a vocabulary $V = \{p_1,...,p_N\}$, where $f(p_i,d_j)$ is the relative frequency of the word p_i of the document d_j in the corpus D.

2.4 Fake news detection

The problem consists of predicting if a news item is fake based on the information provided by it, captured in the form of vectors with the TF-IDF technique. The response variable that we want to predict is defined as a binary variable, this will take the value of 1 if it is a fake news and 0 if it is a real news. For the creation of the models, the usual process of training of supervised classification algorithms was followed, partitioning the data set in training,

validation and tests to obtain the optimal hyperparameters and the algorithms that presented the best performance to new data sets.

In several investigations, it has been shown that classical supervised classification methods solve the problem of detecting fake news in the English language with excellent results, the ones that stand out the most are: SVM, random forests and boosting algorithms (Zhou and Zafarani, 2020). On the other hand, in the research carried out by Posadas-Durán et al. (2019) they solve the detection problem in the Spanish language with the same classifiers and additionally logistic regression obtaining good results. For these reasons, the classifiers mentioned above were chosen, except for regression, which was replaced by the naive bayes classifier since it has less demanding assumptions.

SVM

The support vector machine is a generalization of a linear classifier called the *maximal margin classifier*, that has the objective of solving binary classification problems in which the data set has two classes but these are not separable by a linear boundary. To achieve this, the space of the training data is transformed into one of higher dimension in which a hyperplane can be fitted that maximizes the separation of the two classes. The problem solved is the following:

$$\begin{aligned} &\max_{\beta_0,\beta} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.a} \quad &y_i(\beta_0 + \beta\cdot\phi(x_i)) \geq 1-\xi_i, \forall i \in \{1,...,n\} \\ &\xi_i \geq 0, \quad C>0 \end{aligned} \tag{2}$$

The hyperplane that solves the classification problem is represented by $\beta_0 + \beta\cdot\phi(x_i)$ where ϕ is the function with which the space of the training data is transformed. $\xi_1,...,\xi_n$ are variables that allow the problem to be more flexible and allow a given number of observations to be on wrong sides of the hyperplane controlled by the parameter C.

The classification of an observation x is given by the expression $sign\{\beta_0 + \sum_{i=1}^{n}\alpha_i y_i \phi(x_i)\cdot\phi(x)\}$ and here is the importance of the kernels, since it is not necessary to have an explicit representation of the function ϕ, it is enough to define the inner product of the transformed data, so the kernel functions are represented as follows:

$$K(x,x') = \phi(x)\cdot\phi(x') \tag{3}$$

There are several kernel functions that are used to solve classification problems with SVM, the most used and the ones we chose for this investigation are: linear, sigmoid and gaussian.

Random Forest

The Random Forest algorithm involves a sequence of models that segment the predictor space into simple regions and its classification rules can be modeled by a series of related decisions known as *decision trees*. The most common way to measure the power of division in trained classification trees is the classification error rate, this measures the proportion of observations that do not belong to the most common class, in practice more sensitive measures are used, such as gini index or *cross-entropy* coefficient, since these are a way of measuring *impurity* of the segmentation. The algorithm is shown below:

Algorithm 1: Random Forest for classification

For $m = 1$ **to** M **:**

(a) Take a bootstrap sample Z of size N from the training space

(b) Fit a tree T_m from the sample Z in which each division of the tree is carried out with the following steps

1. Select m variables from the total of variables
2. Take the best variable of the selected m to split the tree based on the metric used (gini or cross-entropy)
3. Segment the space

Output: Collection of trees $\{T_m\}_1^M$

The classification for an observation x will be the class to which x belongs that is repeated the most in the predictions of the trees $\{T_m\}_1^M$

Boosting Trees

Similar to random forests, boosting models are based on the use of a series of weak that together provide a great predictive power and are widely used for regression and classification problems. Unlike random forests, decision trees are not trained on bootstrap samples, instead they are trained on sequential modifications of the training space, each decision that is fitted uses the information from its predecessors.

The objective is to repeatedly apply a weak classifier to modifications of the training set, generating M new weak classifiers and these will be used for the final prediction. The classification of an observation x will be the weighted combination of the predictions of these weak classifiers.

$$G(x) = sign\left\{\sum_{m=1}^{M}\alpha_m G_m(x)\right\} \tag{4}$$

where α_m controls the contribution that the decision tree G_m makes to the final model, giving more relevance to those that are more accurate. The modifications made to the training set consist of applying weights $w_1,...,w_N$ to the N observations $x_1,...,x_N$, the weights are initialized with $w_i = \frac{1}{N} \quad \forall i = 1,..,N$ and as the iterations go by, the weights are modified one by one, making the algorithm classification uses the new weighed data.

In each iteration, the observations that were erroneously classified increase their weight, while those correctly classified have their weight reduced. Thus, as the process progresses, the observations with problems being classified acquire more relevance for the following iterations, forcing to the following classifiers to concentrate on those that did not obtain an accurate prediction by the previous classifiers.

Naive Bayes Classifier

The naive bayes classifier is based on Bayes theorem, which is used to estimate the conditional probability of the occurrence of an event given a certain amount of information about the event. For a training space with p predictors and response variable Y can take K two classes, using Bayes the prediction that the observation x belongs to the class K is the following:

$$p_k(x) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)} = \frac{\pi_k \prod_{j=1}^{p} f_{jk}(x_j)}{\sum_{i=1}^{K} \pi_i f_i(x)} \quad (5)$$

where $\pi_k = P(Y = k)$ is the prior probability that an observation taken at random belongs to class k, $f_k(X) = P(X|Y = k)$ denotes the density function of X for observations from classes k. This model assumes that given a class k the marginal distributions of the predictors $X_1, ..., X_p$ are independent and normal distributed.

2.5 Model evaluation

The evaluation of the models was measured by the F-score, area under the receiver operating characteristic curve (ROC) and the estimation of the general prediction error. By having an imbalance in the classes of the data, that is, there is much more true news than fake news, just measuring the accuracy of the model could lead to biased results (Zaki and Meira, 2014). For this reason, the F-score measure was chosen since it combines both the precision and the recall of the model, thus having a more real measure of the prediction power.

$$F_P = \frac{2TP}{2TP + FP + FN} \quad F_N = \frac{2TN}{2TN + FN + FP}$$
$$F = \frac{F_P + F_N}{2} \quad (6)$$

Where TP, TN, FP and FN are the true positives, true negatives, false positives and false negatives respectively of the model in the test set.

Algorithm 2: General test error estimation (K-fold cross validation)

Partition Corpus D in $[D_1, ..., D_K]$ equal parts
for $i = 1$ **to** K **do**
  $G_i \leftarrow$ train the classifier on $D \setminus D_i$
  $\hat{Err}_i \leftarrow$ calculate prediction error of the classifier G_i on $D \setminus D_i$
end
$\hat{Err} = \frac{1}{K} \sum_{i=1}^{K} \hat{Err}_i$
Output: $\hat{Err}$

3. RESULTS

Collecting the text of the pages on Facebook was done manually by copying the content, on the other hand, the texts collected from the pages of the newspapers was done with the help of web scrapping. Pages of national newspapers chosen were: "El Comercio" and "El Universo" for the real news group and political satire pages were "El Uninverso", "El Culimercio" and "El Mercioco" for the fake news group.

Table 1. National newspaper pages and political satire pages

Page	Website	Type of news
El Universo	www.eluniverso.com	National newspaper
El Comercio	www.elcomercio.com	National newspaper
El Uninverso	www.facebook.com/DiarioUninverso	Political satire
El Culimercio	www.facebook.com/elculimercioec	Political satire
El Mercioco	www.facebook.com/mercioco	Political satire

The news corpus consists of a total of 1629 news items, approximately 59% of these belong to "El Comercio", since due to the structure of its website, it facilitated the process of collecting the news through webscrapping, so it was possible to extract large amount of news with great speed. In general, the pages that create false information tend to have a short lifetime (Allcott y Gentzkow, 2017), so the information obtained from two of the satire pages is much less than that obtained from the pages of national newspapers, except from the "El Uninverso" page, which is a fairly active page that has published a lot of content since its creation; the advantage of this was that the time invested for the extraction of the false texts was not excessive and it was feasible to execute it manually.

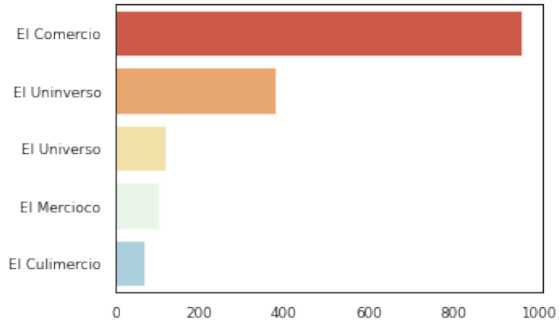

Figure 2. Distribution of the news collected according to the page where they come from

Of the total news we have that 34% corresponds to fake news and the remaining 66% to real news. Analyzing the real news and the fake news from a point of view at the level of terms, we can see that the distribution of the number of terms is quite similar for both groups, the average of terms for the real news is approximately 41 terms while for the fake news is about 42. On the other hand, the number of terms present in the fake news are a little more variable than the real ones, having 15.7 and 19.6 the approximate deviations respectively. The resulting news set was stored in a comma separated values (csv) file with utf-8 encoding to be able to store words in Spanish without losing the accents and the letter ñ.

Vocabulary consists of 9953 words, as this technique depends on the total number of words in a corpus, if the resulting vocabulary consists of an excessive number of words, this quantity will be the number of variables when representing the texts numerically, so remove stop words and lemmatization helped us reduce the number of variables without losing the structure of the text. The original database created has 1629 records with 9953 variables, but applying the mentioned reduction techniques, the number of

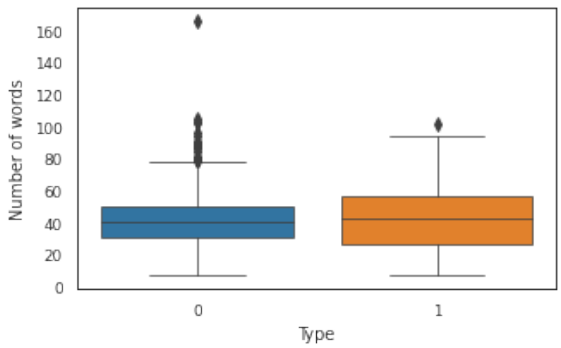

Figure 3. Distribution of the number of terms by news groups

variables decreases to 6336. It was analyzed whether reducing the number of variables with these techniques affects the performance of the classification models, so these two data sets were used to train the algorithms.

Since there is only 34% fake news, the train database was balanced in order not to obtain biased results. To balance the data sets, a technique called SMOTE (Bowyer et al., 2011) was used, which unlike the traditional ones based on resampling with replacement, this creates synthetic observations of the minority class using its K closest neighbors. The minority class is oversampled by taking each of the observations and introducing synthetic observations along the segments that join them with one or all of the K closest neighbors, that is, the difference between an observation and its closest neighbor is taken and this difference is multiplied by a random number between 0 and 1, thereby selecting a random point along the length of the two observations.

F-score values for all the models in general are above 80% which is very good since this indicates that the models make correct predictions most of the time. In the same way, the area under the ROC curve for all the models in general is excellent, having values very close to the best of the cases, this indicates that the models can easily differentiate fake news from real news. The results are shown in Table 2 and models trained from unprocessed texts are represented by $(*)$.

Reducing the number of variables in the data set when processing the news texts could reduce the predictive power of the models, which is true, but it is clear that in this case the reduction is minimal, since the values F-score and area under ROC curve for models without stop words and lemmatization are practically equal to models with stop words and non-lemmatization.

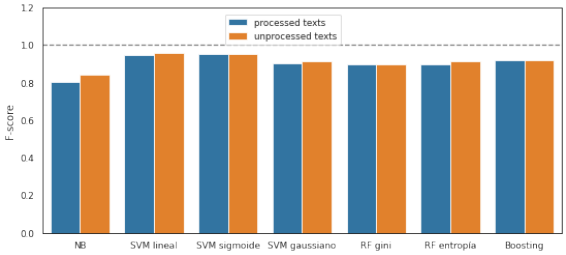

Figure 4. Results of the F-score measure for processed texts and unprocessed texts

The estimate of the expected prediction error was obtained through the cross-validation process with $K = 5$ and $K = 10$ partition groups of the data set. With $K = 5$ of the 1629 news of the corpus, 5 groups of approximately 326 news each were randomly taken, with which the 5 models were trained with a set of approximately 1304 and the predictions were made on approximately 326 news. In a similar way, the estimates were made with $K = 10$, thus, from the 1629 news of the corpus, 10 groups of approximately 163 news each were randomly taken, with which the 10 models were trained with a set of approximately 1630 and made the predictions on approximately 163 news items.

Table 2. National newspaper pages and political satire pages

model	Error	AUC	F-score
SVM linear	4.91%	99,24%	94.41%
SVM sigmoid	4.29%	99,2%	95.13%
SVM gaussian	7.98%	99,31%	90.35%
SVM lineal*	3.68%	99.40%	95.81%
SVM sigmoid*	4.29%	99,30%	95.08%
SVM gaussian*	7.36%	99.20%	91.35%
NB gaussian	17.79%	81.09%	80.36%
NB gaussian*	13.5%	83.37%	84.31%
RF gini	8.59%	97.47%	89.67%
RF entropy	8.59%	97.63%	89.91%
RF gini*	8.59%	98.30%	89.91%
RF entropy*	7.36%	97.93%	91.44%
Boosting trees	6.75%	97.27%	92.11%
Boosting trees*	7.36%	97.60%	91.61%

Based on the results obtained, the models are able to accurately predict whether a news item is real or fake, achieving high values for the F-score measure, being the models with vector support machines with linear and sigmoid kernels that offer the best predictions, reaching accuracy above 95%. Apparently the preprocessing of the texts does not have much relevance on the predictive power of the models, but if we analyze Figure 5, this process does have a direct impact on the estimates of the prediction errors, these are much less variable and in lower value for the models with stop words and non-lemmatized texts than for the models without stop words and lemmatized texts, which is related to the reduction of information as a consequence of reducing the number of variables in the data set; This is true for the models with vector support machines and the two models with a naive bay classifier, since for the models trained with processed texts that use decision trees (boosting and random forests), they present estimates of prediction errors even better than models with decision trees with raw texts.

4. CONCLUSIONS

The study carried out demonstrates the capacity of the supervised classification algorithms to identify with great precision fake news, with F-score scores above 90%, on political satire pages focused on Ecuadorian problems that circulate on social networks, whose objective is not malicious, it is humorous, but the information that is created and shared can be misinterpreted out of context and can mislead people. Social networks have proven to be a powerful mass communication tool, but they are also an excellent source to obtain data from both users and the pages created on these platforms. Information extracted from Facebook

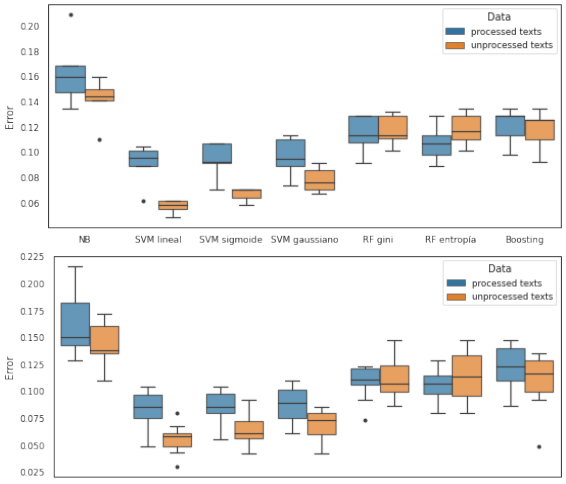

Figure 5. Estimation of the expected prediction error using cross validation with 5 (first) and 10 (second) groups for each of the proposed models with their variants

Table 3. Comparison of mean prediction error estimate

comparation		statistic	p-value
SVM linear	NB gaussian	7.522058	2.138598e-06
	RF gini	3.495373	0.002619
	RF entropy	3.092006	0.006365
	Boosting trees	4.199826	0.000630
SVM sigmoid	NB gaussian	7.507324	3.397924e-06
	RF gini	3.370072	0.003411
	RF entropy	2.940428	0.008746
	Boosting trees	4.093132	0.000895
SVM gaussian	NB gaussian	7.227926	4.703457e-06
	RF gini	2.973300	0.008146
	RF entropy	2.547322	0.020219
	Boosting trees	3.782543	0.001662
SVM linear*	NB gaussian*	12.078495	1.391164e-09
	RF gini*	7.655993	8.332308e-07
	RF entropy*	6.752743	7.627163e-06
	Boosting trees*	6.125672	1.846512e-05
SVM sigmoid*	NB gaussian*	10.544055	6.033120e-09
	RF gini*	6.304513	7.377372e-06
	RF entropy*	5.638357	4.115911e-05
	Boosting trees*	4.966191	1.356008e-04
SVM gaussian*	NB gaussian*	9.976353	1.608514e-08
	RF gini*	5.671745	2.765814e-05
	RF entropy*	5.069808	1.329329e-04
	Boosting trees*	4.363540	4.935202e-04

was vitally important to create the news corpus used to train the models, since without an expert verified fake news database, these pages were the only easily verifiable source for news with fake content and accessible to anyone.

Numerical representation of texts in the form of vectors was a very important aspect in this work, since it was the main piece on which it was based to obtain the necessary data for the training process. Inverse term frequency technique captured the form of writing to a fake news since as we saw this is based on calculating the importance of each of the words both at the level of the news and at a more general level within the corpus. The writing style of fake news is similar to that of a real news story since it pretends to simulate being one but the type of words that are used to elaborate them are different, especially because they make use of words typical of the Ecuadorian slang and it is precisely this what the

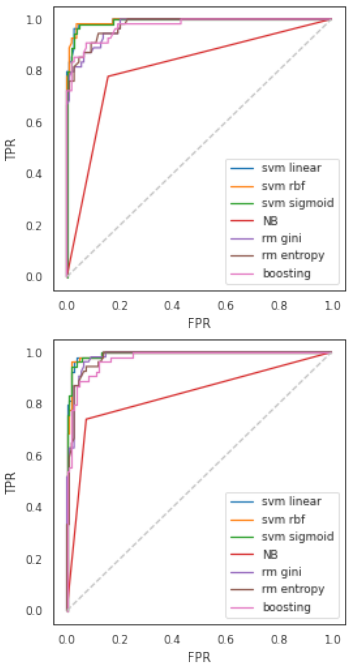

Figure 6. Roc curves of the models. Models trained with lemmatized texts and without stop words (left) models trained with non-lemmatized texts and with stop words (right)

inverse term frequency manages to represent numerically. The number of variables resulting from the numerical representation of the news of the corpus for this case was close to ten thousand, which translates to a large computational expense, so that previously processing the texts was of great help to reduce the time of execution of the algorithms without losing almost any information and maintaining excellent performance of the models when predictions are made.

This study has been carried out with public data from social networks and from national newspaper pages, the investigation could be deepened by creating a more sophisticated model with the help of fact-checking experts to create a larger news corpus addressing more topics in order to generalize the detection of fake news. The model created has a great application within the technological field, thus being able to be used as a tool to help both individuals and companies to directly combat misinformation, especially in crisis situations in which social networks play a fundamental role within communication, which would achieve a better understanding of the behavior of users on social networks during critical events. This work is an advance with respect to the detection of fake news and natural language processing in Spanish, so it is recommended to deepen with different methodologies and adding elements such as images, videos or audios and new text processing techniques in the Spanish language.

REFERENCES

Allcott, H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* (31), 211-236. https://doi.org/10.1257/jep.31.2.211

Bergström, A. and Jervelycke Belfrage, M. (2018). News in Social Media: Incidental consumption and the

role of opinion leaders. *Digital Journalism* (6), 1-16. http://dx.doi.org/10.1080/21670811.2018.1423625

Bondielli, A. and Marcelloni, F. (2019). A Survey on Fake News and Rumour Detection Techniques. *Information Sciences* (497), 38-55. https://doi.org/10.1016/j.ins.2019.05.035

Bowyer, K., Chawla, N., Hall, L., Kegelmeyer, W. (2011). SMOTE: Synthetic Minority Over-sampling Technique *J. Artif. Intell. Res. (JAIR)* (16), 321-357. https://doi.org/10.1613/jair.953

Ciampaglia, G., Shiralkar, P., Rocha, L., Bollen, J., Menczer, F., Flammini, A. (2015). Computational Fact Checking from Knowledge Networks. *PloS one* (10). http://dx.doi.org/10.1371/journal.pone.0128193

Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S., Sunstein, C., Thorson, E., Watts, D., Zittrain, J. (2018). The science of fake news. *Science* (359), 1094-1096. http://dx.doi.org/10.1126/science.aao2998

López-Borrull, A., Vives-Gràcia, J. and Badell, J. (2018). Fake news, ¿amenaza u oportunidad para los profesionales de la información y la documentación? *El Profesional de la Información* (27), 1346. http://dx.doi.org/10.3145/epi.2018.nov.17

Posadas Durán, J., Gomez Adorno, H., Sidorov, G., Moreno, J. (2019). Detection of fake news in a new corpus for the Spanish language *Journal of Intelligent & Fuzzy Systems* (36), 4869-4876. http://dx.doi.org/10.3233/JIFS-179034

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* (12), 2825-2830.

Pulido CM, Ruiz-Eugenio L, Redondo-Sama G, Villarejo-Carballido B. (2020). A New Application of Social Impact in Social Media for Overcoming Fake News in Health. *International journal of environmental research and public health* (17), 2430. http://dx.doi.org/10.3390/ijerph17072430

Shelomi, M. Opinion: Using Pokémon to Detect Scientific Misinformation. Obtained from: https://www.the-scientist.com/. (November, 2020).

Singhania S., Fernandez N., Rao S. (2017). 3HAN: A Deep Neural Network for Fake News Detection. *Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science* (10635).

Soll, J. The Long and Brutal History of Fake News: Bogus news has been around a lot longer than real news. And it's left a lot of destruction behind. Obtained from: https://www.politico.com/magazine/. (November, 2016).

Vajjala, S., Majumder, B., Gupta, A., Surana, H. (2020). *Practical Natural Language Processing*.

Zaki, M. and Meira, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. *Cambridge University Press*, Cambridge University Press, USA.

Zhang, X. and Ghorbani, A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management* (57), 102025. https://doi.org/10.1016/j.ipm.2019.03.0045

Zhou, X. and Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* (53), 40. https://doi.org/10.1145/3395046

BIOGRAPHIES

Nicolás Mafla. Born in Quito, in 2015 he entered to Mathematical Engineering with a mention in Statistics and Operations Research at EPN. During this period he was part of the CLAVEMAT project as a volunteer tutor, providing academic support in fundamental subjects for students of initial engineering levels. He obtained his degree as a Mathematical Engineer in 2021 and currently works as a data scientist in an Ecuadorian fin-tech.

Miguel Flores. Ph.D. in Statistics and Operations Research, Master in Statistical Techniques (University of La Coruña). Has experience in higher education and professional training, university and business in the field of Statistics & Machine Learning. Full Professor of the Probability and Statistics chair at EPN. Member of the Multidisciplinary Research Group on Information Systems, Technology Management and Innovation (SIGTI) of the National Polytechnic School and of the Modeling, Optimization and Statistical Inference Group (MODES) of the University of La Coruña.

Sergio Castillo. Basic academic formation is in Mathematical Engineering with a major in Statistics, Finance and Business Management, from the EPN; with a Higher Specialization in Finance from the Universidad Andina Simón Bolívar; a Master's Degree in University Teaching from ESPE, and then thanks to a

SENESCYT scholarship, He completed his PhD studies in Statistics and Operations Research, at the University of Vigo in Spain, specializing in research fields related to Geostatistics and Non-parametric Statistics.

Roberto Andrade. PhD student in Security Systems at the Faculty of Systems Engineering at EPN, his master's degree is in Network and Telecommunications Management at the Army Polytechnic School in 2013 and his engineering degree is in Electronics and Telecommunications at the National Polytechnic School (EPN) in 2007. Security Officer of the Ecuadorian Ministry of Education (MINEDUC) in 2015, Technological Infrastructure Coordinator at the National Planning Secretariat SENPLADES 2013-2014, Data Center, security and network administration in SENPLADES and Tecnología Sucre 2009-2013 and Technical Engineering for VoIP systems in SERATVoIP 2007-2011. He is a Certified CCNA, CCNP and CCNA Security Technical Instructor at EPN from 2010 to date.

Modelos de Series de Tiempo para Predecir el Número de Casos de Variantes Dominantes del SARS-COV-2 Durante las Olas Epidémicas en Chile

Barría-Sandoval, Claudia [1,2]; **Salas, Patricio** [3]; **Ferreira, Guillermo** [3,*]

[1] *Universidad de las Américas, Escuela de Enfermería, Concepción, Chile*
[2] *Universidad de Concepción, Facultad de Enfermería, Concepción, Chile*
[3] *Universidad de Concepción, Departamento de Estadística, Concepción, Chile*

Resumen: El COVID-19 y sus variantes han creado una pandemia a nivel global. En Chile, hasta el 28 de febrero del 2022, ya se han infectado más de 3 millones de personas y han muerto más de 42 mil personas. En este artículo, se realiza un estudio comparativo de diferentes modelos matemáticos utilizados para modelar y predecir el número de casos diarios confirmados de COVID-19 en Chile. Esta investigación considera los registros diarios de casos confirmados desde el inicio de la pandemia y por lo tanto incluye los contagiados por las distintas variantes del virus (Delta, Gamma y Omicron), estas variantes han dominado la evolución de los contagios diarios en Chile, siendo la variante Omicron la que ha demostrado tener una mayor tasa de contagios a nivel nacional. El objetivo de este estudio es brindar información relevante sobre la evolución de la pandemia por COVID-19 en Chile mediante modelos de series de tiempo que han sido validados en distintas investigaciones y evaluar su precisión frente a la variante Omicron del virus SARS-CoV-2.

Palabras clave: COVID-19; SARS-COV-2; Delta; Gamma; Omicron; Modelo de Series de Tiempo

Time Series Models for Forecasting the Number of Cases of SARS-COV-2 Dominant Variants During the Epidemic Waves in Chile

Abstract: COVID-19 and its variants have created a global pandemic. In Chile, as of February 28 2022, more than 3 million people have been infected and more than 42 thousand people have died. In this article, a comparative study of different mathematical models used to model and predict the number of daily confirmed cases of COVID-19 in Chile is carried out. This research considers the daily records of confirmed cases since the beginning of the pandemic and therefore, includes those infected by the different variants of the virus (Delta, Gamma and Omicron), these variants have dominated the evolution of daily infections in Chile, being the Omicron variant the one that has shown to have a higher rate of infection at national level. The objective of this study is to provide relevant information on the evolution of the COVID-19 pandemic in Chile through time series models that have been validated in different investigations and to assess their validity with the appearance of the Omicron variant of the SARS-CoV-2 virus.

Keywords: COVID-19; SARS-COV-2; Delta; Gamma; Omicron; Time Series Models

1. Introducción

En la ciudad China de Wuhan durante el mes de diciembre del año 2019, comenzó la propagación de un nuevo virus SARS-CoV-2 que causa la enfermedad infectocontagiosa por coronavirus, y desde ese momento se expandió rápidamente por el mundo. El 11 de marzo del año 2020 la OMS, Organización Mundial de la Salud (2020) declaró al COVID-19 como pandemia. El COVID-19 ha tenido un fuerte impacto en la salud a nivel mundial. Dado que este virus ha infectado un gran número de personas causando complicaciones severas, secuelas a largo plazo, defunciones e incremento de mortalidad tanto en Chile como en todo el mundo. En este sentido, los gobiernos y los sistemas de salud han enfocado todo su trabajo y recursos financieros en controlar la propagación de la pandemia por COVID-19. Lo anterior refleja la necesidad de estudiar el comportamiento de la velocidad de propagación y de esta manera brindar información oportuna para tomar decisiones informadas y aplicar las medidas de control pertinentes.

Desde el comienzo de la pandemia, se han propuesto diferentes

*gferreir@udec.cl
Recibido: 10/03/2022
Aceptado: 01/07/2022
Publicado en línea: 23/12/2022
10.33333/rp.vol50n3.02

metodologías para modelar el comportamiento del COVID-19. Entre los modelos más utilizados se encuentran el enfoque econométrico basado en modelos de series de tiempo, los modelos Susceptible-Exposed-Infectious-Removed (SEIR) y el enfoque de aprendizaje automático. Diversos estudios han propuesto modelos del tipo Autoregresivo Integrado de Media Moviles (ARIMA) para predecir el comportamiento del número de contagios por COVID-19. Por ejemplo, Ibrahim et al. (2020) propusieron un modelo del tipo ARIMA de orden $(1,1,0)$ para modelar y realizar predicciones de la propagación del COVID-19 en Nigeria. Talkhi et al. (2021) han realizado un estudio comparativo de diferentes técnicas de series de tiempo y concluyeron que el modelo más adecuado para los datos de casos confirmados por COVID-19 en Irán, fue el Multilayer Perceptron (MLP) y el modelo Holt-Winter para predecir los casos de muerte por COVID-19. En la misma línea, Yonar et al. (2020) propusieron un estudio para predecir y modelar el número de casos de COVID-19 utilizando dos metodologías: ARIMA y Métodos de Suavizado Exponencial. En este trabajo, los autores mencionan que, para los diferentes países en estudio no existe un único modelo para describir el comportamiento del número de casos, pero según las características de los datos, ambos métodos son efectivos para describir las curvas de propagación del virus. Como una alternativa a los métodos econométricos tradicionales están los métodos provenientes del campo de aprendizaje automático (AA). Ghafouri-Fard et al. (2021) realizan una exhaustiva revisión de trabajos de AA que se enfocan en la predicción de la tendencia de propagación del COVID-19. En esta revisión, destacan el beneficio de usar el modelo *Long short-term memory* (LSTM) propuesto por Hochreiter et al. (1997), el cual pertenece a la familia de las redes neuronales recurrentes. Finalmente, Roda et al. (2020) señalaron que las predicciones de la pandemia de COVID-19 que utilizan modelos más complejos pueden no ser más confiables que el uso de un modelo más simple. Se remite al lector a los siguientes autores y sus referencias: Perone (2020), Sarkar (2020), Tran et al. (2020) para complementar la revisión de otros modelos utilizados en las predicciones de COVID-19.

Existen diferentes estudios que han analizado la propagación del COVID-19 en Chile. Por ejemplo, Vicuña et al. (2020) han especificado un modelo de crecimiento logístico generalizado multiparámetro. Entre sus conclusiones mencionan que la política de confinamiento implementada en la mayor parte del país ha demostrado ser eficaz para detener la propagación, y las políticas de confinamiento-relajación, aunque graduales, parecen haber provocado un quiebre al alza en la tendencia. Tariq et al. (2021) han empleado ecuaciones de renovación para estimar el número de reproducción (R) para la fase ascendente temprana de la epidemia de COVID-19 y sus resultados indican una transmisión sostenida temprana de SARS-CoV-2. Por otro lado, Freire-Flores et al. (2021) han utilizado un modelo SEIRD multigrupo para estudiar la propagación de COVID-19 en las diferentes regiones de Chile. Barría-Sandoval et al. (2021) proponen un estudio de diferentes técnicas de series de tiempo para predecir tanto el número de casos confirmados como el número de muertes por COVID-19 en Chile. Barría-Sandoval et al. (2022) proponen un estudio de diferentes modelos de datos de panel para analizar el efecto de la pandemia del COVID-19 sobre las muertes por enfermedades respiratorias en las regiones de Chile.

A diferencia de los estudios antes mencionados, este artículo considera las diferentes variantes del virus que han sido dominantes en Chile, las cuales han sido declaradas como variantes de preocupación (VOC, siglas en inglés de Variant of Concern) por la Organización Mundial de la Salud (2021). De acuerdo a esta clasificación y a los informes del Instituto de Salud Pública (2020) de Chile, las variantes dominantes que han circulado y causado peak de la pandemia en Chile son: "Delta", "Gamma" y "Omicron". A la fecha de la realización de este estudio, la variante Omicron es la dominante y desde su detección el 04 de diciembre de 2021 el número de contagios ha aumentado exponencialmente llegando a cifras de contagios históricas en lo que lleva la pandemia a nivel nacional, pasando de 2060 casos a 22845 casos el 15 de febrero de 2022 (Secretaría de Comunicaciones-Gobierno de Chile, 2021); lo que representa un incremento del 1009 % de los casos confirmados de COVID-19 en los últimos dos meses. De lo anterior, surge el cuestionamiento sobre la efectividad de los modelos de series de tiempo estudiados hasta este momento para predecir la propagación de la pandemia por COVID-19: ¿Son efectivos estos modelos ante el quiebre de tendencia provocado por la variante Omicron?. Esta es la interrogante que esta investigación busca responder.

2. Materiales y Métodos

2.1 Conjunto de Datos y Definición de Variable

En este estudio, se analizó la cantidad de casos diarios confirmados por COVID-19 en Chile desde el 2 de marzo de 2020 al 15 de febrero de 2022. Los datos fueron obtenidos del Ministerio de Ciencia y Tecnología, Conocimiento e Innovación disponibles en el sitio web http://www.minciencia.gob.cl/covid19. La Figura 1 señala el número de casos confirmados de COVID-19 en Chile durante el inicio de la pandemia hasta la fecha de término de este estudio. En esta figura, se observan tres olas de contagios que han producido 3 peaks, los cuales están resaltados por las bandas de color gris claro. La primera ola tiene un peak el 15 de junio del 2020 (bloque gris (a)) con variante dominante Delta. Luego la segunda ola ocurre entre los meses de abril y junio del 2021 con dos peak, uno el 09 de abril y el otro el 05 de junio respectivamente (bloque gris (b)), en este caso la variante dominante fue Gamma. Finalmente, el bloque gris (c) representa el aumento significativo de la Variante Omicron, la cual se inició a principios de enero del 2022.

2.2 Modelos Empíricos

Se realizó una revisión de las metodologías estadísticas utilizadas para modelar los datos registrados sobre el número de contagios diarios confirmados por COVID-19 en Chile a lo largo del tiempo. Y entre ellas se proponen diferentes modelos que permiten capturar la dependencia temporal entre observaciones, incluyendo un modelo proveniente del campo de Machine Learning. A continuación, se describen cada uno de estos modelos:

- **Modelo 1**: ARIMA(p,d,q)

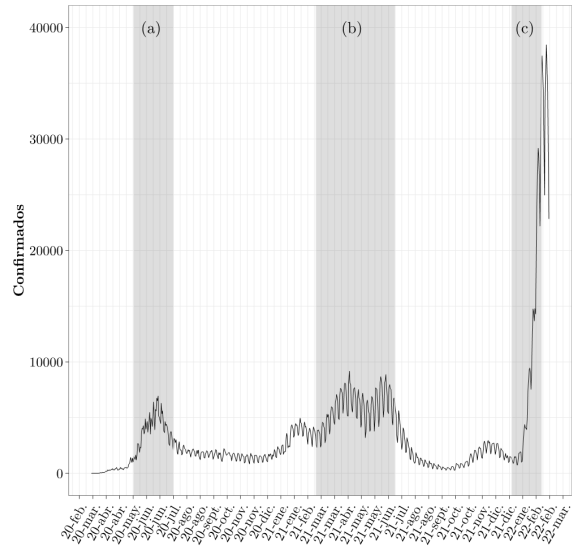

Figura 1. *COVID-19 en Chile:* Casos Confirmados del 03-marzo-2020 al 15-febrero-2022. Paneles (a) Variante Delta; (b) Variante Gamma; (c) Variante Omicron

El modelo autoregresivo integrado de medias móvil (ARIMA) es un modelo del tipo Box-Jenkins. Este tipo de modelo ha sido utilizado en diferentes áreas de la ciencia para estudiar el comportamiento de datos registrados secuencialmente en un periodo de tiempo y fue propuesto por Box et al. (2015) en su libro seminal de 1970 "Time Series Analysis: Forecasting and Control".

Definition 2.1 *Si d es un entero no negativos, entonces* $\{X_t\}$ *es un proceso ARIMA*(p,d,q) *si la serie diferenciada* $(1-B)^d X_t$ *es un proceso causal ARMA. Este proceso puede ser escrito como:*

$$\phi(B)\,(1-B)^d X_t = \theta(B)\,\varepsilon_t \text{ con, } \{\varepsilon_t\} \sim RB(0,\sigma^2)$$

donde $\phi(z) = 1-\phi_1 z-\cdots-\phi_p z^p$, *y* $\theta(z) = 1+\theta_1 z+\cdots+\theta_p z^q$ *corresponden a los polinomios autorregresivo y de medias móviles respectivamente.*

El proceso denotado por *RB* es un proceso de ruido blanco no-correlacionado con media cero y varianza σ^2. El lector puede encontrar una fácil descripción sobre modelos de datos de series temporales en Brockwell et al. (2016).

- **Modelo 2**: Método de Holt-Winters con tendencia amortiguada

Otro modelo para trabajar con datos de series de tiempo es el método de suavizamiento exponencial. Los autores Holt (2004) y Winters (1960) han propuesto una clase más general de métodos para capturar tanto la tendencia como el nivel de la serie de tiempo, la cual es conocida como método de Holt-Winters. Uno de los modelos más simples de este tipo es el suavizamiento exponencial definido por:

$$\begin{aligned} \widehat{X}_{t+h|t} &= \mu_t + hT_t, \\ \mu_t &= \alpha X_t + (1-\alpha)(\mu_{t-1}+T_{t-1}), \qquad \text{Model 2a} \\ T_t &= \beta(\mu_t-\mu_{t-1}) + (1-\beta)T_{t-1}, \end{aligned}$$

donde $0 \le \alpha \le 1$ es el parámetro de suavizamiento y $0 \le \beta \le 1$ parámetro de suavizamiento para la tendencia. Por otro lado, Gardner et al. (1985) desarrollaron un modelo de suavizado exponencial diseñado para amortiguar las tendencias erráticas definido de la siguiente manera:

$$\begin{aligned} \widehat{X}_{t+h|t} &= \mu_t + (\lambda+\lambda^2+\ldots+\lambda^{h-1})T_t, \\ \mu_t &= \alpha X_t + (1-\alpha)(\mu_{t-1}+\lambda T_{t-1}), \qquad \text{Model 2b} \\ T_t &= \beta(\mu_t-\mu_{t-1}) + (1-\beta)\lambda T_{t-1}, \end{aligned}$$

donde $0 < \lambda < 1$ es el parámetro de amortiguación. Los pronósticos de h paso $\widehat{X}_{t+h|t}$ se calculan usando las ecuaciones de suavizado para el nivel μ_t y la tendencia T_t.

- **Modelo 3**: Red Neuronal Recurrente

Una red neuronal recurrente (RNN) es una familia de redes neuronales que permiten procesar datos secuenciales $(x_1,\ldots,x_T)$. Los detalles de la arquitectura de las RNN se pueden encontrar en Goodfellow et al. (2016) y Aggarwal (2018). Una RNN consiste en estados ocultos que se distribuyen de forma temporal y son capaces de predecir los eventos futuros con más precisión que los métodos tradicionales de suavización exponencial Shastri et al. (2020). Los estados ocultos de una RNN son definidos de la siguiente manera:

$$\mathbf{h}^{(t)} = f(\mathbf{h}^{(t-1)}, \mathbf{x}_t; \theta),$$

donde $f(\cdot)$ es una función de activación, x_t son los input y θ son los pesos de la red. En este estudio, se utilizó el modelo Long Short Term Memory (LSTM) desarrollado por Hochreiter et al. (1997), el cual es un tipo especial de RNN capaz de aprender de rezagos o dependencias temporales a largo plazo en los datos. A diferencia de las redes neuronales estándar, este modelo garantiza la conexión de retroalimentación como propone Ghafouri-Fard et al. (2021). Además de procesar puntos de datos individuales, el LSTM puede procesar secuencias completas de datos como lo señala Hochreiter et al. (1997). En la Figura 2, se muestra la arquitectura general de un modelo LSTM.

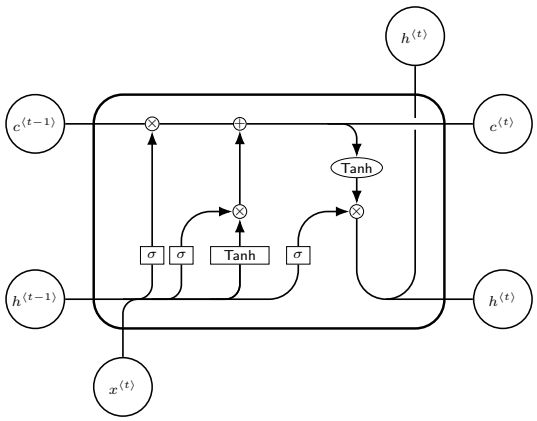

Figura 2. LSTM: Arquitectura del modelo LSTM

El papel central de un modelo LSTM lo desempeña una célula de memoria conocida como "estado de célula" ($c^{\langle t \rangle}$) que mantiene su estado a lo largo del tiempo. El estado de la célula es la línea horizontal que atraviesa la parte superior de la Figura 2. Puede visualizarse como una cinta transportadora a través de la cual la información simplemente fluye, sin cambios.

La información puede añadirse o eliminarse del estado de la célula en la LSTM y se regula mediante puertas □. Estas puertas permiten opcionalmente que la información fluya dentro y fuera de la célula. Por ejemplo, la información se incorpora si la función sigmoide $\sigma(\cdot)$ le asigna un valor 1, se incorpora parcialmente si la $\sigma(\cdot)$ entrega un valor entre 0 y 1 y se descarta si $\sigma(\cdot)$ entrega un valor 0. Luego se procede a actualizar el estado de la célula multiplicando la salida de la puerta $\boxed{\sigma}$ por el estado anterior ($c^{\langle t-1 \rangle}$) mediante la operación de multiplicación puntual ($\otimes$). El proceso continua para actualizar el estado oculto ($h^{\langle t \rangle}$), multiplicando los resultados de la puerta de salida por el resultado que entrega la función de activación $\boxed{\text{Tanh}}$ que tiene como argumento de entrada el estado de la célula actualizado $c^{\langle t \rangle}$. Finalmente, el último estado de celda ($c^{\langle t \rangle}$) y el estado oculto ($h^{\langle t \rangle}$) regresan a la unidad recurrente y el proceso se repite en el paso de tiempo $t+1$. El ciclo continúa hasta que se llega al final de la secuencia.

- **Modelo 4**: Modelo Híbrido
  Los modelos híbridos permiten interactuar con diferentes métodos de predicción mediante una relación lineal ponderada. En particular, el paquete `forecastHybrid` del software libre `R` Team (2013) proporciona predicciones a h pasos ponderando las predicciones de m modelos individuales como sigue:

$$f(i) = \sum_{m=1}^{n} \omega_m f_m(i), \text{ con } 1 \leq i \leq h,$$

  donde $f(i)$ representa la i-ésima predicción total, ω_m son los pesos y $f_m(i)$ son las predicciones de m componentes individuales de los siguientes modelos:

  - **ARIMA**
  - Modelos de suavizamiento exponencial **ETS**, los cuales ajustan datos con una componente de tendencia (T), componente estacional (S) y un término de error (E).
  - El **modelo Theta** propuesto por Assimakopoulos et al. (2000) permite obtener una línea Theta denotada por $Z(\theta)$ la cual se logra con la solución de la ecuación:

$$\nabla^2 Z_t(\theta) = \theta \nabla^2 X_t, \tag{1}$$

    donde $\{X_t\}$ son las observaciones y ∇ es el operador de diferencias (es decir, $\nabla Y_t = Y_t - Y_{t-1}$). Una solución analítica de (1) es dada por:

$$Z_t(\theta) = \theta X_t + (1-\theta)(a+bt),$$

    donde a y b son constantes obtenidas al minimizar $\sum_{t=1}^{n} [X_t - Z_t(\theta)]^2$.

  - Red Neuronal con valores rezagados de la serie de tiempo como entradas y posiblemente algunas otras entradas exógenas, denotado por **NNETAR** en la autoría de Hyndman et al. (2002).
  - Otros modelos tales como: **STL** (Seasonal Decomposition of Time Series by Loess), **TBATS** (Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components) y **SNAIVE**, se pueden revisar en detalle en De Livera et al. (2012).

  Finalmente, el paquete `forecastHybrid` permite utilizar todos los modelos antes mencionados a partir de los comandos `auto.arima()`, `ets()`, `them()`, `nnetar()`, `stlm()`, `tbats()` y `snaive()`; los cuales se pueden combinar con pesos iguales, pesos que utilizan errores en la muestra, o determinados por validación cruzada según lo mencionado por Shaub et al. (2020). Estas formas de ponderación originan tres modelos diferentes denotados en este estudio como Modelo 4a, Modelo 4b y Modelo 4c, respectivamente.

3. RESULTADOS

Se analizó el desempeño de los Modelos $1-4$ descritos en la Sección 2. Los datos analizados fueron la cantidad de casos con un diagnóstico confirmado de COVID-19 en Chile desde 2 de marzo de 2020 al 15 de febrero de 2022 (ver Figura 1). Desde esta figura, se puede apreciar un aumento significativo de la varianza al final del periodo provocado por la variante Omicron del COVID-19. Para estabilizar la varianza se realizó una transformación logarítmica a las observaciones, es decir, $Y_t = \log X_t$. Tanto las estimaciones como las predicciones fueron obtenidas a partir de la serie $\{Y_t\}$ para luego transformar a su escala original por la utilización de la función inversa, es decir $\widehat{X}_t = \exp(\widehat{Y}_t)$. Esta transformación ha sido aplicada por Feroze (2021) para evaluar la progresión del COVID-19 en Irán. A continuación, se describen los tipos de modelos utilizados para los datos, los comandos y paquetes del software libre `R` para la estimación de parámetros.

Para el Modelo 1, se aplicó un modelo ARIMA$(3,1,3)$ con media constante. Tales estimaciones se obtuvieron usando el comando `Arima` del paquete `forecast`. En el caso de los Modelos 2a, las estimaciones fueron obtenidas a través del comando `holt` del paquete `forecast` y para el Modelo 2b se utilizó el mismo comando y agregó el argumento `damped=TRUE`.

El Modelo 3 fue construido en Google Colaboratory utilizando `Python 3.0` con librerías de código abierto como: `Tensorflow` propuesto por Abadi et al. (2016), `Pandas` elaborada por McKinney (2010), `Numpy` desarrollada por Oliphant (2006) y `Keras` construida por Chollet (2018). La arquitectura considerada para el Modelo 3 fue simple; donde la primera capa de entrada de la red estuvo compuesta por 15 neuronas con función de activación tangente hiperbólica. Posteriormente, la capa de salida estuvo conformada por 1 neurona. Para el entrenamiento de este modelo se utilizó el optimizador ADAM y el cuadrado medio del error como función de pérdida para evaluar

el desempeño del mismo.

Para el Modelo 4, se utilizó la función `hybridModel` del paquete `forecastHybrid`, el argumento `weights` que permitió seleccionar las ponderaciones de los métodos que interactúan en los modelos híbridos. En particular, las ponderaciones del tipo `equal, insample.errors, cv.errors` fueron implementadas para generar las predicciones de los Modelos 4a, 4b y 4c respectivamente.

La Tabla 1 muestra las estimaciones de los parámetros y la desviación estándard (d.e.) estimada de los Modelos $1-2$. Para probar la importancia de las estimaciones de los parámetros se aplicó la estadística $t = \widehat{\theta}/de(\widehat{\theta})$ al Modelo 1 para el conjunto de datos. En la última columna de esta tabla, se puede observar que las estadísticas t son altamente significativas al 5% de nivel de confianza. Por otra parte, la Tabla 2 reporta las ponderaciones de los Modelos 4a, 4b y 4c. La Figura 3 muestra los valores ajustados para cada modelo, donde las líneas discontinuas representan los datos reales, mientras que las líneas continuas representan los valores ajustados y se despliegan en las curvas de colores para los modelos $1-4$. A partir de esta figura, se puede concluir que el mejor método que captura la tendencia y la estructura de dependencia temporal son los modelos del grupo 4.

Tabla 1. Casos Confirmados de la serie COVID-19: Parámetros estimados de los Modelos $1-3$

Modelo	Parámetros	Estimaciones	d.e.	$\lvert t\rvert$
Modelo 1	ϕ_1	$-0,23$	$0,07$	$3,25$
	ϕ_2	$0,19$	$0,06$	$3,22$
	ϕ_3	$-0,29$	$0,05$	$5,67$
	θ_1	$-0,92$	$0,06$	$15,65$
	θ_2	$-0,77$	$0,09$	$8,38$
	θ_3	$0,79$	$0,05$	$16,68$
	σ^2	$0,05$	–	–
Modelo 2a	α	$0,99$	–	–
	β	10^{-4}	–	–
Modelo 2b	α	$0,99$	–	–
	β	10^{-4}	–	–
	λ	$0,98$	–	–

Tabla 2. Casos Confirmados de la serie COVID-19: Ponderaciones del Modelo 4

	ARIMA	ETS	Tthetam	NNETAR	TBATS	SNAIVE
				Pesos		
4a	0,17	0,17	0,17	0,17	0,17	0,17
4b	0,13	0,11	0,11	0,43	0,12	0,11
4c	0,19	0,16	0,16	0,14	0,19	0,16

Para la identificación del mejor modelo, en la siguiente sección se propone un análisis de precisión para predicciones realizadas con datos de entrenamiento y datos de prueba.

3.1 Análisis de la precisión del pronóstico ex post

Se evaluó la precisión de las predicciones usando un conjunto de entrenamiento y un conjunto de prueba. Se consideró un conjunto de entrenamiento $\{X_t\}$ desde 02 de marzo 2020 al 10 de febrero 2022 (período de estimación) con un total de $N = 709$ observaciones y datos de prueba $\{X_t\}$ del 11 al 15 de febrero 2022 (período validación) que se utilizó para la predicción paso a paso de m con $m = 5$ observaciones diarias.

Los valores $\widehat{X}_{N+1},\ldots,\widehat{X}_{N+m}$ se denominan pronóstico ex post o pronóstico del período. Los pronósticos m-step-ahead se compararon con el período de validación, lo que da lugar a errores de pronóstico ex post, es decir, $X_{N+h} - \widehat{X}_{N+h}$ para el horizonte $h = 1,\ldots,m$. Los errores fueron evaluados por las estadísticas de los residuales, como el error medio (ME), error cuadrático medio (RMSE), error absoluto medio (MAE), error porcentual medio (MPE) y error porcentual absoluto medio (MAPE), definidos a continuación:

$$\text{MAE:} \quad \frac{1}{n}\sum_{t=1}^{n}|X_t - \widehat{X}_t|,$$

$$\text{RMSE:} \quad \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \widehat{X}_t\right)^2},$$

$$\text{MPE} = \quad \frac{1}{n}\sum_{t=1}^{n}100(X_t - \widehat{X}_t)/X_t,$$

$$\text{MAPE} = \quad \frac{1}{n}\sum_{t=1}^{n}100|(X_t - \widehat{X}_t)/X_t|$$

Los valores pequeños de estas estadísticas reflejan una mejor bondad de ajuste. La Tabla 3 informa las estadísticas de errores de pronóstico tanto para los datos de entrenamiento como los de prueba. En ambos casos, todos los indicadores favorecieron el modelo híbrido con ponderaciones que utilizaron errores en la muestra, es decir, el Modelo 4b con un MAPE de $0,62\%$ en los datos de prueba. Considerando lo señalado por Lewis (1982) (pág. 40) quien estableció criterios de interpretación de los valores del MAPE argumentando que, en un valor MAPE $<10\%$ la predicción tiene una alta precisión. En nuestro caso, todos los modelos tienen un MAPE menor al 10%; por lo tanto, se puede inferir que a pesar de la llegada de la variante Omicron los modelos de series de tiempo utilizados en este estudio siguen siendo válidos. En particular, se puede señalar que el Modelo 4b posse una mejor performance que el resto de los modelos, demostrando la idoneidad para la predicción del COVID-19.

Los hallazgos anteriores están respaldados por la Figura 4, donde se puede observar que las predicciones (puntos en color) del Modelo 4b son más cercanas a los valores reales denotados por los puntos de color negro.

4. CONCLUSIONES

Este estudio consideró el modelamiento de los datos de casos confirmados por COVID-19 en Chile, con el fin de establecer el mejor modelo para pronosticar el comportamiento de esta enfermedad en presencia de la variante Omicron. Para este propósito se utilizaron distintos modelos para ajustar los datos de COVID-19: SARIMA, Holt-Winter, Holt-Winter Damped Trend, Hybrid, BSTS, TBATS y LSTM.

Los hallazgos de este estudio concluyen que el modelo con menor error de pronóstico en datos de casos confirmados de COVID-19 en Chile es el modelo híbrido (Modelo 4b), el que

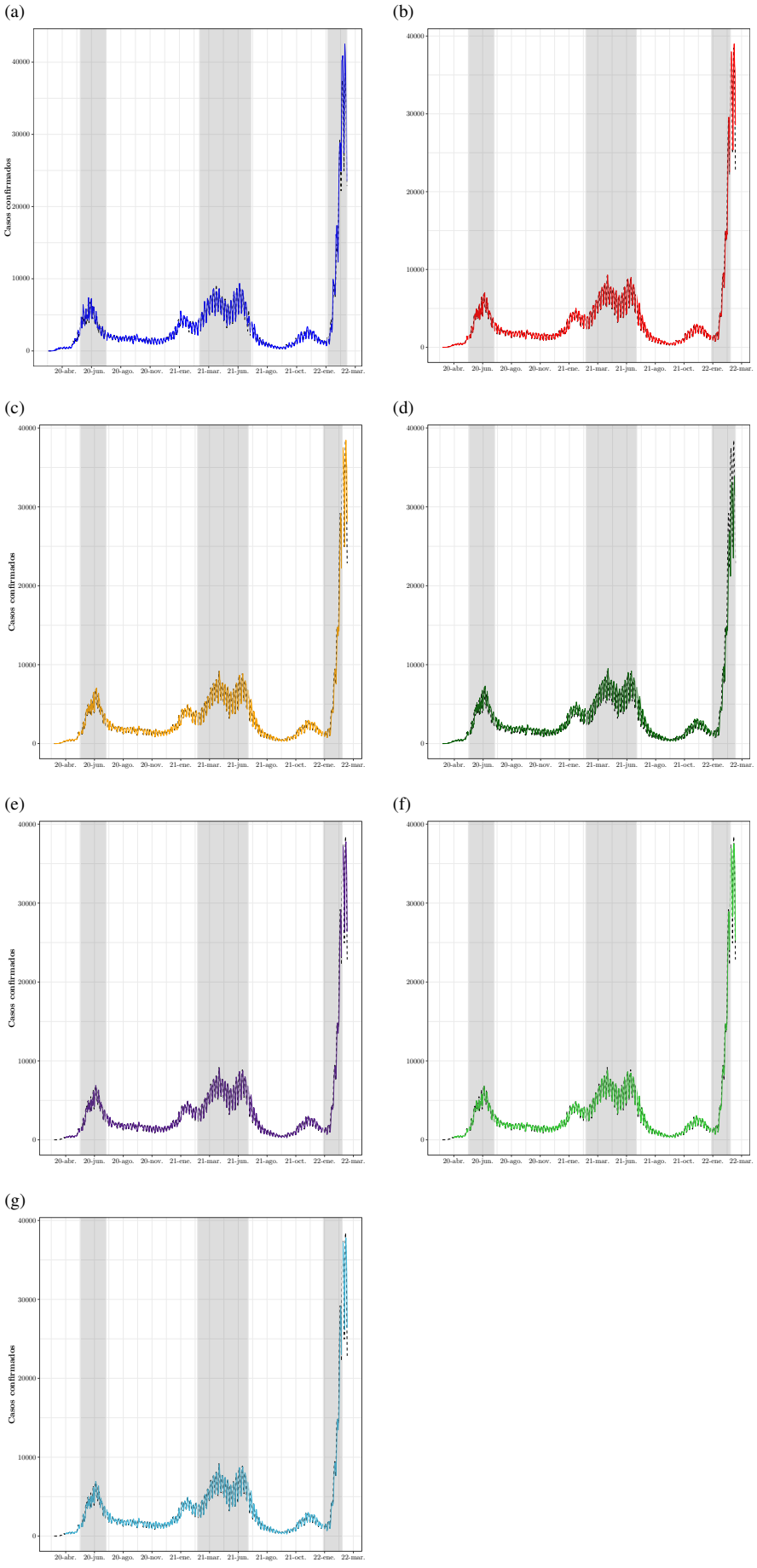

Figura 3. COVID-19 en Chile: Casos confirmados (líneas discontinuas negras) versus valores ajustados (línea continua). (a) Modelo 1. (b) Modelo 2a. (c) Modelo 2b. (d) Modelo 3. (e) Modelo 4a. (f) Modelo 4b. (g) Modelo 4c

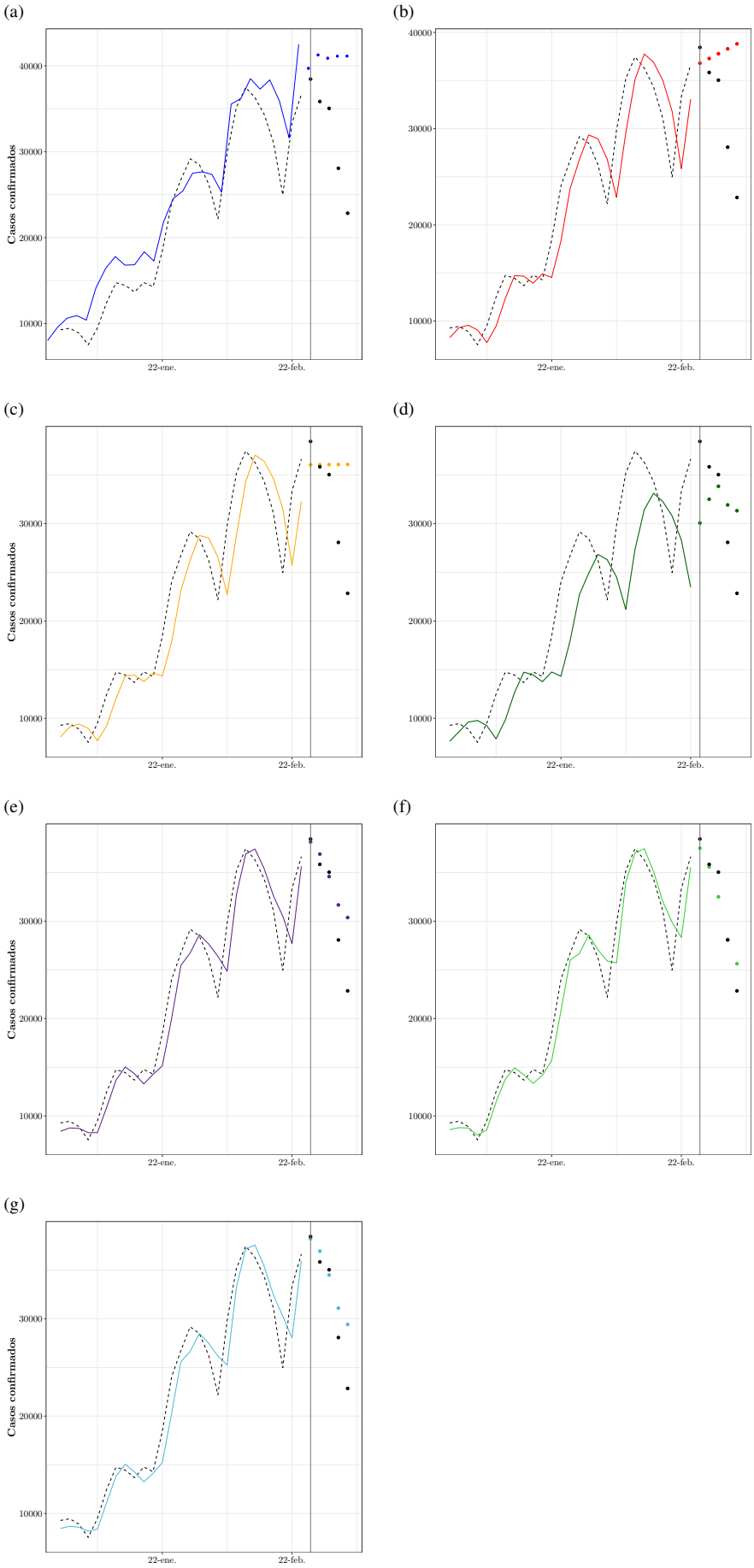

Figura 4. *Predicción para casos confirmados de COVID-19.* La línea continua y los puntos representan valores ajustados y el pronóstico ex post, respectivamente. (a) Modelo 1. (b) Modelo 2a. (c) Modelo 2b. (d) Modelo 3. (e) Modelo 4a. (f) Modelo 4b. (g) Modelo 4c

Tabla 3. Estadísticas descriptivas de los errores de pronóstico ex post

Modelo	ME	RMSE	MAE	MPE	MAPE
		Training Data			
Modelo 1	$-2{,}8\times10^{-3}$	0,21	0,16	-0,11	2,50
Modelo 2a	$5{,}3\times10^{-4}$	0,25	0,18	0,01	2,83
Modelo 2b	$5{,}6\times10^{-3}$	0,25	0,18	-0,11	2,79
Modelo 3	$-3{,}3\times10^{-2}$	0,35	0.28	-0.26	4.38
Modelo 4a	$3{,}6\times10^{-3}$	0,16	0,12	$6{,}8\times10^{-3}$	1,64
Modelo 4b	$\mathbf{2{,}7\times10^{-3}}$	**0,13**	**0.10**	$\mathbf{2{,}1\times10^{-3}}$	**1,37**
Modelo 4c	$3{,}2\times10^{-3}$	0,15	0,12	$3{,}3\times10^{-3}$	1,57
		Testing Data			
Modelo 1	-0,24	0,31	0,24	-2,35	2,35
Modelo 2a	-0,18	0,28	0,20	-1,79	1,97
Modelo 2b	-0,14	0,24	0,16	-1,34	1,59
Modelo 3	-0,01	0,19	0,16	-0,16	1,60
Modelo 4a	$-8{,}9\times10^{-2}$	0,14	0,09	-0,88	0,92
Modelo 4b	$\mathbf{-4.9\times10^{-2}}$	**0,09**	**0,06**	**- 0,49**	**0,62**
Modelo 4c	$-8{,}1\times10^{-2}$	0,13	0,08	-0,80	0,82

permite realizar una predicción con mayor precisión que los otros modelos considerados en este trabajo. En este contexto, podemos sostener que los modelos considerados tienen una alta precisión de predicción, siendo de gran utilidad para analizar el comportamineto dinámico temporal del COVID-19 frente a las distintas variantes presentes en el periodo de estudio. Se debe tener presente que, es difícil encontrar un modelo estadístico que tenga una precisión en la predicción de los datos cercana al 100%, dado que a menudo se exponen a una correlación serial y a estructuras de cambio no-estocásticas. Por lo tanto, no existe un modelo universal capaz de capturar de manera precisa el comportamiento futuro de las olas pandémicas por COVID-19, debido a que en el análisis de series de tiempo se deben considerar componentes de tendencias, ciclos y variables externas como las características biosociodemográficas particulares de cada país.

En particular, en nuestro estudio, los modelos incluidos siguen siendo útiles para capturar la evolución del número de contagios por COVID-19 en presencia de las diferentes variantes dominantes en Chile. A la vez, se destaca que, todos los modelos utilizados han sido validados por otros autores en diversas bases de datos de COVID-19 a nivel mundial. Para complementar lo anterior, se puede leer el trabajo realizado por Chakraborty et al. (2022) quienes han presentado un interesante resumen de los modelos utilizados para predecir el número de casos por COVID-19 en diferentes países e informaron cuales de estos modelos son los más apropiados para capturar la dinámica temporal de la pandemia.

Es relevante mencionar que este estudio, se incorpora al grupo de las escasas investigaciones existentes para analizar modelos de series de tiempo en presencia de la variante Omicron. Según la tendencia de los datos y los resultados de las predicciones, el número de casos confirmados por COVID-19 en Chile, está disminuyendo. Sin embargo, dado que la variante Omicron es más contagiosa se sugiere continuar en alerta y mantener las medidas preventivas indicadas por las autoridades sanitarias de Chile.

REFERENCIAS

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *USENIX Association, OSDI, 16*, 265-283.

Aggarwal, C.C. (2018). *Neural Networks and Deep Learning*. Springer.

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International journal of forecasting, 16*(4), 521-530. https://doi.org/10.1016/S0169-2070(00)00066-2

Barría-Sandoval, C., Ferreira, G., Benz-Parra, K., & López-Flores, P. (2021). Prediction of confirmed cases of and deaths caused by COVID-19 in Chile through time series techniques: A comparative study. *Plos one, 16*(4), e0245414. https://doi.org/ 10.1371/journal.pone.0245414

Barría-Sandoval, C., Ferreira, G., Méndez, A., & Toffoletto, M. C. (2022). Impact of COVID-19 on deaths from respiratory diseases: Panel data evidence from Chile. *Infection Ecology & Epidemiology, 12*(1), 2023939. https://doi.org/10.1080/20008686.2021.2023939

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control. (5^aed.).JohnWiley&Sons.*

Brockwell, Peter J & Davis, Richard A. (2016). *Introduction to Time Series and Forecasting. (3^aed.)*. Springer.

Chakraborty, Tanujit and Ghosh, Indrajit and Mahajan, Tirna and Arora, Tejasvi. (2022). Nowcasting of COVID-19 confirmed cases: Foundations, trends, and challenges. *Modeling, Control and Drug Development for COVID-19 Outbreak Prevention.* Springer, 1023–1064.

Chollet, F. (2018). Keras: The Python Deep Learning library. Astrophysics Source Code Library. *ASCL Code Record.* https://ascl.net/1806.022

De Livera A. M., Hyndman, R. J., & Snyder, R. D. (2012). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American statistical association, 106*(496), 1513-1527. https://doi.org/10.1198/jasa.2011.tm09771

Freire-Flores, D., Llanovarced-Kawles, N., Sanchez-Daza, A., & Olivera-Nappa, Á. (2021). On the heterogeneous spread of COVID-19 in Chile. *Chaos, Solitons & Fractals, 150*, 111156. https://doi.org/10.1016/j.chaos.2021.111156

Feroze, Navid. (2021). Assessing the future progression of COVID-19 in Iran and its neighbors using Bayesian models. *Infectious Disease Modelling, 6*, 343–350.

Gardner Jr, Everette S and McKenzie, ED. (1985). Forecasting trends in time series. *Management Science, 31*(10), 1237-1246. http://dx.doi.org/10.1287/mnsc.31.10.1237

Ghafouri-Fard, S., Mohammad-Rahimi, H., Motie, P., Minabi, M. A., Taheri, M., & Nateghinia, S. (2021). Application of machine learning in the prediction of covid-19 daily new cases: A scoping review. *Heliyon, 7*(10), e08143. https://doi.org/10.1016/j.heliyon.2021.e08143

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holt, CC. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting, 20*(1), 5-10. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., ... & Wang, E. (2020). Forecast: Forecasting functions for time series and linear models. *The R Foundation, package Version 8.16*. https://pkg.robjhyndman.com/forecast/

Ibrahim, R. R., & Oladipo, H. O. (2020). Forecasting the spread of COVID-19 in Nigeria using Box-Jenkins modeling procedure. *medRxiv*, 1-15. https://doi.org/10.1101/2020.05.05.20091686

Instituto de Salud Pública Ministerio de Salud Gobierno de Chile (2020). Vigilancia Genómica SARS-Cov2. Obtenido de: https://vigilancia.ispch.gob.cl/app/varcovid. (Marzo, 2020).

Lewis, C.D. (1982). Industrial and business forecasting methods. *London: Butterworths.*

McKinney, W. (2010). Data structures for statistical computing in python. *In Proceedings of the 9th Python in Science Conference, 445*(1), 51-56.

Oliphant, T. E. (2006). A guide to NumPy. *Trelgol Publishing*. https://web.mit.edu/dvp/Public/numpybook.pdf

Organización Mundial de la Salud (2020, Marzo). Alocución de apertura del Director General de la OMS en la rueda de prensa sobre la COVID-19 celebrada el 11 de marzo de 2020. Obtenido de: https://www.who.int/es/director-general/speeches.

Organización Mundial de la Salud (2021, Marzo). Seguimiento de las variantes del SARS-CoV-2. Obtenido de: https://www.who.int/es/activities/tracking-SARS-CoV-2-variants.

Perone, G.(2020). An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in Italy. *medRxiv* 1-14. https://doi.org/10.1101/2020.04.27.20081539.

Roda, Weston C and Varughese, Marie B and Han, Donglin & Li, Michael Y.(2020).Why is it difficult to accurately predict the COVID-19 epidemic?. *Infectious Disease Modelling.5*: 271–281. https://doi.org/10.1016/j.idm.2020.03.001

Sarkar, D., Biswas M. (2020). COVID 19 Pandemic: A Real-time Forecasts & Prediction of Confirmed Cases, Active Cases using the ARIMA model & Public Health in West Bengal, India. *medRxiv* 1-22. doi: https://doi.org/10.1101/2020.06.06.20124180.

Secretaría de Comunicaciones - MSGG Gobierno de Chile. (2021, Marzo). Cifras Oficiales COVID-19. Obtenido de: https://www.gob.cl/coronavirus/cifrasoficiales/datos.

Shaub, D., & Ellis, P. (2020). forecastHybrid: Convenient Functions for Ensemble Time Series Forecasts. *The R package version 5.0.19*. https://CRAN.R-project.org/package=forecastHybrid.

Shastri, S., Singh, K., Kumar, S., Kour, P., & Mansotra, V. (2020). Time series forecasting of covid-19 using deep learning models: India-usa comparative case study. *Chaos, Solitons & Fractals*, 140:110227. https://doi.org/10.1016/j.chaos.2020.110227

Talkhi, N., Fatemi, N. A., Ataei, Z., & Nooghabi, M. J. (2021). Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods. *Biomedical Signal Processing and Control, 66*(102494), 1-8. https://doi.org/10.1016/j.bspc.2021.102494

Tariq, A., Undurraga, E. A., Laborde, C. C., Vogt-Geisse, K., Luo, R., Rothenberg, R., & Chowell, G. (2021). Transmission dynamics and control of COVID-19 in Chile, March-October, 2020. *PLoS neglected tropical diseases, 15*(1), e0009070. https://doi.org/10.1371/journal. pntd.0009070

Team, R.C. (2013). R: A Language and Environment for Statistical Computing. *The R Foundation*. http://www.R-project.org/

Tran, TT and Pham, LT and Ngo, QX.(2020).Forecasting epidemic spread of SARS-CoV-2 using ARIMA model (Case study: Iran). *Global Journal of Environmental Science and Management. 2020;6*(4), 1-10. https://doi.org/10.22034/GJESM.2019.06.SI.01

Vicuña, M. I., Vásquez, C., & Quiroga, B. F. (2021). Forecasting the 2020 COVID-19 epidemic: A multivariate Quasi-Poisson regression to model the evolution of new cases in Chile. *Frontiers in public health, 9* (610479), 1-7. https://doi.org/10.3389/fpubh.2021.610479

Winters, P.R. (1960). Forecasting sales by exponentially weighted moving averages.*Management Science, 6*(3), 231-362. https://doi.org/10.1287/mnsc.6.3.324

Yonar, Harun and Yonar, Aynur and Tekindal, Mustafa Agah & Tekindal, Melike.(2020). Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods.*Eurasian Journal of Medicine and Oncology 4*(2): 160–165. https://ejmo.org/10.14744/ejmo.2020.28273/

BIOGRAFÍAS

Claudia Barría-Sandoval, Académica de Enfermería de la Universidad de las Américas, Concepción, Chile. PhD (c) en Enfermería, Universidad de Concepción, Chile. Mg. en Salud Pública y Gestión Sanitaria, Universidad de Valencia, España. Sus intereses de investigación están orientados a la Salud Pública, Enfermería y Atención Integral en la Comunidad.

Patricio Salas, Profesor Asistente en el Departamento de Estadística de la Universidad de Concepción, Chile. PhD (c) en Ingeniería Industrial de la Universidad de Concepción, Chile. Mg. en Estadística Aplicada, Universidad de Concepción, Chile. Ingeniero Estadístico, Universidad de Concepción, Chile. Sus intereses de investigación incluyen modelos econométricos de elección discreta, machine learning, modelos basados en agentes (ABM), series de tiempo y simulación de eventos discretos.

Guillermo Ferreira, PhD en Estadística de la Pontificia Universidad Católica de Chile. Académico Asociado en el Departamento de Estadística de la Universidad de Concepción, Chile. Sus intereses de investigación incluyen el análisis y predicción de series de tiempo, modelos econométricos, procesos espacio-temporal y modelos epidemiológicos.

Análisis y Diseño de un Modelo Predictivo para Detección de Phishing Basado en Url y Corpus del Correo Electrónico

Albán, Fernanda [1,*]; **Urvina, Menthor** [2]; **Andrade, Roberto** [3]

[1] *Escuela Politécnica Nacional, Facultad de Ciencias, Quito, Ecuador*
[2] *Escuela Politécnica Nacional, Departamento de Matemática, Quito, Ecuador*
[3] *Escuela Politécnica Nacional, Departamento de Informática y Ciencias de la Computación, Quito, Ecuador*

Resumen: Uno de los delitos cibernéticos más reportados a nivel mundial es el phishing. En la actualidad se están desarrollando diversos sistemas anti-phishing (APS) para identificar este tipo de ataque en sistemas de comunicación en tiempo real. A pesar de los esfuerzos de las organizaciones, este ataque continúa creciendo, teniendo como causas: la detección errónea en el ataque de día cero, el alto costo computacional y las tasas altas de falsificación. Aunque el enfoque de Machine Learning (ML) ha logrado una tasa de precisión favorable, se debe considerar que la elección y el rendimiento del vector de características es un punto clave para obtener un nivel de precisión adecuado. En este trabajo, se propone un modelo predictivo basado en ML y en el análisis de la eficiencia de algunos esquemas anti-phishing que sirvieron para entender esta temática. El modelo propuesto consta de un módulo de selección de características que se utiliza para la construcción del vector final. Estas características se extraen de la URL, las propiedades de la página web y del corpus de correo electrónico. El sistema utiliza los modelos de clasificación, Random Forest (RF) y Naïve Bayes (NB), que han sido entrenados en el vector de características. Los experimentos se basaron en Dataset compuestas por instancias de phishing y benignas. Mediante el uso de la validación cruzada, los resultados experimentales indican una precisión del 97,5 % para los dataset utilizados, mientras que para el abordaje de esta investigación a nivel local se obtuvo una precisión del 96,5 %.

Palabras clave: Anti-phishing; Ataques cibernéticos; Phishing; Middleware; Amenaza

Analysis and Design of a Predictive Model for Phishing Detection Based on Url and Email Corpus

Abstract: One of the most reported cyber crimes worldwide is phishing, and various anti-phishing systems (APS) are currently being developed to identify this type of attack on communication systems in real time. Despite the efforts of organizations, this attack continues to grow, due to the erroneous detection in the zero-day attack: the high computational cost and the high rates of forgery. Although the Machine Learning (ML) approach has achieved a favorable accuracy rate, it should be considered that the choice and performance of the feature vector is a key point to obtain an adequate level of accuracy. In this work, a predictive model based on ML and the analysis of the efficiency of some anti-phishing schemes that served to understand this issue is proposed. The proposed model consists of a feature selection module that is used to build the final vector. These characteristics are extracted from the URL, the properties of the web page, and the email corpus. The system uses the Random Forest (RF) and Naïve Bayes (NB) classification models, which have been trained on the feature vector. The experiments were based on datasets composed of phishing and benign instances. Using cross-validation, the experimental results indicate a precision of 97.5 % for the datasets used, while a precision of 96.5 % was obtained for the approach of this research at the local level.

Keywords: Anti-phishing; Cyberattacks; Phishing; Middleware; Threat

1. Introducción

Phishing es una clase de ataque cibernético, del tipo ingeniería social, que se usa frecuentemente para el robo de datos personales, tales como: las credenciales de inicio de sesión y los números de tarjetas de crédito. Las técnicas de ingeniería social pretenden adquirir la identidad de los usuarios ingenuos o información confidencial sensible mediante el uso de correos electrónicos falsificados, sitios web falsos, anuncios/promociones dudosas en línea, SMS falsos de proveedores de servicios o empresas, entre otros. El objetivo principal de los phishers es: atacar a las grandes corporaciones, instituciones financieras y gubernamentales que

dolores.alban@epn.edu.ec
Recibido: 10/03/2022
Aceptado: 27/05/2022
Publicado en línea: 23/12/2022
10.33333/rp.vol50n3.03

generalmente sufren enormes daños en su credibilidad. Los informes de seguridad de Estadísticas y Tendencias en 2021 indicaron que en enero de ese año se registró el pico más alto a nivel histórico en la cantidad de sitios de phishing a nivel mundial, según reporte del *Anti-Phishing Working Group*.

Este tipo de ataque afecta principalmente al sector financiero, con un porcentaje del 24,9 % durante el primer trimestre, seguido por las redes sociales con el 23,6 % y los distribuidores de servicios de sitios web con el 19,6 %. Otro sitio que presentó un crecimiento acelerado en la cantidad correos electrónicos que registraron textos únicos en el campo asunto de los correos, con un total de 172.793 asuntos diferentes en un solo mes.

En la actualidad, los phishers han desarrollado un *ransomware* que ejecuta un código malicioso que afecta negativamente a los recursos informáticos y exige el pago de un rescate, para restaurar los recursos al estado original. La incidencia de correos con presencia de phishing es del 93 % según *Chief Security Officer*. El informe observó que la mayoría de las víctimas tiende a pagar rápidamente debido a la naturaleza sensible de sus recursos observar en Online report on phishing activities (2016).

Aunque cada día los sistemas se actualizan, los phishers buscan nuevas maneras de atacar y efectuar el robo a los usuarios. Varios de los sistemas de phishing desarrollados se enfocan principalmente en la revisión de URL, o la validación de protocolos de seguridad como https. Sin embargo, estas alternativas de control han sido abordadas por los phishers, obteniendo certificados válidos o cambiando continuamente las URL,lo que dificulta la detección a través de las herramientas de seguridad. Motivados por esta premisa, el objetivo de esta investigación es desarrollar un esquema anti-phishing apoyado en las características presentes en el Corpus del email, URL y Dominios, mediante la construcción de modelos de clasificación, utilizando la base de ajuste de parámetros en RF y NB con tres fuentes de datos para detectar phishing.

Como resultado de esta investigación, los principales aportes son:

1) Determinar características para clasificar entre sitios web de phishing, sospechosos y legítimos, obtenidos de los tres Dataset.

2) Identificar el vector de características final que ha ofrecido el mejor rendimiento en comparación con otros en el campo anti-phishing.

3) Identificar el modelo con mayor precisión mediante métricas usadas frecuentemente en la detección de phishing.

El documento también presenta la ventaja de la detectabilidad en la elección del conjunto de características para el corpus del email.

2. Marco Teórico

En esta sección, se describen las nociones teóricas necesarias para comprender la clasificación de URL utilizando métodos estadísticos para descubrir las propiedades léxicas, basadas en *host* de las URL de sitios web maliciosos, con el propósito de entender como clasificar la presencia de un ataque malicioso a gran escala para la construcción del modelo de predicción.

Una vez que los datos han sido entendidos, se puede establecer una idea referente al camino a tomar o sobre la técnica a emplear. Para predecir la probabilidad de ser víctima de robo de información o suplantación de identidad se ha dividido en dos secciones principales: metodología de investigación, análisis de características para detección de phishing, problemas de clasificación. Las referencias principales son Calva (2020), Orunsolu et al. (2019), Rosero (2020) y Hastie et al. (2017), que serán continuamente utilizadas en este trabajo.

2.1 Metodología de Investigación

La problemática que se tiene para esta investigación, toma en cuenta que el phishing es un conflicto social y en la actualidad se busca una solución eficiente. Para este trabajo, se utilizará la metodología CRISP-DM que es un modelo de proceso independiente para la minería de datos. Consta de seis fases iterativas que van desde la comprensión del problema hasta el desarrollo del escrito final, como indica Creswell (2015).

A continuación, se comienza con la investigación, es importante mencionar que para el análisis de características referente a la detección de phishing se realizó una revisión de literatura teniendo como principales fuentes Adebowale et al. (2018), Chin et al. (2018), Orunsolu et al. (2019) y Gansterer and Polz (2009), para posteriormente realizar una comparación de resultados de otras investigaciones relacionadas, esto se visualizará en la sección de resultados.

2.2 Análisis de características para detección de phishing

La definición de ataque phishing es un caso típico de clasificación binaria, ya que una comunicación en línea, por ejemplo: (correo electrónico, sitio web y chat electrónico) puede ser clasificada como: Phishing o benigna.

Tratándolo más formalmente, sea w una solicitud que necesita clasificación, es decir

$$w \underset{\rightarrow}{x} \{\text{Phish, benigna}\}. \tag{1}$$

Entonces x es el sistema anti-phishing que toma características, $f_i \in w$ donde

$$w_i = (f_1, f_2, ...f_i, ..., f_m),$$

Es decir, w_i es un vector no vació.
Por lo tanto, una solicitud contiene al menos una característica, por ejemplo: (enlaces, etiquetas HTML, scripts, certificado SSL, etc.) sobre la cual se puede consultar o clasificar la predicción de su estado. Debido a que estas características pueden variar de simples a complejas, el modelo propuesto utiliza una evaluación de frecuencia de características para la composición de vectores de características representada por $x = \{x_1, x_2, ..., x_n\}$ que asignan la etiqueta y a cada $f_i \in w$, de modo que la etiqueta y es una clase binaria representada como:

$$y = \begin{cases} 1, & \text{si es phishing,} \\ 0, & \text{si es benigna,} \end{cases} \quad (2)$$

Representado (2) como

$$x_i : f(w) \rightarrow y$$

La ecuación (1) describe el problema de clasificación donde, dado un dato de entrenamiento D, que contiene $(w_1, w_2, ..., w_n)$ y cada w_i contiene un conjunto de características $(f_1, f_2, ..., f_m)$. Además, los datos de entrenamiento son un conjunto de clases. $C = (c_1, c_2)$ que representa sitios legítimos y de suplantación de identidad que:

$$c_1 = \{w_i, f_i \mid w_i \in D, y = benigna, i = 1, ..., m\},$$
$$c_2 = \{w_i, f_i \mid w_i \in d, y = phishing, i = m+1, ..., p\}.$$

Por tanto, cada caso $w_i \in D$ se le puede dar una clase $c_i \in C$ y se representa como un par $(w_i, (c_i))$ dónde c_i es una clase de C asociada con el caso w_i en los datos de entrenamiento. Sea H el conjunto de clasificadores para $D \rightarrow C$, donde cada caso $c_i \in C$ se le da una clase y el objetivo es encontrar un clasificador $h_i \in H$ que maximiza la probabilidad de que $h(c_i) = C$ para cada caso de prueba. En el sistema propuesto, se eligen dos clasificadores de aprendizaje automático más comunes para la clasificación de phishing.

2.3 Módulo de selección de funciones

Un proceso de extracción de características implica la identificación de ciertas características en un conjunto particular de datos, por ejemplo: (phishing o benignas). Tales características generalmente se marcan como "huellas digitales", ocurren con poca o ninguna probabilidad, en la mayoría de los casos, estos rasgos suelen excluirse mutuamente de las otras. En este enfoque, se utiliza la evaluación de características basada en el análisis de frecuencia de varias características recopiladas de literatura existente. Esto se define como un módulo de selección de características (FSM) que consta de:

- Las características de la URL
- Las propiedades del documento web
- Las características del email

Estos tres componentes se consideran un filtro en FSM y cada factor se organiza en el enfoque para tener un sistema apoyado en componentes. Con base en esto, los tres filtros se construyen como un filtro unitario y uno compuesto para lograr gradualmente un planteamiento de detección eficiente.

2.4 Las características de la URL (filtro F1)

Las características de la URL representan las características asociadas con las direcciones web donde se puede recuperar una página en particular de Internet. Las características de la URL se extraen ya sea una URL absoluta o una URL relativa mediante el análisis de la estructura de enlaces en el DOM (Dominio). Para la extracción de identidad de URL, FSM considera **href** y **src** atributos de los enlaces de anclaje, en particular las etiquetas $< a >$, $< area >$, $< link >$, $< img >$ y $< script >$ del árbol de DOM, una página web donde normalmente se encuentran las direcciones web. Basándose en el estudio preliminar, se construyó el módulo de selección de características mediante una serie de consultas sobre ciertos rasgos de la URL seleccionadas de las investigaciones existentes (ver, Aburrous et al. (2010), Gowtham and Krishnamurthi (2014), Sonowal and Kuppusamy (2020) y Zouina and Outtaj (2017)).

Basado en la metodología de evaluación de características de frecuencia, se presenta el algoritmo 1.

Algorithm 1 Análisis de frecuencia de evaluación de características de URL

Require: Corpus de phishing actualizado, d_{ph}, URL, d_{be}, Valor umbral predefinido,θ.
Ensure: Vector de dimensión de características basado en URL S_m
Empezar
1. Para $i = 1$ hasta n:
2. $F_{URL(n)} \leftarrow$ el conjunto de todas las n funciones de URL.
3. Si $F_{url\_i} \in d_{ph}$ ó d_{be} Luego
4. $S \leftarrow$ nueva lista de funciones
5. Calcular la frecuencia de $F_{url\_i} \in d_{ph}.d_{be}$
6. Calcular la información de frecuencia, FI, de F_{url_i}
7. Si $FI_{F_{url\_i}} > \theta$, Luego
8. Adjuntar F_{url_i} a S
9. **Caso Contrario**
Rechazar F_{url_i}
10. $i = i + 1$
Continuar
11. Rango $F_{url\_i} \in S$
12. Seleccionar la parte superior F_{url_i} características $\in S$
13. Obtener una medida de desempeño de S
14. Identificar la mejor medida de desempeño como las mejores características
15. $S_m \leftarrow$ mejores características
16. Fin

La Tabla 6 presenta el significado de las notaciones utilizadas en el 1, esta se encuentra disponible en el Apéndice A.

2.5 Las propiedades del documento web (filtro F2)

Las propiedades del documento web de una página web se extraen de la etiqueta de documento. Este proceso de extracción se basa en el concepto de método Término-Frecuencia de un Documento de Frecuencia Inversa (TF-IDF). El método se utiliza para extraer un conjunto de palabras clave del documento d, que se recopila de varias partes de una página web. El TF-IDF refleja la estadística numérica de cuán relevante es una característica para un documento en un corpus de datos. El valor de TF-IDF aumenta proporcionalmente al número de veces que aparece una característica en el documento, pero se compensa con la frecuencia de la característica en el corpus. Por tanto, un término particular definido como t tiene un peso TF-IDF alto si el término tiene una frecuencia alta en un documento D dado y una frecuencia baja si el término es relativamente poco común.

Dado un documento dy su conjunto de identidad de términos t, el FSM usa la medición de la tasa de frecuencia para determinar la inclusión de una característica en la clase discriminatoria. A continuación, se presenta el Algoritmo de Análisis de frecuencia de evaluación de características.

Algorithm 2 Análisis de frecuencia de evaluación de características

Require: Tamaño de datos(p), conjunto de características original(n), umbral(θ), clase(C).
Ensure: Subconjunto de características principales de dimensión $m(fs)$
Empezar
1. Para $i = 1$ hasta n:
2. Para $j = 1$ hasta p:
3. a=1
4. Seleccioanr $s_i \in H_P$
5. Calcular CFS usando {
6. $s(f_i, f_2, ..., f_n, c)$ como entrada
7. Para $i = 1$ hasta n:
8. Inicializar el factor de correlación apropiado, t
9. $r = calcular\_correlacion(f, c)$
10. Si $(t > r)$ luego
11. Adjuntar $f_i \in s_n$ en m
12. Fin
13. Ordenar m en valor descendente de t
14. Quitar f con rango inferior
15. Devolver f predominante como $f(s)$
16. Fin}
17. Si $(f_i(s) > \theta)$ luego agregar a $m(fs)$
18. $c = c + 1$
19. Si $(a \leq n)$ volver a 3
20. Fin para
21. Fin para
22. Conjunto de características de retorno $m(fs)$
23. Fin

El algoritmo 2 presenta el flujo de la metodología del sistema, y las notaciones para este algoritmo se encuentran la Tabla 6.

2.6 Las propiedades del corpus del email (filtro F3)

A veces, hay casos en los que el filtrado de *spam* no tiene éxito en evitar que el *malware* de productos básicos u otros correos electrónicos no solicitados lleguen.

Es por esta razón que, para este filtro se analizará el contenido de los correos tanto en los que presentan phishing como los emails reales mediante la técnica de Text Mining con el propósito de detectar patrones adicionales a los que se conocen por defecto que son:

- El remitente no corresponde con el servicio que envía el correo
- La gramática con fallos
- URL_s falsas camufladas en hipervínculos
- Archivos adjuntos que no son lo que parecen
- Correo de un servicio no utilizado o no contratado

2.7 Problemas de clasificación y su modelamiento

Para este trabajo de investigación, se trabajará con los problemas de clasificación, frecuentemente se divide la base en dos conjuntos de datos que son: entrenamiento/training y prueba/test, siguiendo normalmente el criterio de Pareto de 80 % y 20 %.

2.8 Random Forests

El Bagging o bootstrap es una técnica para procedimientos de varianza grandes y sesgos pequeños, como lo son los árboles que son más simples de entrenar y ajustar. La idea esencial del bagging es promediar muchos modelos ruidosos, pero aproximadamente insesgados y, por lo tanto, reducir la varianza. Los árboles son candidatos ideales para el bagging, ya que pueden capturar interacciones complejas.

Dado que los árboles son notoriamente ruidosos, se beneficia enormemente el promedio. Además, dado que cada árbol generado en el bagging se distribuye de manera idéntica $(i.d.)$, la expectativa del promedio de B árboles es lo mismo que la expectativa de cualquiera de ellos. Esto contrasta con el impulso, donde los árboles se generan de forma adaptativa para eliminar el sesgo y, por lo tanto, no son $i.d.$

2.9 Naïve Bayes

El NB es un algoritmo de clasificación de texto simple y efectivo que usa las probabilidades conjuntas de palabras y categorías para estimar las probabilidades de categorías dado un documento como se menciona en Anwar et al. (2017). El supuesto de independencia condicional se puede expresar formalmente como:

$$P(A \mid C = c) = \prod_{i=1}^{n} P(A_i \mid C = c),$$

donde cada conjunto de atributos o conjunto de características $A = \{A_1, A_2, ..., A_n\}$ consta de n valores de atributo. Con el supuesto de independencia condicional, en lugar de calcular la probabilidad condicional de clase para cada agrupación de A, solo estime la probabilidad condicional de cada A_i, dado C. Para clasificar una muestra de prueba, el clasificador NB calcula la probabilidad posterior para cada clase C como:

$$P(C \mid A) = \frac{P(C) \prod_{i=1}^{n} P(A_i \mid C)}{P(A)}. \quad (3)$$

La ecuación (3) indica que al observar el valor de una característica particular, A_i, la probabilidad previa de una categoría particular, C_i, $P(C_i)$ se puede convertir a la probabilidad posterior, $P(C_i \mid A_i)$, que representa la probabilidad de una característica en particular, A_i siendo una categoría particular, C_i .

3. Marco Metodológico

En esta sección, se centrará en la segunda etapa de la metodología CRISP-DM que se basa en la recopilación de datos para poder abordar el problema, como siguiente paso el análisis de mismos para llegar a verificar la calidad de los datos, en la siguiente etapa se procede a la construcción del vector de características esenciales para la generación del modelo. Como fuente principal para esta sección, véase, Rosero (2020), Martínez (2018) y Zhao et al. (2019).

Para este proceso se ha utilizado el método *Feature Selection*, que permite preprocesar las características de las URL_S y correos electrónicos, teniendo una peculiaridad que es imputar los innecesarios sin correr el riesgo de pérdida de información. Además, se utilizó RF y NB para la construcción del vector de aprendizaje automático. A continuación, en la Figura 1 se ilustra el diagrama de flujo del modelo a realizar.

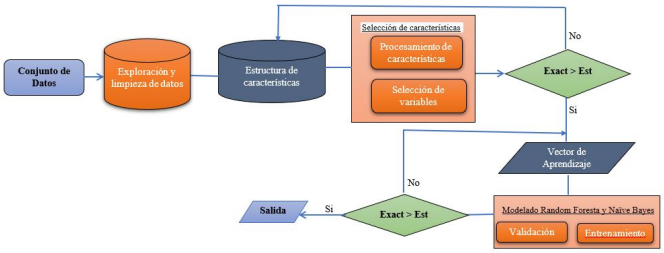

Figura 1. Diagrama de flujo del modelo de detección y mitigación de phishing

Es de vital importancia mencionar que para este proyecto se seleccionó los lenguajes de programación Java para la construcción del vector de características y el software Python para la implementación de los modelos predictivos, se escogió estos lenguajes porque proporcionan librerías eficientes en la manipulación y tratamiento de este tipo de registros.

En la siguiente sección, se va a describir los datos descargables con que se trabajará para el proyecto de investigación.

3.1 Descripción del conjunto de datos

Se utilizan tres conjuntos de datos que se encuentran disponibles al público, estas bases contienen URL malignas y reales, correos con presencia de phishing y correos benignos para evaluar el rendimiento de la arquitectura de detección de phishing propuesta. Los conjuntos experimentales se obtienen de un corpus de datos extraídos de Monkey (2020), Phishtank (2021) y William and Cohen (2019).

El número de datos recolectados en cada conjunto es: 14200 en PhishTank, 3700 en Monkey y para Enron se tiene 16800, dándonos un total general de 31084; estos datos posteriormente se los tratará para obtener una data limpia lista para realizar análisis y visualizar algunos resultados.

Ahora bien, estas tres fuentes de datos se eligen porque contienen conjuntos de datos verificados que se utilizan en la mayoría de las investigaciones anti-phishing para comparar los resultados de la evaluación.

3.2 Tratamiento de datos

Una vez identificados los datos, se sigue con la exploración y análisis de datos, que se lleva a cabo con el uso de técnicas estadísticas, que nos dirán si los datos extraídos son válidos. La etapa de exploración se hizo a través de búsquedas básicas de los Dataset con extensión M.box y CSV.

Los documentos descargables contienen las URL_s y correos electrónicos con Phishing y sin Phishing, por tal razón no requiere de una discriminación entonces se lleva a cabo un proceso aparte, debido a que, aunque se cuenta con Datasets completamente diferentes en sus categorías, y sus variables son parecidas. Se eligen estos datos para hacer el análisis.

1). header['From']

2). header['Subject']

3). header ['Content-Type']

4). header['Date']

5). header['Body']

Siguiendo con este proceso se limpia los datos únicamente para las variables seleccionadas, lo primero que se va a corregir es el porcentaje de valores nulos. Es así que se tiene los siguientes resultados, para PhishTank se tiene 5,9 %, Monkey con 10,8 % y Enron tiene 10 % de valores nulos, estos valores son aceptables y no necesitan un tratamiento especial.

Con los datos que se obtuvieron de este análisis, se podrá determinar que dominios predominan en la falsificación de identidad. Seguidamente, se visualizan estos resultados de las principales fuentes utilizadas para este estudio que son: PhishTank y Monkey.

Se observa en la Tabla 1 que el dominio Bank presenta la frecuencia más alta en el contexto de ser víctima de Phishing para la data de PhishTank con un porcentaje de ocurrencia del 18 %, le sigue Gmail con el 16 %. Ahora bien, para la base Monkey el dominio predominante es Gmail con una representatividad del 45 % y esto se sobreentiende dado que esta data en particular contiene correos infectados. De manera general los phishers apuntan a tener una buena ganancia con el menor tiempo de ejecución.

Posteriormente, se procede a identificar las palabras más usadas tanto en los emails con presencia de phishing y los reales. Mediante el método de exploración de texto (Text Mining) en las bases, esta técnica ayudó a identificar las palabras que presentan una alta frecuencia en los correos electrónicos con presencia de Phishing o benignas, Una vez obtenido el conjunto de palabras, se las tomará como variables en el modelo de detección. Ahora se observarán las palabras identificadas, la frecuencia absoluta y la frecuencia relativa en los emails, lo que permite observar la diferencia para la detección con más claridad.

Nota: Se utilizó esta técnica dado que se puede explorar y descubrir relaciones ocultas dentro de datos no estructurados. Dado que

Tabla 1. Dominios detectados como los más utilizados en la suplantación de identidad (Phishing) en PhishTank

Dominios	PhishTank	R.PhisT	Monkey	R.Monkey
Bank	1300	0,1831	227	0,0289
Gmail	1200	0,1690	3512	0,4472
Run escape	1039	0,1463	0	0,0000
Google	620	0,0873	231	0,0294
PayPal	609	0,0858	1020	0,1299
eBay	388	0,0546	1	0,0001
Facebook	345	0,0486	25	0,0032
office	259	0,0365	138	0,0176
Microsoft	217	0,0306	124	0,0158
Halifax	124	0,0175	0	0,0000
Amazon	119	0,0168	24	0,0031
Android	99	0,0139	393	0,0500
Apple	99	0,0139	695	0,0885
Netflix	93	0,0131	100	0,0127
Adobe	84	0,0118	13	0,0017
WhatsApp	75	0,0106	1	0,0001
YouTube	60	0,0084	54	0,0069
Steam	57	0,0080	0	0,0000
Yahoo!	55	0,0077	993	0,1264
Outlook	52	0,0073	61	0,0078
LinkedIn	40	0,0056	0	0,0000
Instagram	38	0,0054	9	0,0011
Virus total	34	0,0048	0	0,0000
Twitter	29	0,0041	17	0,0022
JPM Chase and Co.	22	0,0031	0	0,0000
Hotmail	12	0,0017	56	0,0071
American express	10	0,0014	100	0,0127
Vodafone	9	0,0013	0	0,0000
HSBCgruop	7	0,0010	0	0,0000
Windows	6	0,0008	60	0,0076

Tabla 2. Frecuencia absoluta de las palabras que se determinaron como características para detección de Phishing.

Palabra	Phishing	F.Rphis	No-Phishing	F.R.Nphis
Actualizar	233	0,006	746	0,073
Confirmar	121	0,003	197	0,019
Usuario	244	0,006	220	0,021
Cliente	45	0,001	245	0,024
Querido	112	0,003	65	0,006
Miembro	44	0,001	258	0,025
Restringir	365	0,009	257	0,025
Sostener	120	0,003	871	0,085
Verificar	242	0,006	18	0,002
Cuenta	402	0,010	179	0,017
Notificación	343	0,009	80	0,008
Login	236	0,006	12	0,001
Sesión	500	0,012	81	0,008
Clic aquí	1200	0,030	194	0,019
Contraseña	930	0,023	81	0,008
Felicidades	700	0,017	38	0,004
Felicitaciones	523	0,013	22	0,0012
Ganaste	189	0,005	0	0,000
Gratis	1200	0,030	934	0,091
Seguridad	118	0,003	40	0,004
Importante	1500	0,037	125	0,012
Aviso	47	0,001	135	0,013
Crédito	5800	0,144	833	0,081
Banco	14500	0,360	219	0,021
En línea	124	0,003	818	0,080
Enviar	448	0,011	593	0,058
Transferir	945	0,023	2057	0,201
Acceso	743	0,018	369	0,036
Contagio	38	0,001	3	0,000
Brote	638	0,016	95	0,0084
Epidemia	735	0,021	1057	0,102
Suspender	68	0,002	26	0,003
Tarjeta	6054	0,150	278	0,027
Sospechosa	435	0,011	89	0,009
Financiera	657	0,016	56	0,005
Actividad	767	0,021	99	0,008
Vulnerable	57	0,005	36	0,004
Pandemia	965	0,036	209	0,011
Vacuna	1260	0,053	509	0,015
Covid-19	25750	0,597	7576	0,191

el 80 % de los datos en el mundo reside en un formato no estructurado, la minería de texto es una práctica extremadamente valiosa dentro de este tipo de procesos.
Con los resultados obtenidos, se observó la influencia del uso de estas palabras en los correos falsos y verídicos. A continuación, se visualiza una gráfica de las frecuencias obtenidas.

Figura 2. Número de términos usados en los emails con Phishing y sin Phishing

3.3 Verificación la calidad de los datos

Luego de realizar el análisis de datos, se concluye que los Datasets obtenidos son adecuados, dado que nos brindan la información requerida para la extracción de características, las mismas que contribuyen para la modelación eficiente, pretendiendo obtener una buena precisión. Con respecto a los correos electrónicos, provienen de una fuente fidedigna, hay que notar que los registros están actualizados, por lo que se puede asegurar que los resultados son un buen punto de referencia para nuevos estudios.

El conjunto de datos que contiene correos no infectados, ayuda a balancear los datos y con ello obtener un mejor resultado en la predicción del modelo de aprendizaje, dando un mejor pronóstico en la identificación de correos electrónicos infectados, debido a que nos brinda información para determinar las características que ayudan a determinar entre un correo infectado y otro no.

PhishTank es una fuente confiable mencionada en varios artículos referentes a este tema en especial Orunsolu et al. (2019), la información proporcionada nos ayuda en demasía en el análisis de URL_s y de dominios. Los registros recopilados, brindan datos muy relevantes y precisos, lo que garantiza la ausencia de riesgo de tener un porcentaje alto de ruido en el procesamiento, y que los resultados serán veraces.

Preparar los datos para ajustarlos de forma adecuada para el proceso de modelamiento, es algo principal y frecuentemente conlleva más tiempo, debido a que es necesario agrupar y elegir de manera correcta las variables que van a ser procesadas.

4. Construcción del vector de características

Para el desarrollo de esta parte del proyecto, se seleccionó el lenguaje de programación Java 11, dado que se tiene datos no estructurados en formato M.box y se procesan para determinar las características para la detección de Phishing, para lo cual se utiliza librerías para parsear el HTML y extraer los datos necesarios. Con esto se exporta la base de mensajes completa del formato JSON a CSV.

4.1 Selección de características

Como se expuso al inicio de esta sección para el diseño del modelo, hay que seleccionar las características apropiadas, basándonos en estudios referentes a la temática y el análisis que se realizó en el capítulo anterior. Además, se utiliza una biblioteca Java llamada Secure Socket Extension para extraer información de terceros relacionada con un dominio en particular durante el proceso de extracción de características. Esto proporciona una forma eficaz de examinar todas las características y etiquetas relevantes de la página analizada para examinar su estado.

Para las funciones de URL, se extraen 7 funciones. En las características del documento web, se extraen otras 17 características, que se extraen para mejorar la detección de phishing en los diversos corpus del email. Aunque algunas otras características todavía están disponibles, se elige especialmente esas características porque las que se omiten se puede deducir de las elegidas (por ejemplo, el número de puntos en la mayoría de las URL) de phishing está asociado con nombres de dominio alargados.

Se designa un nombre a las características extraídas. La fase del clasificador de aprendizaje automático utilizó las características elegidas para entrenar los algoritmos (RF y NB) de aprendizaje automático y seguir con el proceso de validación.

4.2 Identificación del ataque Phishing en emails

Para el entendimiento sobre los patrones que tiene un correo electrónico con presencia de Phishing o a su vez un correo legítimo, se realizó una revisión exhaustiva de documentación relacionada con el tema de estudio, con esto se procede a realizar una comparación de palabras entre correos electrónicos benignos (reales) e infectados mediante la técnica de text mining.

Este proceso permitió la correcta caracterización de las palabras que predominan en email con legítimo o con Phishing. La formación de características, basadas en los términos textuales son unidas y esquematizadas de tal forma que cada palabra tiene homogeneidad entre ellas y así formar un grupo de estas. Se debe mencionar que esta etapa es una parte primordial para la construcción del Dataset que posteriormente se usará para el modelamiento y la validación. Se conformaron 9 grupos de palabras para este proceso.

La variable dependiente y representa la identificación de correos mediante ciertas características, por tanto, será una variable binaria que toma los valores de 1 para los correos con etiqueta de Phishing y 0 para los correos con etiqueta de benignos: de modo que la etiqueta y es una clase binaria representada como:

$$y = \begin{cases} 1, & \text{si es phishing,} \\ 0, & \text{si es benigna.} \end{cases}$$

5. Construcción del modelo Predictivo

En esta sección, se describe el modelo utilizado para el cumplimiento del objetivo planteado, por lo que se utilizará las métricas que indicarán el porcentaje de precisión de los algoritmos de clasificación, en la siguiente subsección, se evalúa estos modelos y se observará si se cumple con los criterios establecidos para los datos de prueba y locales.

5.1 Implementación del Algoritmo Predictivo.

Haciendo uso del software Python explicado a profundidad en una sección anterior, en esta implementación se utilizarán las siguientes librerías para la construcción del Algoritmo Predictivo: *sklearn.metrics*, *matplotlib.pyplot* y *numpy*.
A continuación, se muestran los detalles y observaciones del algoritmo.

En el software Python, se carga el paquete que contiene librerías, de NB, Árboles de decisión y RF son los que se utilizarán para la modelización de esta problemática que tiene como punto fundamental predecir si un correo tiene presencia o no de este ataque.

En la elaboración del modelo, se toma como entrada el documento que contiene la matriz de 0_s y 1_s, estos datos se dividirán mediante un muestreo aleatorio simple, en dos submuestras aleatorias, una para el desarrollo del modelo (train) y la otra para su validación (testeo).

La muestra de modelamiento corresponde aproximadamente al 80 % de la muestra original. En cambio, la muestra de validación corresponde aproximadamente al 20 % de la muestra original, siguiendo el principio de Pareto. Teniendo como objetivo principal detectar la presencia de phishing, las entradas que se consideran para los algoritmos de clasificación son las que se muestran a continuación.

- train_inputs: Variable que abarca el 80 % de los datos para la caracterización de un email con presencia o no de Phishing.
- test_inputs: Variable que abarca el 80 % de los datos para la caracterización de un email con presencia o no de Phishing
- train_outputs: Variable que abarca el 20 % de los datos para la caracterización de un email con presencia o no de Phishing.
- test_outputs: Variable que abarca el 20 % de los datos para la caracterización de un email con presencia o no de Phishing.

En el problema de la detección de phishing, el uso de RF y NB es común en el caso de tener vectores de características con diferentes volúmenes de dimensionalidad de datos (Moghimi and Varjani , 2016). En la literatura, el análisis de frecuencia de diferentes clasificadores indica una alta adopción de estos dos clasificadores, especialmente en la definición de problemas de phishing debido a su simplicidad y alta precisión(Anwar et al., 2017; Dhanalakshmi and Chellappan, 2013). Motivado por las investigaciones anteriores de RF y NB sobre conjuntos de datos de phishing, nuestro método de clasificación emplea estos dos clasificadores en el mismo conjunto de funciones para evaluar su rendimiento.

5.2 Modelo Random Forests

Como otros clasificadores, los clasificadores de bosque deben estar equipados con dos matrices: una matriz X de forma dispersa o densa que contiene las muestras de entrenamiento, y una matriz Y de forma que contiene los valores objetivo, para este modelo se utilizará la librería RandomForestClassifiery, donde cada árbol en el conjunto se crea a partir de una proporción extraída con reemplazo del conjunto de entrenamiento. Sin embargo, RandomForestRegressor usa un número predeterminado de árboles de 100, que normalmente no es suficiente. En consecuencia, esta se subió hasta 1.000 árboles en primera instancia para posteriormente dejarlo en 3.000 árboles. La profundidad predeterminada de cada árbol (max_depth) es 5, lo que significa que se ensambla árboles con profundidad máxima de 5.

Para determinar los parámetros óptimos, primero se ha ejecutado la forma predeterminada, es decir sin cambiar lo que por defecto está determinado para observar los resultados que se obtienen, posteriormente se va jugando con los parámetros para encontrar los que hacen mínima la tasa de mal clasificados.

El propósito de usar RF es que se consigue una variación pequeña al combinar varios árboles, ocasionalmente a costa de un incremento pequeño en el sesgo. En la práctica, la reducción de la varianza es a menudo significativa, por lo que se obtiene un mejor modelo.

5.3 Modelo Naive Bayes

El módulo sklearn. ensemble tiene la librería MultinomialNB que se usará para la implementación del algoritmo de Bayes para datos distribuidos multinominalmente, siendo estas las dos variaciones clásicas de NB, para el uso en la clasificación de texto en el que los datos son recuentos de vectores de palabras normalmente, aunque también se sabe que los vectores tf-idf funcionan bien. La distribución está parametrizada por vectores $\theta_y = (\theta_{y_1}, ..., \theta_{y_n})$ para cada clase y, donde n es el número de características (en la clasificación del texto, el tamaño de la palabra) y θ_{y_i} es la probabilidad $P(x_i \mid y)$ de característica i apareciendo en una porción perteneciente a la clase y.

Para esta implementación no fue necesario utilizar (partial_fit) dado que el conjunto de entrenamiento completo no presentó ningún inconveniente en la memoria.

De forma general, el tiempo de ejecución de este modelamiento depende de los datos que se utiliza por ejemplo, para el conjunto de datos descargables en la construcción de características tuvo una hora y media de procesamiento mientras que para el entrenamiento se demoró alrededor de una hora, ahora bien, para los datos locales (cuentas personales) el tiempo varió sustancialmente, dado que era necesario construir un programa en JAVA 11 para indexar los mensajes de las cuentas, por tal razón, se requirió de cuatro horas para completar el proceso. Hay que mencionar que este periodo de tiempo varía dependiendo del equipo que se utilice en la modelización.

5.4 Evaluación el modelo

En la fase de evaluación, se compara el rendimiento del sistema propuesto a través de un experimento de validación cruzada de 10 veces antes del proceso de evaluación en el conjunto de datos de prueba. Esto implica la división aleatoria del conjunto de datos de prueba en diez submuestras iguales, de las cuales una sola submuestra se usa para la validación final del modelo, mientras que las otras submuestras son usadas para el entrenamiento del sistema. Por tanto, el modelo predictivo propuesto se basó en el 80 % del conjunto de datos y se validó en el 20 % restante.

Las razones para usar la validación cruzada en este modelo son para:

i. i. Verificar el comportamiento del error del modelo predictivo, en este caso, los errores asociados con el modelo predictivo de RF y NB en la detección de phishing.

ii. Validar el Dataset de entrenamiento mediante la validación de cada subconjunto. Esto es para tener un nivel alto de confianza en el modelo entrenado.

En el experimento, se usará tanto conjuntos de datos de prueba como los locales, estas datas contienen sitios web legítimos y de phishing no superpuestos que se procesara previamente para posteriormente obtener un archivo CVS, que contiene el vector de características finales para procesar en el modelo predictivo. A continuación, se explica cómo se recolectaron los datos locales en forma resumida sin dejar de lado lo primordial.

Dado que el enfoque de esta investigación es a nivel local del país, los dataset disponibles en la red pública son demasiado antiguos o en otro idioma, por tal motivo se opta por utilizar los correos de 3 cuentas, una cuenta de Gmail personal, Outlook personal y una institución educativa nacional.

- Se implementa un programa en JAVA 11 que permite ejecutar un cron que se encargará de indexar todos los emails de las cuentas mencionadas para obtener tanto correos categorizados como verdaderos, así como de la carpeta de spam mediante la lectura directa del buzón de mensajes con el uso del protocolo POP3.

- Esos registros se guardan en un Dataset no estructurado (mongodb) en formato JSON y se procesan para determinar las características necesarias para la ejecución, para lo cual se utiliza librerías como JSOUP para parsear el HTML y extraer los datos necesarios.

Con esto se exporta la base de mensajes completa del formato JSON a CSV, las librerías que se utilizaron para este proceso están descritas en la sección “Construcción de características"; En el Servicio "MensajeSrv"se encuentra la lógica descrita en Indexado de mensajes.

Posteriormente, se realiza el análisis respectivo, para esto se procede a dividir el proceso en tres etapas, la primera el entrenado del modelo, como siguiente fase se tiene la predicción, y para finalizar los resultados conseguidos al ejecutar el modelo en un entorno moderado para la detección de phishing en los emails en

las cuentas personales antes mencionadas.

A continuación, se detalla los resultados y el análisis del modelo.

Etapa uno: (Entrenado del modelo) Para esta etapa se tiene alrededor de 7043 correos para el entrenamiento del modelo, teniendo 3620 con presencia de phishing y 3423 correos reales.

Nota: Estos datos representan el 80 % del dataset obtenido.

Hay que mencionar, antes de seguir con la segunda etapa, que en la data de testeo se tiene 1761 emails teniendo 920 emails con phishing y 841 emails sin presencia de phishing, estos datos se utilizarán para la predicción, teniendo un total de 8804 correos en la base original.

Segunda etapa: (Predicción del modelo) Para la fase de predicción, se trabajará con 1629 emails para esta etapa, se hará uso de los modelos de clasificación RF y NB. El rendimiento del sistema propuesto se evalúa mediante el uso de cinco parámetros estándar que consisten en Accuracy, Precisión, Recall_score y Puntación F1. Estas son las métricas de rendimiento estándar para evaluar cualquier sistema de detección de phishing para la evaluación de los resultados.

Adicionalmente, el coeficiente de correlación de Mathew (MCC) nos permite determinar el poder predictivo del modelo de aprendizaje automático en el experimento de validación cruzada.
Estos consisten en:

- Verdadero Positivo (TP): que señala el número de correos identificados como phishing, siendo estos phishing.
- Verdadero Negativo (TN): son los señalados como benigno, siendo estos phishing.
- Falsos Positivos (FP): son los señalados como phishing, siendo estos benignos/reales.
- Falsos Negativos (FN): son señalados como benignos, siendo estos benignos.

Para el cálculo del MCC y determinar la calidad del modelo de predicción, al aproximarse MCC a la unidad, indica que el sistema tiene una predicción casi perfecta y, por lo tanto, es un sistema de detección confiable. La siguiente ecuación representa el MCC.

Donde:

TPRxTNR-FPRxFNR $= Factor_1$
y
(TPR+FPR)(TPR+FNR)(TNR+FPR)(TNR+FNR) $= Factor_2$

$$MCC = \frac{Factor\_1}{\sqrt{Factor\_2}}$$

En el siguiente capitulo, se presentan los resultados para los dos conjuntos de datos, adicionalmente se presenta la comparación con otros trabajos que siguen la misma línea de investigación y de esta manera observar el nivel de predictibilidad del modelo.

6. Resultados y discusión

El rendimiento del sistema propuesto se evalúa utilizando cinco parámetros estándar y la validación cruzada, esto se menciona en el anterior capítulo. Para el proceso de VC se repitió 10 veces y después de la validación, se calcula una sola estimación. Esta estimación es el promedio de las diez iteraciones.

El incentivo para aplicar el experimento de validación cruzada es ajustar el rendimiento de un modelo fuera del conjunto de entrenamiento, a continuación, se analizarán los resultados de los dos conjuntos de datos.

6.1 Resultado para los datos descargables

Tabla 3. Tabla de resultados para el conjunto de datos Locales (Cuentas personales: Gmail, Outlook e Institucional)

Métricas \Modelo	RF	NB	Trees
Accuracy	97,70 %	92,53 %	94,50 %
Precisión	97,24 %	90,57 %	94,30 %
Recall_score	98,55 %	93,06 %	95,70 %
Puntuación F1	97,54 %	92,06 %	95,92 %
Roc_auc	97,50 %	93,50 %	94,60 %
Coef. Mathew	0,971	0,923	0,959

Se procede a interpretar los datos de las métricas obtenidas, se visualiza que el mejor modelo es Random Forests a comparación de NB como un plus se calculó para árboles de decisión, teniendo como resultado que RF presenta el que mejor nivel de predicción para este investigación, ahora bien se analiza las métricas obtenidas, teniendo la exactitud que representa el porcentaje en el cual el modelo ha acertado, obteniendo un valor de 97,70 % de exactitud, el valor obtenido para la precisión es de un 97,24 %. Dado que se tiene un conjunto relativamente balanceado se puede decir que la métrica Accuracy proporciona un resultado confiable para nuestro trabajo.

En el caso de la métrica de Recall es la capacidad del modelo para detectar los casos significativos. En nuestro caso, se tiene 98,55 % que es claramente un valor muy bueno para una métrica. Se puede afirmar que nuestro algoritmo de clasificación es muy sensible.

Para la Puntuación F1, se observa un porcentaje de 97,54 % dado que nos esquematiza la precisión y sensibilidad en una sola métrica, dándonos una gran ayuda cuando la distribución es desigual, así al tener una alta precisión y alto recall, implica que, el modelo elegido maneja perfectamente esa clase.

En el caso del coeficiente de correlación de Mathew, se tiene 0,971 con una buena calidad para el modelo de predicción RF.

En la Figura 3, se presenta la curva ROC para los datos de testeo, se procede a interpretar. Dado que el valor es cercano a uno, se puede decir que el rendimiento del modelo es bastante bueno. Así, se ha encontrado un clasificador con un rendimiento muy bueno sin tener el riesgo de sobreajuste en los datos de las bases: PhishTank, Monkey y Enron.

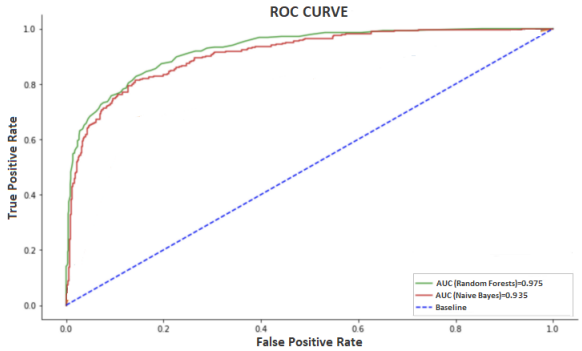

Figura 3. Curva Roc y AUC de datos descargables

Ahora se procede a calcular el valor real del error y el porcentaje de error que tiene el modelo. La ecuación 4 determina el valor real (ERR) y en la ecuación 5, la tasa (TERR), es en otras palabras, el porcentaje que el modelo no etiquetó los correos correctamente.

$$ERR = \frac{\text{FN+FP}}{\text{TN+FN+FP+TP}}, \tag{4}$$

donde, la tasa de error (TERR) se calcula de la siguiente manera:

$$TERR = \frac{\text{FN+FP}}{\text{TN+FN+FP+TP}} \cdot 100, \tag{5}$$

$$= 2,3\,\%.$$

Finalmente, de los resultados que se obtuvieron, se observa que, de los 6217 correos analizados, el 97,7 % que corresponde a 6074 emails fueron clasificados correctamente, por el contrario, el 2,3 % correspondiente a 143 correos fueron etiquetados de forma incorrecta. Por consiguiente, se puede decir que el modelo presenta una predictibilidad alta, teniendo como consecuencia resultados fiables en la clasificación de phishing.

6.2 Resultado para los datos Locales

En esta sección, se analiza los resultados para los datos extraídos de las tres cuentas personales. A continuación, se presenta la Tabla 4 que son las métricas obtenidas en el modelo con este conjunto de datos.

Tabla 4. Tabla de resultados para el conjunto de datos Locales (Cuentas personales: Gmail, Outlook e Institucional)

Métricas\Modelo	RF	NB	Trees
Accuracy	95,40 %	89,35 %	91,50 %
Precisión	92,14 %	85,98 %	89,03 %
Recall_score	96,55 %	89,62 %	92,70 %
Puntuación F1	94,74 %	87,08 %	91,22 %
Roc_auc	94,90 %	90,20 %	91,26 %
Coef. Mathew	0,967	0,932	0,954

De la misma forma que se hizo en la sección anterior para los registros descargables, se procede a interpretar los datos de la Tabla 4 obtenidos, se puede visualizar que el mejor modelo es RF al igual que en el caso anterior, ahora bien, se analiza las métricas obtenidas, teniendo la exactitud que representa el porcentaje de predicciones correctas frente al total por tanto se tiene 95,40 % de exactitud para este modelo, el valor obtenido para la precisión es de un 92,14 %. Por tanto, nuestro modelo es más preciso que exacto, coincidiendo con el experimento anterior.

En el caso de la métrica de Recall (Sensibilidad) es la habilidad del modelo para detectar los casos relevantes. En nuestro caso, se tiene 96,55 % es claramente un valor bueno para una métrica. Se puede decir que nuestro algoritmo de clasificación es sensible.

Para la Puntuación F1, se observa un porcentaje de 94,74 %, dado que nos esquematiza la precisión y sensibilidad en una sola métrica, dándonos una gran ayuda cuando la distribución es desigual, así al tener una alta precisión y alto recall, implica que, el modelo elegido maneja perfectamente esa clase.

En el caso del coeficiente de correlación de Mathew, se tiene 0,967 obteniendo una buena calidad para el modelo de predicción RF. De la misma manera que se visualizó en los datos experimentales, en la Figura 4 se presenta la curva ROC para los datos de testeo, se procede a interpretar. Dado que el valor es cercano a uno, se puede decir que el rendimiento del modelo es bastante bueno. Así, se encontró un clasificador con un rendimiento muy bueno sin tener el riesgo de sobreajuste en los datos locales.

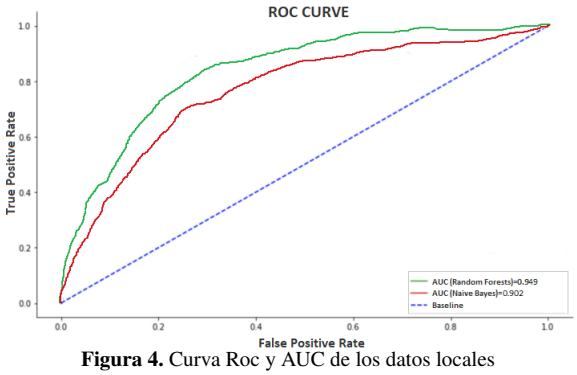

Figura 4. Curva Roc y AUC de los datos locales

Al igual que la sección anterior, se determina el porcentaje de error presente en el modelo, y esto nos proporciona una validez buena para el análisis de resultados.

Se procede a calcular el valor real del error y el porcentaje de error que tiene el modelo. La ecuación 4 determina el valor real (ERR) y en la ecuación 5 la tasa (TERR), es en otras palabras, el porcentaje que el modelo no etiquetó los correos de correctamente.

$$ERRL = \frac{\text{FN+FP}}{\text{TN+FN+FP+TP}}, \tag{6}$$

donde, la tasa de error (TERRL) se calcula de la siguiente manera:

$$TERRL = \frac{\text{FN+FP}}{\text{TN+FN+FP+TP}} \cdot 100, \tag{7}$$

$$TERRL = 4,6\,\%.$$

De los resultados alcanzados, se observa que, de 1.761 correos analizados, el 95,4 % que corresponde a 1680 correos se clasificó de forma correcta, mientras que el 4.6 % correspondiente a 81 correos fueron clasificados de forma errada. Por tal motivo, se determina que el modelo tiene un porcentaje de predicción relativamente alto, dando como consecuencia resultados fiables en la clasificación de phishing.

7. Comparación de resultados obtenidos

Resumiendo, de los resultados obtenidos en los dos conjuntos de datos se observa que, el modelo construido logró un 97.7 % en la detección de phishing en URL_s y en los correos electrónicos para los datos experimentales como son: PhishTank, Monkey y Enron. Teniendo como observación principal que al variar el conjunto de características se puede obtener un mayor o menor porcentaje de predicción, pero esto se debe a la actualización de técnicas de ataque en este tipo de delito, el vector de características debe actualizándose.

Obteniendo que la aplicación de un modelo ayuda a la detección de correos electrónicos con presencia de phishing. Pese a esto también se debe mencionar que la posibilidad de tener el 2,3 % de error al momento de la clasificación, aunque es pequeño no deja de ser un dato que hay que tomar en cuenta para futuros trabajos.

Para los resultados obtenidos en el conjunto de datos locales se visualiza que, en diversas métricas aplicadas, el modelo creado logró un 95.4 % en la determinación de existencia de phishing en URL_s y en los correos electrónicos. Los porcentajes tienen una gran diferencia entre los datos experimentales y los locales, dado que los datos que se tiene en nuestro país no son los mejores o como se menciona anteriormente la detección de delitos cibernéticos en Ecuador está comenzando, es por esta razón que los datos no presentan ciertas características para obtener un porcentaje mayor.

La tasa de error es de 4,6 % dándonos como indicativo que, de 1000 correos analizados, 46 de ellos no se clasificarán de forma adecuada dando un margen de perdida de información o ser víctimas de phishing.

En la siguiente subsección, se mostrarán los resultados de trabajos anteriores, al aplicar técnicas de ML para la detección de este tipo de delito.

7.1 Resultados para el sistema y los trabajos anteriores

El enfoque propuesto en este trabajo de investigación aborda las limitaciones de otras técnicas anti-phishing en términos mediciones de parámetros como la eficiencia computacional y la robustez. En esta sección, la evaluación del desempeño en parámetros de evaluación se compara con otras técnicas existentes. Esta comparación se basa en los trabajos relacionados a técnicas anti-phishing. La Tabla 5 presenta las estadísticas de rendimiento dando a notar que el método propuesto es uno de los mejores modelos anti-phishing existente en los anteriores trabajos.

Tabla 5. Comparación de trabajos relacionados con el método propuesto

Trabajo	Datos phish	Datos benig	Total	TPR	FPR	Exact
Sonowal y Kup	667	995	1.662	90,54	5,82	92,72
Kaur y Kalra	1.078	846	1.924	99,44	0,56	97,51
Chin	3.718	1.185	4.903	96,9	0,03	97,39
Tan y col.	500	500	1.000	99,2	7,8	91,41
El método propuesto	15.964	15.120	31.084	98,7	0,79	97,7

7.2 Validación Estadística

La validación, que se plantea desde el punto de vista estadístico tiene el objetivo de verificar qué modelo de clasificación es el adecuado para el APS teniendo como base el tiempo de ejecución para el desarrollo del mismo.

Las hipótesis contrastadas en un ANOVA para el tiempo de ejecución de los modelos utilizados en este trabajo son:

- H_0: No hay diferencias entre las medias de los tiempos de ejecución de los modelos: $\mu_1 = \mu_2 = \mu$.
- H_1: Las medias son significativamente distintas la una de la otra.

Se tiene 2 grupos y la cantidad de observaciones por grupo es de 10, por lo tanto, se tiene un modelo equilibrado. A continuación, se calcula las medias y las desviaciones típicas de cada grupo.

La media para el primer grupo es de (μ_1): 90 y la desviación estándar (σ_1): 0,83066238629146, para el grupo dos la media (μ_2): 94 y la desviación estándar (σ_2): 1,4494897427832 Por procesos anteriores, de esta investigación se puede decir que existen datos atípicos o diferencia de varianzas. En este caso, los 2 grupos parecen seguir una distribución simétrica.

Se continua con la verificación de las condiciones para un ANOVA, dentro de cada grupo los datos son independientes entre ellos ya que se ha hecho un probado aleatorio. La variable cuantitativa se distribuye de forma normal en cada uno de los grupos, para este estudio de normalidad se realizó mediante la forma gráfica y con el Test de normalidad *Shapiro-Wilk*, dado que el número de observaciones es menor a 50, se tiene para el grupo 1 el valor de W: 0,987166 y un p_val: 0,905192 mientras, que para el grupo 2 el valor de W: 0.970674 y un p_val: 0.706497.

Por lo tanto, no se tiene evidencia de falta de normalidad, para el siguiente paso se procede con la prueba ANOVA obteniendo como resultado el valor del estadístico de prueba, F=1,994349. Es significativamente distinto de 1 para cualquier nivel de significación y por tal razón, se rechaza la hipótesis nula de igualdad de medias. Además, el valor de eta cuadrado (η^2) es de 0.018, lo indica un tamaño de efecto pequeño.

Por lo tanto, las medias son significativamente distintas la una de la otra y por ende se llegaría a concluir que el modelo con el mejor tiempo de ejecución para esta investigación es el que presenta una media de 90 minutos siendo este RF, Ahora se puede aseverar que tanto con el test ANOVA y mediante la matriz de confusión para obtener las métricas RF es el modelo de clasificación que presenta mejor estabilidad para este tipo de desarrollo.

7.3 Discusión

Los esfuerzos que deben seguirse para la detección de phishing deben ser aún mayor, dado la creciente ola de tecnología que no nos deja otra opción de ser precavidos con la información que se divulga en la web, por esta razón se examinó: las limitaciones, el éxito y el impacto de esta investigación a nivel práctico y

teórico. En un enfoque práctico, las personas que revisen este artículo pueden tener una idea del cómo se construyó el vector de características y se aplicó en el modelo, para posteriormente mejorar su resiliencia en la seguridad cibernética al tratar de explicar cómo funciona este delito y prepararse para evitar caer en el phishing.

Desde el punto de vista teórico, esta investigación combina dos mundos el práctico y académico donde los informes de inteligencia de amenazas cibernéticas y los trabajos de investigación académica se utilizaron para aprender y desarrollar la comprensión de las capacidades, tácticas, técnicas y procedimientos de los atacantes.

Las limitaciones de esta investigación están relacionadas con los datos que se tienen para la ejecución de este modelo, dado que son registros que tienen una alta presencia de *NaN* y esto no permite tener los resultados que se tenía en mente. Además, hay que recalcar que los datos utilizados para la evaluación son de 3 cuentas personales, dando como resultado un conjunto de muestra pequeña para el entrenamiento de los modelos RF Y NB y se obtuvo un desempeño menor al que se obtuvo con los datos de experimentales. Sin embargo, durante este trabajo se procuró analizar las características que se están presentando actualmente en los ataques de phishing.

El éxito de esta investigación se basa principalmente en el análisis general de técnicas, tácticas y procedimientos de atacantes modernos que son comúnmente utilizados por los actores de amenazas del mundo real que se obtuvieron a través de la revisión de la literatura. Esto proporcionó más información sobre las acciones específicas y qué capacidades técnicas existen para eludir los controles de seguridad modernos, como el uso del candado de seguridad *Https* para el robo de datos personales. Aunque esta investigación analiza diferentes modelos para visualizar el que nos otorga un mejor nivel de predictibilidad, exactitud y precisión y así dar una pauta a la hora de diseñar e implementar controles de seguridad de compensación, detecciones y procedimientos de respuesta para proteger los datos críticos.

Los resultados y el contenido de este trabajo podrían ser utilizados por estudiantes que deseen aprender más sobre este delito cibernético; las técnicas, tácticas y procedimientos que usan comúnmente los actores de amenazas.

7.4 Abreviaturas y Siglas

APS	Anti-Phishing System
CRISP-DM	Cross Industry Standard Process for Data Mining
FSM	Módulo de selección de características
TF-IDF	Término-Frecuencia Documento Frecuencia Inversa
ROC	Receiver Performance Characteristic
URL	Uniform Resource Locator
TP	Verdadero Positivo
TN	Verdadero Negativo
FP	Falsos Positivos
FN	Falsos Negativos

8. Conclusiones y Recomendaciones

8.1 Conclusiones

(a) Según el análisis del material recopilado para esta investigación, los actores de amenazas modernos se basan más comúnmente en técnicas de phishing para obtener acceso inicial. Algunos de los medios más comunes para obtener este acceso parecen ser la implementación de malware en un documento de Office, que luego se ejecuta una vez que el usuario habilita las macros. Otro método que los atacantes parecen estar usando constantemente es el abuso de vulnerabilidades conocidas públicamente, ya que es mucho más rentable que descubrir vulnerabilidades previamente desconocidas.

(b) En este trabajo, se considera el problema de la detección de phishing utilizando enfoques de aprendizaje automático. En nuestro primer intento, llamado conjunto de prueba, se identifican varias características interesantes al analizar el problema desde el punto de vista estadístico. Se pudo visualizar que ver el problema desde un punto puramente, es insuficiente para resolverlo de manera efectiva. La intención del atacante de phishing también debe tenerse en cuenta para una solución eficaz.

(c) Describir un enfoque hacia el diseño de Funciones basadas en nombres de dominio, análisis tanto de URL como el corpus del email para la detección de phishing mediante aprendizaje automático. Nuestro diseño de funciones hizo hincapié en la eliminación del posible sesgo en la clasificación debido a conjuntos de datos de phishing y páginas legítimas elegidas. Nuestro enfoque difiere de otros trabajos en este espacio, ya que explora la relación del corpus del email con su intención de phishing. Con un conjunto de características balanceado, se obtuvo una tasa de clasificación de 97 % con datos de validación cruzada. Además, se mostró una tasa de detección de 97-97.7 % para URL activas en la lista negra.

(d) Nuestros tiempos de extracción y clasificación de características son muy bajos y muestran que nuestro enfoque es adecuado para la implementación en tiempo real. Es probable que nuestro enfoque sea muy eficaz en las estrategias de phishing modernas, como el phishing extremo, que están diseñadas para engañar incluso a los usuarios experimentados.

8.2 Recomendaciones

Las recomendaciones que se presentan en la sección final son de ayuda para trabajos futuros, teniendo en cuenta que esta investigación es un pequeño aporte a este campo tan extenso y relativamente nuevo.

(a) Los phishers buscan vulnerabilidades constantemente para atacar y a su vez no ser detectados al momento de sus ataques, por tal razón se recomienda ejecutar un análisis constante de las técnicas empleadas por estos atacantes, y así determinar características nuevas o que ayuden a mejorar el vector final y por ende obtener un modelo preciso con un nivel de predictibilidad más elevado.

(b) Se recomienda utilizar Datasets actuales para tener una mejor visión de este tipo de delitos cibernéticos, pero este punto es primordial dado que el tiempo de vida de estos dominios y URL_s es muy corto, por tal razón el conjunto de datos debe ser actualizado al menos cada 6 meses. Así se obtendrán resultados confiables y actuales.

(c) Revisar el trabajo de Orunsolu, Sodiya y Akinwale, en Orunsolu et al. (2019), dado que tiene una buena revisión de literatura para este tema de investigación.

REFERENCIAS

Aburrous, M., Hossain, M., Dahal, K. and Thabtah, F. (2010). Experimental case studies for investigating e-banking phishing techniques and attack strategies *Cognit. Comput.* (2), 242–253 https://doi.org/10.1007/s12559-010-9042-7

Adebowale, M., Lwin, K., Sanchez, E. and Hossain, M. (2018). Intelligent Web-Phishing Detection and Protection Scheme using integrated Features of Images, Frames and Text. *Expert System with Applications.* (115), 300-313 https://doi.org/10.1016/j.eswa.2018.07.067

Amat Rodrigo, Joaquín. (2020). Análisis de texto (text mining) con Python, cienciadedatos.net. Obtenido de: https://www.cienciadedatos.net/. (Diciembre, 2020).

Anwar, T., Abu-Kresha, M. and Bakry A. (2017). An efficient method for web page classification based on text. *International J. Eng. Comput. Sci.*

Barraclough, P. & Sexton, G. (2015). Phishing website detection fuzzy system modelling, *IEEE, London, UK*, 1384-1386, 10.1109/SAI.2015.7237323.

Breiman, L. (2001). Random Forests. Machine Learning *SpringerLink, 45*, 5–32, https://doi.org/10.1023/A:1010933404324

Calva Yaguana, Karen Priscilla. (2020). *Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado automatizado en R.* [Tesis de pregrado, Escuela Politécnica Nacional]. Repositorio institucional de la Escuela Politécnica Nacional. https://bibdigital.epn.edu.ec/

Chin, T., Xiong, K. and Hu, C. (2015). PhishLimiter: A Phishing Detection and Mitigation Approach using Software-Defined Networking, *IEEE Access, 6*, 42516-42531, 10.1109/ACCESS.2018.2837889

Cortina, V. G. (2015). *Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario*. [Universidad Carlos III de Madrid]. Departamento de Informática.

Creswell. (2015). Educational research. Planning, conducting and evaluating quantitative and qualitative research. *USA*.

Dhanalakshmi, R. & Chellappan, C. (2013). Detecting Malicious URLs in E-mails- An Implementation. *AASRI Procedia, 4*, 125-131, https://doi.org/10.1016/j.aasri.2013.10.020

Gansterer, W.N. & Polz, D. (2009). E-mail classification for phishing defense, in Advances in Information Retrieval. *Heidelberg: Springer Berlin Heidelberg*, 449–460.

Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., Caihuelas Quiles, R. (2020). *Minería de datos Modelos y algoritmos*, Editorial UOC.

Gowtham, R., Gupta, J. and Gamya, P.G. (2017). Identification of phishing web pages and their target domains by analyzing the feign relationship *J. Informat. Secur. Appl, 35*, 75-84

Gowtham,R. & Krishnamurthi, I. (2014). PhishTackle-a web services architecture for anti-phishing *Cluster Compt, 17*, 1051–1068. https://doi.org/10.1007/s10586-013-0320-5

Gupta, B.B., Tewari, A., Jain, A.K. and Agrawal, P. (2017). Fighting against phishing attacks: state of the art and future challenges. *Neural Comput. Applic, 28*, 3629–3654 https://doi.org/10.1007/s00521-016-2275-y

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Hota, H.S., Shrivas, A.K. and Hota, R. (2018). An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique *Procedia Comput. Sci, 132*, 900-907, 10.1016/j.procs.2018.05.103

Isa, D., Lee, L., Kallimani, V. and Rajkumar, R. (2016). Text document pre-processing using bayes formula for classification based on the vector space model Comput. *Informat. Sci. J, 1* (4), 79-90.

Jain, AK & Gupta, BB. (2016). A novel approach to protect against phishing attacks at client side using auto-updated. *EURASIP Journal on Information Security*(1), 1-11.

Kittler, J., Hatef, M. and Duin, R.P.W. (1998). On Combining Classifiers. Transactions on pattern analysis and machine intelligence. *IEEE, 20*.

Khonji, M., Iraqi, Y. and Jones, A. (2013). Phishing Detection: A Literature Survey. *IEEE Communications Surveys & Tutorials, 15*(4), 2091-2121.

Kuncheva, L. (2004). Combining Pattern Classifiers. Methods and algorithms. *John Wiley & Sons, New Jersey*.

Martínez, M. B. (2018). Minería de Datos. web: http://bbeltran.cs.buap.mx/NotasMD.pdf.

Moghimi, M. & Varjani, A.Y. (2016). New rule-based phishing detection method *Expert systems with applications, 53*, 231-242.

Monkey.org. (2020). Data de correos electrónicos Monkey.org. Web de Monkey.org: https://monkey.org/ jose/phishing/; (2020)

CSO Online report on phishing activities. Accessed 2016 http://www.csoonline.com/articles

Orunsolu, A.A., Afolabi, O., Sodiya, A.S. and Akinwale, A.T. (2019). A Users' Awareness Study and Influence of Socio-Demography Perception of Anti-Phishing Security Tips. Acta Informatica Pragensia, 7(2), 138-151.

Pedregosa. (2011). Scikit-learn: Machine Learning in Python *JMLR 12*, 2825-2830

Phishtank dataset. (2021). http://www.phishtank.com. (2021).

Qabajeh, I., Thabtah, F. and Chiclana, F. (2018). A recent review of conventional vs. automated cybersecurity anti-phishing techniques *Computer Science Review, 29*, 44-55.

Rosero Gomezcoello, Johanna Mishell. (2020). *Detección y mitigación de ataques de ingeniería social tipo Phishing utilizando minería de datos* [Tesis de pregrado, Escuela Superior Politécnica del Ejercito]. Repositorio institucional de la Escuela Superior Politécnica del Ejercito. http://repositorio.espe.edu.ec/

Segal, M. (2004). *Machine learning benchmarks and random forest regression* [Tesis, University of California].

Shabtai, A., Kanonov, U., Elovici, Y., Glezer, C. and Weiss, Y. (2016). Andromaly: a behavioural malware detection framework for android devices. *Journal of Intelligent Information Systems, 38*(1), 161-190.

Sonowal, G. & Kuppusamy, K.S. (2020). PhiDMA- A phishing detection model with a multi-filter approach *Journal of King Saud University-Computer and Information Sciences, 32*(1), 99-112.

Sonowal, G. & Kuppusamy, K.S. (2018). MMSPhiD: A Phoneme based Phishing Verification Model for Persons with Visual Impairments. *Information and Computer Security Journal*.

Tan, C. L., Chiew, K. L. and Sze, S. N. (2017). Phishing Webpage Detection Using Weighted URL Tokens for Identity Keywords Retrieval. *In 9th International Conference on Robotic, Vision, Signal Processing and Power Applications* (pp. 133-139). Springer, Singapore.

Moghimi, M. and Varjani, A.Y. (2016). New rule-based phishing detection method.*Expert Systems with Applications*, (pp. 231-242), https://doi.org/10.1016/j.eswa.2016.01.028..

William, W. and Cohen, MLD.CMU. (2019). Base de datos de correos electrónicos de Enron. Web de https://www.cs.cmu.edu/ ./enron/.

Zhao, J., Wang, N., Ma, Q. and Cheng, Z. (2019). Classifying Malicious URLs Using Gated Recurrent Neural Networks. *Springer, Cham*, 385–394.

Zouina, M. & Outtaj, B. (2017). A novel lightweight URL phishing detection system using SVM and similarity index *Human-centric Computing and Information Sciences, 7*(1), 1-13.

BIOGRAFÍA

Fernanda Albán Toapanta. Egresada en Ingeniería Matemática mención Estadística e Investigación Operativa (Escuela Politécnica Nacional, 2021). Dicto charlas en Arcotel, Sercop, CNT, Women in Data Science y en la sesión de Ciberseguridad del Congreso de Investigación Aplicada a Ciencia de Datos – II Congreso Nacional de R Users Group. Realizo análisis estadísticos para Jedal In Software & Al, We Trust y Sercop. Para su trabajo de titulación me ha enfocado en el área de Ciberseguridad específicamente la detección de Phishing.

Ménthor Urvina Mayorga. Matemático (Escuela Politécnica Nacional, 1990). Magister en Investigación Operativa, mención Gerencia, 1998 (Universidad Andina Simón Bolívar – Escuela Politécnica nacional). Profesor Principal a Tiempo Completo de la Escuela Politécnica Nacional, adscrito al Departamento de Matemática. Autor de libros de Cálculo Vectorial y Ecuaciones Diferenciales Ordinarias. Imparte las cátedras de Cálculo Vectorial, Ecuaciones Diferenciales Ordinarias, Probabilidad y Estadística. Áreas de interés: Cálculo, Ecuaciones Diferenciales Ordinarias, Estadística, Probabilidades, Análisis Numérico, Investigación de Operaciones.

Roberto Omar Andrade. Ingeniero en Electrónica y Telecomunicaciones (Escuela Politécnica Nacional (EPN), 2007). Magister en Gestión de Redes y Telecomunicaciones (Escuela Politécnica del Ejército, 2013), en la actualidad es estudiante de doctorado en Sistemas de Seguridad en la EPN. Oficial de Seguridad del Ministerio de Educación de Ecuador (MINEDUC, 2015), Coordinador de Infraestructura Tecnológica en la Secretaría Nacional de Planificación SENPLADES 2013-2014. Centro de datos, seguridad y administración de redes en SENPLADES y Tecnología Sucre 2009-2013 e Ingeniería Técnica para sistemas VoIP en SERATVoIP 2007-2011. Es instructor técnico certificado de CCNA, CCNP y CCNA Security en EPN desde 2010 hasta la fecha.

Apéndice A. PRIMER APÉNDICE

Tabla 6. Lista de notaciones y sus significados.

Notaciones	Descripciones
n	Número de funciones para el análisis de frecuencia
F_{url_i}	Una instancia de la función de URL
d_{ph}	Una base de datos de URL de phishing confirmadas
d_{be}	Una base de datos de URL legítimas confirmadas
θ	El umbral para el análisis de características
p	Tamaño del conjunto de datos
S_i	Categoría de función para la función F2
H_P	Funciones de phishing de alto impacto
$CFS(s)$	Función de selección de característica de correlación para s
f	Una instancia de característica en una categoría de característica particular
t	Factor de correlación (incertidumbre de simetría o correlación Coeficiente de Pearson)
a	Encimera
$f(s)$	El conjunto de características de alto impacto seleccionadas
$m(fs)$	El subconjunto de funciones de alto impacto seleccionadas para la detección de phishing

Análisis de Correspondencias Múltiples para el Estudio de los Homicidios Intencionales en el Ecuador

Abril, Mauricio[1,*]; Chariguamán, Nancy[2]; Aguilar, Johanna[3]

[1]Escuela Superior Politécnica de Chimborazo, Sede Orellana, Coca, Ecuador
[2]Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Riobamba, Ecuador
[3]Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Riobamba, Ecuador

Resumen: El presente estudio permite analizar las relaciones entre las categorías de variables asociadas con los homicidios intencionales en el Ecuador, por medio del Análisis de Correspondencias Múltiples, para determinar las relaciones entre las modalidades que influyen y contribuyen al cometimiento de este tipo de violencia. Esto permite optimizar la elaboración e implementación de política pública para reducir y erradicar los niveles de violencia que se tiene en el país. Se obtiene que el tipo de muerte violenta depende principalmente de la provincia en la que se produce, con un tipo de arma y en rangos etarios de personas jóvenes. Estos resultados permiten el tratamiento desde diferentes áreas del conocimiento para mitigar este fenómeno de violencia y garantizar un adecuado nivel de seguridad a la ciudadanía.

Palabras clave: Homicidios, correspondencias múltiples, políticas

Multiple Correspondence Analysis for the Study of Intentional Homicides in Ecuador

Abstract: The present study allows to analyze the relationships between the categories of variables associated with intentional homicides in Ecuador, through Multiple Correspondence Analysis to determine the relationships between the modalities that influence and contribute to the committing of this type of violence. This allows to optimize the elaboration and implementation of public policy to reduce and eradicate the levels of violence that exist in Ecuador. It is obtained that the type of violent death depends mainly on the province in which it occurs, with a type of weapon and in range age groups of young people. These results allow the treatment from different areas of knowledge to mitigate this phenomenon of violence and guarantee an adequate level of security in citizenship.

Keywords: Homicides, correspondence análisis, politics

1. INTRODUCCIÓN

Desde el aparecimiento humano en el planeta, se ha evidenciado el cometimiento de la muerte no voluntaria de un ser humano, lo que haría pensar, que este comportamiento es natural de la convivencia de la especie. Esta práctica violenta ha evolucionado al ritmo del desarrollo de la humanidad, con las particularidades y lógicas asociadas a cada cultura o régimen económico.

Al focalizar este fenómeno en América Latina, se puede indicar dos hitos importantes que lo han suscitado. El primero es la producción de sustancias sujetas a fiscalización (Drogas ilegales), seguido de la distribución y comercialización de estas sustancias a países consumidores. Esto ha generado entre otros, los principales clientes del sistema penitenciario en la mayoría de los países de Latinoamérica (Wacquant, 2007).

El segundo es el tráfico de armas para grupos que operan al margen de la ley, que por lo general están asociados con el mercado ilegal del negocio de drogas. Estos fenómenos, en el Ecuador, son tipificados como delitos en el Código Orgánico Integral Penal (COIP); es decir, al estar relacionados con economías ilegales globales y crimen organizado.

Los mercados ilegales, administrados y gestionados por el crimen organizado formalizaron en los noventas una cantidad

mauricioabrildonoso@gmail.com
Recibido: 14/03/2022
Aceptado: 03/08/2022
Publicado en línea: 23/12/2022
10.33333/rpvol50n3.04

de por lo menos 800.000 millones de dólares anuales lo que equivale al 15% del total del comercio mundial (Curbet, 2007).

Por esto, el rol del Estado ecuatoriano para tratar este tipo de violencia incluyendo las relacionadas, tiene que ser reconsiderado, dada la permeabilidad territorial a través de sus fronteras y la difuminación de las barreras entre los ámbitos de Policía Exterior e Interior (Rivera, 2011). Como evidencia, se tiene los hechos de violencia que se han producido en nuestro país, aquellos ocurridos en los Centros de Rehabilitación Social. Así, desde el año 2000 hasta el 2018 se ha producido un incremento de alrededor del 120% de la población penitenciaria en la región Latinoamericana (Kaleidos, 2021), lo que produce pugnas violentas entre los internos por controlar los Centros de Rehabilitación Social.

De las primeras investigaciones judiciales y reportes de prensa, estos hechos obedecerían a pugnas entre bandas delictuales asociadas con el tráfico y expendio de sustancias sujetas a fiscalización, relacionadas principalmente con carteles mexicanos y colombianos, con el objetivo de captar el almacenamiento, envío de estas sustancias hacia países consumidores y la distribución y comercialización al interior del Ecuador. Esta disputa violenta se genera como consecuencia de la economía ilegal a la que está asociada esta actividad ilícita.

Otro aspecto importante a considerar, asociado a los homicidios intencionales, es el aparecimiento de delitos considerados conexos como: delitos transnacionales, tráfico de personas, armas, delitos aduaneros, lavado de activos, entre otros.

Con base en estos antecedentes, se realiza un estudio que determine la relación entre las diferentes variables: espacio temporal y las asociadas a las víctimas del homicidio intencional.

Desde la criminología, el delito se produce donde tiene la facilidad de producirse, sea esto por falta de presencia del Estado, limitaciones en el accionar de las instituciones de Justicia.

En América Latina y particularmente en nuestro país, el delito está asociado con el conocimiento limitado de las causas que facilitan la realización de estos hechos (Fiscalía General del Estado, 2015).

Situación que permitirá a las autoridades encargadas de la seguridad interna y externa, el poder formular planes, programas y proyectos focalizados a reducir los niveles de violencia que se vienen presentando.

Sin duda, una eficiente y eficaz respuesta de la justicia criminal debe sustentarse en la medida de lo posible, en un número de indicadores, tales como: homicidios resueltos por la Policía y Fiscalía, personas arrestadas y sentenciadas por homicidios, entre otros (United Nations Office on Drugs and Crime, 2013). Además, para la formulación de política pública orientada a garantizar la seguridad e integridad de todos y todas las personas en el país, ya que es un mandato constitucional.

Los homicidios y asesinatos, al ser un fenómeno social y de comportamiento, es claro que no se puede encontrar soluciones ni en lo puramente educativo, en la atención social, ni en lo represivo (Wacquant, 2004), por lo que tiene que ser analizado en conjunto con técnicas y metodologías apropiadas y adecuadas al tipo de datos que se dispone.

En el Ecuador, la Policía Nacional es la institución encargada de realizar el levantamiento de un cadáver cuando este hecho se produce de manera violenta o por causas externas, para luego con la dirección investigativa de la Fiscalía, esclarecer estos actos de violencia hasta obtener una reparación integral de las víctimas y resolverlos según jurisprudencia.

Para lo cual, a partir del año 2010, se tiene una reforma policial en el Ecuador, donde la principal acción fue el desconcentrar el accionar de esta institución en territorio, adoptando un modelo de zonas, circuitos y distritos (Pontón y Rivera, 2016).

Una de las mejoras que se obtuvo fue el contar con información válida, confiable y consistente tratada técnicamente por todas las instituciones del Sector Seguridad, la misma que servirá como materia prima del presente estudio.

2. MATERIALES Y MÉTODOS

Este estudio inicia con un análisis exploratorio, que permita conocer a priori, el tipo de cada una de las variables y sus categorías, las cuales se recaban principalmente, de la realización del Protocolo de Autopsia que es un impulso fiscal hacia Medicina Legal como parte integral de la investigación criminal (Código Orgánico Integral Penal, 2014) y el comportamiento univariante.

La Tabla 1 muestra la descripción de 13 variables asociadas con las muertes violentas que se produjeron en el país, información recabada de la Comisión de Seguridad.

Las variables: edad, hora, y fecha, fueron transformadas a variables categóricas, debido a que, en este tipo de estudios se orienta al uso de variables cualitativas en lugar de cuantitativas (Wackerly et al., 2010), además de que el Análisis de Correspondencias Múltiples (ACM) se lo realiza sobre variables cualitativas y categóricas.

Este tipo de muertes por lo general están relacionadas con el tipo de arma que se emplea. Así, más de medio millón de homicidios a nivel mundial es producto del uso de armas ligeras por personas civiles (Dammert, 2006).

En este sentido, se realizó un análisis descriptivo. Asimismo, se realizaron tablas de frecuencia, conocidas también como tablas de contingencia, cruzando todas las variables con el tipo de arma.

Tabla 1. Variables que describen el fenómeno de los homicidios intencionales en el Ecuador

Variable	**Blanca, N = 4,301[1]**	**Fuego, N = 9,562[1]**	**Otras, N = 2,466[1]**	**p-value[2]**
Tipo de delito				<0.001
Asesinato	2,267 (22%)	6,871 (67%)	1,114 (11%)	
Femicidio	225 (48%)	65 (14%)	176 (38%)	
Homicidio	1,806 (33%)	2,535 (46%)	1,173 (21%)	
Sicariato	3 (3.1%)	91 (94%)	3 (3.1%)	
Provincia				
Azuay	140 (36%)	135 (34%)	118 (30%)	
Bolívar	25 (24%)	47 (44%)	34 (32%)	
Cañar	49 (40%)	50 (40%)	25 (20%)	
Carchi	43 (43%)	21 (21%)	35 (35%)	
Chimborazo	63 (40%)	12 (7.5%)	84 (53%)	
Cotopaxi	97 (43%)	30 (13%)	98 (44%)	
El Oro	185 (18%)	736 (71%)	115 (11%)	
Esmeraldas	385 (27%)	919 (65%)	107 (7.6%)	
Galápagos	1 (50%)	0 (0%)	1 (50%)	
Guayas	829 (16%)	3,725 (73%)	563 (11%)	
Imbabura	112 (43%)	47 (18%)	100 (39%)	
Loja	62 (36%)	46 (27%)	63 (37%)	
Los Ríos	326 (20%)	1,142 (72%)	124 (7.8%)	
Manabí	242 (15%)	1,235 (76%)	156 (9.6%)	
Morona Santiago	37 (35%)	30 (29%)	38 (36%)	
Napo	22 (37%)	19 (32%)	19 (32%)	
Orellana	70 (42%)	56 (34%)	40 (24%)	
Pastaza	33 (51%)	10 (15%)	22 (34%)	
Pichincha	1,088 (55%)	433 (22%)	450 (23%)	
Santa Elena	32 (21%)	81 (53%)	40 (26%)	
Santo Domingo de los Tsáchilas	216 (32%)	377 (56%)	82 (12%)	
Sucumbíos	126 (24%)	337 (65%)	54 (10%)	
Tungurahua	93 (45%)	29 (14%)	86 (41%)	
Zamora Chinchipe	15 (38%)	13 (33%)	11 (28%)	
Zona no delimitada	10 (23%)	32 (74%)	1 (2.3%)	
Área				<0.001
Rural	1,356 (26%)	2,934 (56%)	940 (18%)	
SD	54 (23%)	127 (54%)	53 (23%)	
Urbano	2,891 (27%)	6,501 (60%)	1,473 (14%)	
Sexo de la víctima				
Hombre	3,572 (25%)	8,928 (63%)	1,680 (12%)	
Mujer	728 (34%)	634 (30%)	782 (36%)	
SD	1 (20%)	0 (0%)	4 (80%)	
Nacionalidad de la víctima				<0.001
Colombiana	87 (22%)	276 (70%)	34 (8.6%)	
Ecuatoriana	4,116 (26%)	9,177 (59%)	2,373 (15%)	
Otra	68 (33%)	85 (41%)	52 (25%)	
Venezolana	30 (49%)	24 (39%)	7 (11%)	
Estado civil				<0.001
Casado	794 (27%)	1,669 (56%)	524 (18%)	

Variable	**Blanca, N = 4,301[1]**	**Fuego, N = 9,562[1]**	**Otras, N = 2,466[1]**	**p-value[2]**
Divorciado	138 (29%)	227 (49%)	103 (22%)	
SD	292 (28%)	585 (56%)	177 (17%)	
Soltero	2,677 (27%)	5,776 (59%)	1,417 (14%)	
Unión libre	351 (20%)	1,263 (71%)	159 (9.0%)	
Viudo	49 (28%)	42 (24%)	86 (49%)	
Etnia				<0.001
Afroecuatoriano	337 (25%)	923 (68%)	100 (7.4%)	
Blanca	64 (21%)	189 (62%)	53 (17%)	
Indígena	107 (36%)	62 (21%)	125 (43%)	
Mestizo	3,491 (27%)	7,418 (57%)	1,996 (15%)	
Montubio	96 (21%)	317 (69%)	44 (9.6%)	
Otra	206 (20%)	653 (65%)	148 (15%)	
Lugar del evento				<0.001
Centro de rehabilitación social	70 (46%)	37 (25%)	44 (29%)	
Hogar	1,300 (34%)	1,604 (42%)	952 (25%)	
Otro	782 (25%)	1,733 (55%)	640 (20%)	
Via pública	2,149 (23%)	6,188 (68%)	830 (9.1%)	
Año				<0.001
2010	538 (21%)	1,759 (67%)	323 (12%)	
2011	486 (21%)	1,548 (67%)	284 (12%)	
2012	484 (25%)	1,188 (62%)	237 (12%)	
2013	413 (24%)	981 (57%)	326 (19%)	
2014	403 (32%)	681 (54%)	187 (15%)	
2015	298 (28%)	546 (52%)	211 (20%)	
2016	329 (34%)	446 (47%)	180 (19%)	
2017	297 (31%)	505 (52%)	171 (18%)	
2018	342 (35%)	460 (47%)	170 (17%)	
2019	359 (31%)	647 (55%)	166 (14%)	
2020	352 (26%)	801 (59%)	211 (15%)	
Mes				0.009
Enero	389 (27%)	881 (60%)	196 (13%)	
Febrero	334 (25%)	781 (58%)	230 (17%)	
Marzo	413 (28%)	844 (57%)	226 (15%)	
Abril	347 (25%)	847 (60%)	207 (15%)	
Mayo	400 (28%)	809 (57%)	206 (15%)	
Junio	370 (27%)	843 (61%)	173 (12%)	
Julio	321 (26%)	763 (61%)	163 (13%)	
Agosto	363 (28%)	746 (57%)	203 (15%)	
Septiembre	334 (26%)	739 (58%)	211 (16%)	
Octubre	322 (25%)	750 (58%)	214 (17%)	
Noviembre	318 (24%)	761 (59%)	220 (17%)	
Diciembre	390 (28%)	798 (57%)	217 (15%)	
Día de la semana				<0.001
Lunes	531 (27%)	1,121 (56%)	351 (18%)	
Martes	427 (25%)	1,005 (58%)	306 (18%)	
Miércoles	381 (22%)	1,051 (61%)	300 (17%)	
Jueves	398 (22%)	1,141 (63%)	274 (15%)	
Viernes	554 (25%)	1,337 (60%)	334 (15%)	

Variable	**Blanca, N = 4,301[1]**	**Fuego, N = 9,562[1]**	**Otras, N = 2,466[1]**	**p-value[2]**
Sábado	820 (28%)	1,742 (59%)	415 (14%)	
Domingo	1,190 (31%)	2,165 (56%)	486 (13%)	
Rango horario				<0.001
Madrugada	1,339 (32%)	2,188 (52%)	720 (17%)	
Mañana	734 (28%)	1,488 (56%)	419 (16%)	
Noche	1,432 (24%)	3,715 (62%)	812 (14%)	
Tarde	796 (23%)	2,171 (62%)	515 (15%)	
Rango de edad				<0.001
20-45	2,990 (26%)	7,277 (63%)	1,316 (11%)	
45-65	698 (28%)	1,313 (53%)	473 (19%)	
Mayor a 65	228 (35%)	164 (25%)	268 (41%)	
Menor a 20	385 (24%)	808 (50%)	409 (26%)	

[1]n (%)

[2]Pearson's Chi-squared test

La Tabla 1 muestra también el P-valor asociado con el análisis de tablas de contingencia para la prueba de bondad de ajuste Chic cuadrado (Brandt, 2014). Datos que permiten evidenciar dependencia entre las variables, en todas se tiene significancia estadística, ya que su p-valor es menor al 5%, esto indica, que están relacionadas entre sí (Agresti, 2002), el tipo de arma con las variables temporales, geográficas, y las asociadas a la víctima.

De esta primera aproximación se puede proponer actividades para prevenir el cometimiento de este tipo de violencia, sin embargo, estas no toman en cuenta la interacción entre las categorías, por lo que se aplicará la metodología que se describe a continuación.

Se verifica a través de técnicas multivariantes de manera global las relaciones entre variables y también las relaciones entre sus categorías, por medio del ACM. Una aproximación de este análisis para la ciudad de Quito se puede consultar en (Abril y Castro, 2015). En seguida, se presenta en detalle el ACM, técnica utilizada para este estudio.

2.1. Análisis de correspondencia múltiple (ACM)

El ACM estudia las relaciones entre cualquier número de variables, cada una de ellas con varias modalidades. Estas relaciones se representan generalmente en un gráfico bidimensional.

El ACM está diseñado para analizar tablas disyuntivas completas, que son tablas de contingencia de variables cualitativas. Se tiene que considerar que, las modalidades de cada variable son mutuamente excluyentes, y cada individuo pertenece a una y solo una de ellas.

Otro aspecto que aborda el ACM es el poder reducir las dimensiones de la tabla de datos (Hardle y Simar, 2015), lo que es similar al Análisis de Componentes Principales (ACP), pero en este caso con variables categóricas, y la descomposición de la tabla se realiza en sus factores.

Una tabla disyuntiva completa **Z** queda descrita mediante los siguientes parámetros:

- Un conjunto de individuos $I = 1, \cdots, i, \cdots, n$
- Un conjunto de variables categóricas (cualitativas u ordinales) $J_1, \dots, J_k, \dots, J_Q$
- Un conjunto de categorías para cada variable $1, \dots, m_k$

El número total de modalidades viene dado por:

$$J = \sum_k m_k \tag{1}$$

Con la Ecuación (1) se obtiene la siguiente matriz de individuos, variables y modalidades, para aplicar la reducción de dimensiones para así visualizar las categorías en el plano bidimensional.

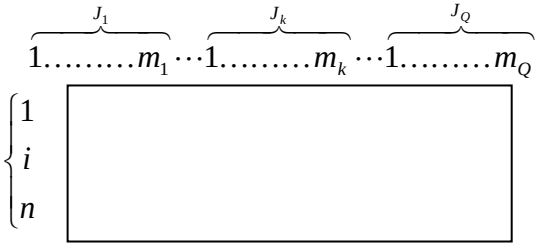

Figura 1. Diseño de la matriz Z

La Figura 1 muestra que la matriz Z es una tabla $I * J$. El elemento z_{ij} puede tomar el valor de 0 o 1, según el individuo i tome la modalidad j o no, por lo que, es una variable dicotómica.

2.2. Metodología para realizar un ACM

A continuación, se muestran las particularidades de un ACM aplicado a una tabla disyuntiva completa.

a) Significado de la terminología

Los elementos de $\boldsymbol{Z}$ $z_{ij} = k_{ij}$ son 0 o 1, con esto se tiene, el número de variables que describen el fenómeno:

$$k_i = \sum_i k_{ij} = Q \quad (2)$$

De la Ecuación (2) se muestra el inverso del número de preguntas que toma el valor de 0 o 1 según el individuo haya elegido o no la modalidad *j*, como en la Ecuación (3).

$$f_{ij}/f_i = k_{ij}/k_i = 1/Q \quad (3)$$

Donde $f_{ij}/f_i = k_{ij}/k_i$ son los perfiles fila (Lebart et al., 1985).

Los $k_j = \sum_i k_{ij}$, corresponden al número de individuos que poseen la modalidad *j*.

b) Matriz a diagonalizar

Para obtener los factores es necesario diagonalizar la matriz *V*, en caso de requerir más detalles ver (Grande y Abascal, 1989), en este caso particular se convierte en:

$$V = \frac{1}{Q} D^{-1} B \quad (4)$$

En la Ecuación (4), se tiene la matriz $B = Z'Z$, que es la tabla de Burt y es una matriz simétrica formada por Q^2 bloques:

- Los bloques de la diagonal son tablas diagonales (Díaz y Morales, 2012) que cruzan una pregunta con sí misma $Z_k' Z_k$.
- Los bloques fuera de la diagonal son tablas de contingencia obtenidas cruzando las preguntas de dos en dos $Z_k' Z_{k'}$.

0	C_{12}		C_{1Q}
C_{21}	0		C_{2Q}
		0	
C_{Q1}	C_{Q2}		0

Figura 2. Matriz de bloques que se obtiene de la tabla de Burt

De la Figura 2, se tiene la matriz *D*, que es diagonal cuyos elementos son los de la matriz de Burt, los efectivos de cada modalidad.

c) Fórmulas de transición

Sustituyendo los valores de la Ecuación (2) en las fórmulas de transición del Análisis Factorial de Correspondencias (AFC) simples, se tiene:

$$F_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_j \frac{f_{ij}}{f_i} G_\alpha(j) \quad (5)$$

$$G_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i \frac{f_{ij}}{f_i} F_\alpha(i) \quad (6)$$

De las Ecuaciones (5) y (6) se obtiene las siguientes fórmulas para el ACM

$$F_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{Q} \sum_j k_{ij}\, G(j) \quad (7)$$

$$G_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \frac{1}{k_j} \sum_i k_{ij}\, F(i) \quad (8)$$

Las Ecuaciones (7) y (8) representan las fórmulas de transición del ACM.

d) Centros de gravedad en las nubes y subnubes

1. El centro de gravedad de los puntos de variables $N(J)$ en AFC es $\sqrt{f_i}$. En este caso, es la distribución uniforme $U(1/\sqrt{n})$. En efecto de (2) se tiene la Ecuación (9)

$$\sqrt{f_i} = 1/\sqrt{n} \quad (9)$$

2. El centro de gravedad de las modalidades de cada pregunta, es el mismo que el de la nube de modalidades $N(J)$, $1/\sqrt{n}$. En efecto, el centro de gravedad de la sub tabla se obtiene a partir de su distribución marginal.

3. Como el AFC es centrado y el centro de gravedad de las modalidades de una pregunta coincide con el conjunto J, y con el origen, las modalidades de cada cuestión están centradas en torno al origen, no pueden tener todos los mismos signos.

e) Ayudas a la interpretación

Se define la Ecuación (10) como la contribución de una variable J_k al factor α como la suma de las contribuciones de las modalidades de la variable:

$$CTA_\alpha(J_k) = \sum_{j \in J_k} CTA_\alpha(j) \quad (10)$$

f) Las inercias

1.- Si *G* representa el centro de gravedad, la inercia debida a la modalidad *j* es:

$$I(j) = f_j d^2(j, G) = f_j \sum_i \left(\frac{f_{ij}}{f_i \sqrt{f_i}} - \sqrt{f_i} \right)^2 = \frac{1}{Q}\left(1 - \frac{k_j}{n}\right) \quad (11)$$

La Ecuación (11) depende de las variables y los perfiles fila.

2.- La inercia de una variable es la suma de inercias de las modalidades que se expresa en la Ecuación (12).

$$I(J_k) = \sum_{j \in J_k} I(j) = \sum_{j \in J_k} \frac{1}{Q}\left(1 - \frac{k_j}{n}\right) = \frac{1}{Q}(m_k - 1) \quad (12)$$

3.- La inercia total es la suma de las inercias de todas las preguntas.

$$I = \sum_k I(J_k) = \sum_k \frac{1}{Q}(m_k - 1) = \frac{J}{Q} - 1 \quad (13)$$

La inercia total de la Ecuación (13) depende del número de modalidades totales y el número de variables, note que, mientras más modalidades tengan las variables categóricas, menor será la inercia total.

3. RESULTADOS Y DISCUSIÓN

Se muestran los resultados obtenidos al emplear esta técnica a través de las librerías de R "FactoMineR" (Husson et al., 2020) y "factoextra" (Kassambara, 2020), las que permiten realizar un ACM.

Se inicia el estudio, estimando las contribuciones de las dimensiones que explica la varianza de las variables y categorías, lo cual permite identificar el número de ejes donde se representaran las variables y categorías.

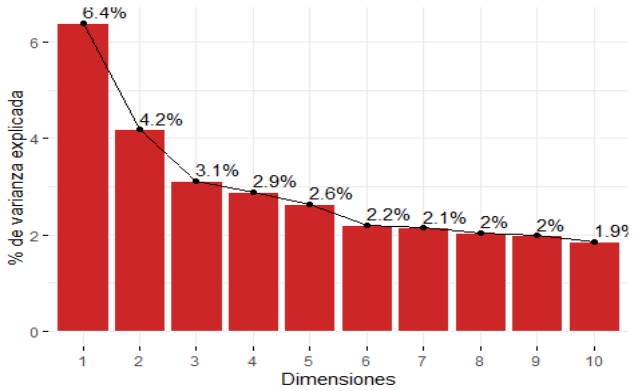

Figura 3. Varianza explicada por los valores y vectores propios

Con estas consideraciones, la Figura 3 muestra que se debe considerar tres ejes para representar las variables y categorías, esto en función de la varianza explicada, la interpretación es similar al que se tiene en el Análisis de Componentes Principales (ACP) para variables cuantitativas (Deisenroth et al., 2020). Es claro que la forma en que se calcula el porcentaje de la varianza explicada está relacionada con la naturaleza de las variables, y la distancia que se emplee en cada una de las metodologías.

La varianza explicada por el ACM es menor en relación a la varianza que explica un ACP, esto debido a que un ACP estudia las relaciones lineales, mientras que en un ACM se estudian relaciones mucho más generales, por lo que, se requiere de al menos un $min(J_k, J_q) - 1$, dimensiones para representar la relación entre dos variables con *k y q* categorías respectivamente (Husson y Jérôme, 2017).

Por consiguiente, se estudian más dimensiones en el ACM que en un ACP, en el presente estudio, el número de dimensiones a considerar sería de alrededor de 23. Se obtiene de la *i-ésima* dimensión la cual representa la fracción de la variabilidad total, que está dado por:

$$D_i = \lambda_i / \sum_{j=1}^{p} \lambda_j \quad (14)$$

Donde en la Ecuación (14) los λ_j son los valores propios asociados a la matriz de Burt (Zelterman, 2015).

A continuación, se presenta el comportamiento de las variables categóricas en las dimensiones establecidas y determinadas en la Figura 3.

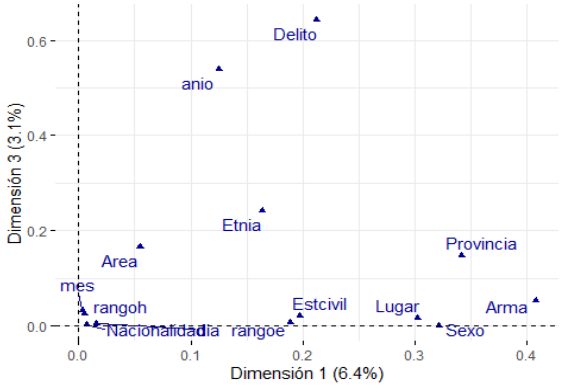

Figura 4. Representación de todas las similitudes de las variables en los ejes 1-3

La Figura 4 muestra que la dimensión 1 y la dimensión 3 explican aproximadamente el 9% de la variabilidad. Es importante indicar que las variables que se ubican al lado derecho del origen aportan con las contribuciones más significativas a la inercia. De ahí que, son aquellas variables que se encuentran a mayor distancia del origen (Beh y Lombardo, 2014).

Así, las que más representan al conjunto de variables, en este análisis son: Provincia, sexo de la víctima y el arma utilizada en estas muertes.

Otro aspecto a considerar es que las variables área, mes, rango horario, nacionalidad, están correlacionadas por la ubicación respecto a los ejes, pero se ubican muy cerca del origen, esto indica que no aportan significativamente la explicación de la varianza.

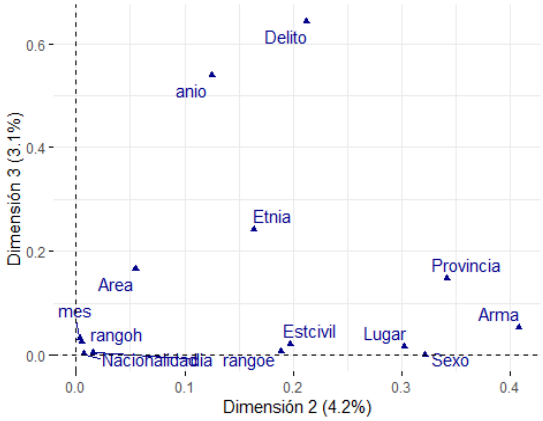

Figura 5. Representación de todas las similitudes de las variables en los ejes 2-3

La Figura 5 indica la contribución en la explicación de la varianza de las dimensiones 2-3.

Se observa que estas contribuyen a explicar aproximadamente un 7% de la varianza total. Muestra también similitudes semejantes a las descritas en el apartado anterior, salvo que, para la variable delito, se relaciona de cierta forma en la dimensión 2 con el estado civil, el rango de edad y el lugar donde ocurrió el evento.

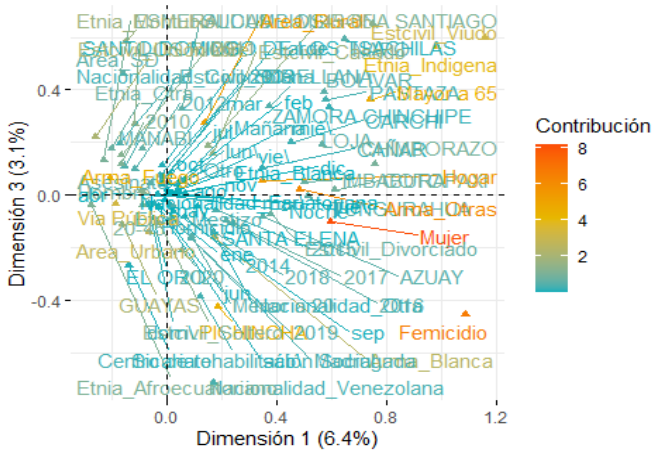

Figura 6. Representación de las similitudes de las categorías de las variables en los ejes 2-3

La Figura 6 presenta la relación de las categorías en los ejes 1-3, de esta se observa que el arma de fuego se relaciona en la primera dimensión con provincias como Guayas, El Oro, Santo Domingo de los Tsáchilas, Los Ríos y Esmeraldas. Además, interacciona con los asesinatos, hombres en edades de 20 a 45 años.

Contrariamente, se tiene que los homicidios relacionan la provincia de Pichincha, Carchi, Imbabura, Chimborazo, Cotopaxi, con sexo mujer, arma blanca, edades entre 40 y 65 años y el hogar.

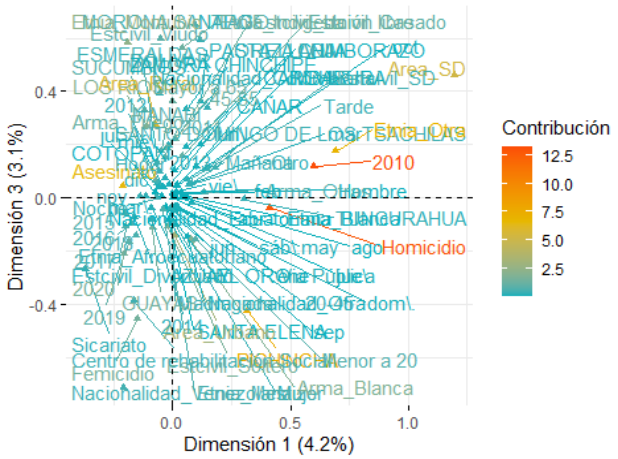

Figura 7. Representación de las similitudes de las categorías de las variables en los ejes 2-3

La Figura 7 muestra las relaciones entre categorías en las dimensiones 2-3, así, se tiene que, los asesinatos se relacionan con las provincias de la costa. El sicariato con los hombres en edades entre los 20 y 45 años, se produce en la vía pública en etnias Afro ecuatoriano, Montubio y Mestizo. Por otra parte, se tiene que los homicidios relacionan las provincias de Pichincha, Tungurahua, en mujeres de menos de 20 años.

4. CONCLUSIONES

Las relaciones entre variables asociadas a los homicidios intencionales son del tipo temporal cuando y en qué hora se produjo, la ubicación donde se produjo, las relacionadas con las víctimas, como el rango de edad, el sexo, estado civil, nacionalidad, la etnia y las que describen como se produjo este tipo de violencia, donde es muy representativo el arma utilizada, el tipo de delito. De esta manera, se garantiza que, al analizarlas en conjunto con la metodología utilizada, se determina claramente las relaciones significativas estadísticamente entre variables y categorías, lo que, permite generar conocimiento sobre los diferentes factores y correspondencias asociados con este tipo de violencia.

Este tipo de violencia intencional es tratada desde diferentes ámbitos del conocimiento como la Sociología, Criminología, Derecho Penal, la Probabilidad, la Estadística, etc. Con el análisis realizado se determina la relación entre diferentes variables, lo que, permitirá abordar de forma concisa y con validez comprobada los planes, programas y proyectos que se orienten a evitar el cometimiento de este tipo de violencia, desde lo preventivo, lo reactivo y la aplicación de justicia con el objeto de la reparación a las víctimas asociadas con este fenómeno social y criminal.

Este análisis al enmarcarse en técnicas multivariantes, permite determinar que las siguientes variables con sus categorías presentan asociación estadística, estas son: el arma utilizada siendo el arma de fuego la que mayor representatividad tiene, el tipo de delito asesinato, el rango de edad se tiene en jóvenes entre 20 y 45 años, el sexo de la víctima con mayor riesgo son los hombres, las provincias en donde se producen son en las de la costa ecuatoriana como Guayas, El Oro, Manabí, Los Ríos y Esmeraldas, otras variables relacionadas son el rango horario y el día de ocurrencia. De esta forma, de las trece variables iniciales, con el ACM aplicado, se identifican seis variables junto con sus categorías, para describir de forma adecuada el fenómeno analizado.

REFERENCIAS

Abril, M., y Castro, A. (2015). *Victimización en Quito mediante análisis de datos categóricos* (Primera ed.). Editorial Académica Española.

Agresti, A. (2002). *Categorical Data Analysis* (Segunda ed.). John Wiley & Sons, Inc.

Beh, E., y Lombardo, R. (2014). *Correspondence Analysis Theory, Practice and New Strategies* (Primera ed.). WILEY.

Brandt, S. (2014). *Data Analysis* (Cuarta ed.). Springer.

Código Orgánico Integral Penal. (10 de Febrero de 2014). *Código Orgánico Integral Penal.* Suplemento Código Orgánico Penal: https://tbinternet.ohchr.org/Treaties/CEDAW/Shared%20Documents/ECU/INT_CEDAW_ARL_ECU_18950_S.pdf

Curbet, J. (2007). *Conflictos globales Violencias locales* (Primera ed.). FLACSO.

Dammert, M. (2 de Febrero de 2006). *Política y armas.* FLACSO. Retrieved 4 de Septiembre de 2021, from https://repositorio.flacsoandes.edu.ec/handle/10469/2427

Deisenroth, P., Faisal, A., y Ong, C. (2020). *Mathematics for Machine Learning* (Primera ed.). Cambridge University.

Díaz, G., y Morales, M. (2012). *Análisis Estadístico de Datos Categóricos* (Primera ed.). Editorial Universidad Nacional de Colombia.

Fiscalía General del Estado. (5 de Febrero de 2015). *Los delitos en Ecuador una mirada desde las cifras.* https://issuu.com/fiscaliaecuador/docs/libro_fiscalia_horizontal_publicado

Grande, I., y Abascal, E. (1989). *Métodos Multivariantes para la Investigación Comercial* (Primera ed.). Ariel.

Hardle, W., y Simar, L. (2015). *Applied Multivariate Statistical Analysis* (Cuarta ed.). Springer. https://doi.org/10.1007/978-3-662-45171-7

Husson, F., Josse, J., Le, S., y Jeremy, M. (11 de Diciembre de 2020). *The R Project for Statistical Computing.* The R Project for Statistical Computing: https://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf

Husson, F., y Jérôme, S. (2017). *Exploratory Multivariate Analysis by Example Uning R* (Segunda ed.). Chapman & Hall/CRC.

Kaleidos. (30 de Octubre de 2021). *Kaleidos Centro de Etnografía Interdisciplinaria*. https://www.kaleidos.ec/diagnostico-del-sistema-penitenciario-del-ecuador-2021/

Kassambara, A. (1 de Abril de 2020). *The R Project for Statistical Computing.* The R Project for Statistical Computing: https://cran.r-project.org/web/packages/factoextra/factoextra.pdf

Lebart, L., Morineau, A., y Jean, F. (1985). *Tratamiento Estadístico de Datos* (Tercera ed.). Marcombo Boixareu Editores.

Pontón, D., y Rivera, F. (2016). Postneoliberalismo y policía: caso de Ecuador 2007-2013. *Desafíos, 28*(2), 213-253. https://doi.org/http://dx.doi.Org/10.12804/desafios28.2.2016.06

Rivera, F. (2011). *Inteligencia estratégica y Prospectiva* (Primera ed.). FLACSO-Sede Ecuador.

United Nations Office on Drugs and Crime. (2013). *Global Study on Homicide 2013.* Vienna: UNODC. Retrieved 3 de Septiembre de 2021, from https://www.unodc.org/documents/gsh/pdfs/2014_GLOBAL_HOMICIDE_BOOK_web.pdf

Wackerly, D., Mendenhall, W., y Sheaffer, R. (2010). *Estadística Matemática con aplicaciones* (Séptima ed.). Cengage Learning™.

Wacquant, L. (2004). *Las cárceles de la miseria* (Segunda ed.). Manatial.

Wacquant, L. (2007). *Parias Urbanos* (Segunda ed.). Manantial.

Zelterman, D. (2015). *Applied Multivariate Statistics with R.* Springer. https://doi.org/10.1007/978-3-319-14093-3

BIOGRAFÍA

Mauricio Enrique Abril Donoso, Ingeniero Matemático por la Escuela Politécnica Nacional
Magister en Estadística Aplicada por la Escuela Politécnica Nacional, se ha desempeñado como:
Estadístico Observatorio Metropolitano de Seguridad Ciudadana de Quito
Estadístico Ministerio del Interior Plan Nacional de Seguridad
Asesor en Estadística Fiscalía General del Estado
Coordinador de Operaciones Empresa de Pasajeros de Quito
Docente Universitario.
Consultor en Minería de Datos y Modelos Estadísticos, Director Nacional de Estadística y Análisis de Información de Salud en el Ministerio de Salud Pública, Oficial de Seguridad de la Información en el Ministerio de Salud Pública.
.

Nancy Elizabeth Chariguamán Maurisaca, Ingeniera en Estadística Informática (ESPOCH). Magister en Matemática Básica (ESPOCH). Docente-Investigador en la Escuela Superior Politécnica de Chimborazo, actualmente se desempeña como investigadora del grupo de investigación ESTADISMATICA y Coordinadora del grupo de investigación ESTADISMATICA

Johanna Enith Aguilar Reyes, Ingeniera en Estadística Informática (ESPOCH). Magister en Gestión y liderazgo educacional (UTPL). Docente-Investigador en la Escuela Superior Politécnica de Chimborazo, Actualmente investigadora del grupo de investigación MOODELING – ESPOCH.

Identificación de Clusters Espaciales de Empresas y la Influencia de Factores Externos en su Constitución

Jácome, Jorge [1]; **Flores, Miguel** [2]

[1]*Escuela Politécnica Nacional, Facultad de Ciencias, Quito, Ecuador*
[2]*Escuela Politécnica Nacional, Departamento de Matemática, Quito, Ecuador*

Resumen: El presente trabajo se centra en el uso de distintas técnicas de minería de datos, basadas en métodos estadísticos geoespaciales para la identificación de patrones con respecto a las actividades económicas de las empresas, registradas en la Superintendencia de Compañías en el Distrito Metropolitano de Quito, al igual que definir factores externos que influyen en la constitución de estas. Primero, para la creación de los clusters, se utilizan los indicadores locales de asociación espacial (LISA, Local Indicators of Spatial Association), los cuales definirán los barrios que posean alta densidad y se caracteriza a estos con variables auxiliares que describen la locación. Esto también ayuda a identificar potenciales barrios con similares características, pero sin una alta densidad de empresas. Después, se utilizan modelos de regresión espacial, para identificar la relación que existe entre el número de empresas y las variables auxiliares determinadas, analizando el coeficiente asociado a cada una. Finalmente al complementar ambos resultados, se obtiene el listado de los barrios con alta densidad en los cuales se debería trabajar con el factor de crecimiento o disminución de empresas de cada una de las variables auxiliares identificadas.

Palabras clave: Estadística Espacial, Indicadores Locales de Asociación Espacial, Modelos de Regresión Espacial

Identification of Spatial Clusters of Companies and the Influence of External Factors in their Constitution

Abstract: This work focuses on the use of different data mining techniques, based on geospatial statistical methods for the identification of patterns with respect to the economic activities of companies, registered in the Superintendencia de Compañías in the Metropolitan District of Quito, as well as defining external factors that influence the constitution of these. First, to create the clusters, the Local Indicators of Spatial Association (LISA) are used, which will define the neighborhoods that have high density and are characterized with auxiliary variables that describe the location. This also helps to identify potential neighborhoods with similar characteristics, but without a high density of businesses. Then, spatial regression models are used to identify the relationship that exists between the companies and the auxiliary variables found, analyzing the coefficient associated with each one. Finally by complementing both results, the list of high-density neighborhoods in which work should be done is obtained with the factor of growth or decrease of companies for each of the auxiliary variables identified.

Keywords: Spatial Statistics, Local Indicators of Spatial Association, Spatial Regression Models

1. INTRODUCCIÓN

La distribución dentro del territorio nacional se ha vuelto relevante para el estado ecuatoriano. Por este motivo, la Asamblea Nacional del Ecuador (2016) ha desarrollado la Ley Órganica de Ordenamiento Territorial, Uso y Gestión del Suelo; al igual que los disgnósticos y planes de ordenamiento territorial de cada Gobierno Autónomo Descentralizado.

Dentro del Distrito Metropolitano de Quito (DMQ), la Alcaldia Metropolitana de Quito (2015) realiza el Plan de Desarrollo y Ordenamiento Territorial, en el cual se obtiene un diagnóstico de la situación actual del territorio de la ciudad tomando en cuenta varios aspectos como: el económico, social, ambiental y movilidad. En este trabajo, se propone el uso de indicadores espaciales, los cuales ayudan a identificar centros de aglomeraciones utilizando su posición. La variable a utilizar es el número de empresas dentro del DMQ, segun el tipo de actividad económica que ejercen. Al utilizar los indicadores de asociación espacial, se buscan todos los sectores que posean aglomeración de las distintas actividades a las que se dedican las empresas.

Martori y Hoberg (2008) analizaron a la población migrante de España con los indicadores locales de asociación espacial (LISA), clasificando los distintos sectores de Barcelona como alto o bajo en densidad según la vivienda y el país de procedencia. Con

jorgejacome95@outlook.com
Recibido: 14/03/2022
Aceptado: 30/06/2022
Publicado en línea: 23/12/2022
10.33333/rp.vol50n3.05

respecto al ámbito de las empresas y en el contexto nacional, Rojas (2015), en la revista Analitika, define posibles centros de empleo en el DMQ, en relación con la densidad poblacional y de empleo utilizando LISA.

Cid (2011) utiliza modelos de regresión espacial para definir la asistencia escolar, con información del Censo Nacional de Argentina. Estos datos son georeferenciados a nivel de cada departamento, donde se relaciona el valor de la asistencia escolar con variables de los hogares como: nivel educativo del jefe de hogar, nivel de desocupación, cantidad de jóvenes, etc. Para encontrar la influencia de estas variables, se realizan varios modelos de regresión incluyendo modelos espaciales.

La importancia del análisis de los factores: económico, ambiental, social y movilidad radica en que las empresas coexisten en un espacio fijo y limitado con todos estos. En este trabajo, se proponen dos cosa. Primero, el uso de indicadores locales de asociación espacial para definir aglomeraciones de empresas. Segundo, el uso de modelos de regresión espacial y el análisis sobre los coeficientes asociados a cada una de las variables. Así se determinan las distintas interacciones de los factores y la densidad de las empresas, esto permitiría tomar mejores direcciones dentro del análisis de las políticas públicas de ordenamiento territorial en el DMQ.

2. METODOLOGÍA

En esta sección, se plantea el caso de estudio al igual que las fuentes de información de los datos utilizados. Al igual se describirán las partes principales de la metodología para la aplicación del análisis espacial.

2.1 Caso de Estudio

El foco de estudio del presente trabajo abarca el ámbito económico dentro del ordenamiento territorial del DMQ, tomando como eje principal a las empresas. Esta información proviene de la Superintendencia de Compañias (2021), la cual es el organismo técnico, con autonomía administrativa y económica, que vigila y controla la organización, actividades, funcionamiento, disolución y liquidación de las compañías y otras entidades en las circunstancias y condiciones establecidas por la Ley. Esta institución lleva un directorio actualizado de forma anual de todas las compañías del país. Hay que recalcar que no todas las empresas deben reportar su información a la superintendencia, más bien solo aquellas que poseen una constitución más estable como sociedad anónima, laboral, etc.

Estos datos pasaron por un proceso previo de georreferenciación para así obtener su ubicación dentro de la ciudad. Con este proceso se obtuvieron 16 117 empresas, distribuidas de la siguiente manera: 24,19 % empresas de comercio, 19,22 % empresas científico técnicas, 8,93 % empresas administrativas, 7,64 % empresas de transporte, 7,54 % empresas manufactureras, 7,20 % empresas de información, 5,86 % empresas de construcción y el resto se agrupa en un sola categoría con 19,42 %. Se definen las 7 actividades económicas con mayor número de empresas dentro de la ciudad, descritas anteriormente. Esto se lo hace para concentrar el estudio, y así descartar actividades que no tienen representatividad. Todas estas empresas restantes se las asocia en Otros.

Al DMQ se lo divide en todos los barrios dentro de la zona urbana. Esta distribución se obtiene de la pagina del Gobierno Abierto de Quito (2021), el cual es un portal que sigue la iniciativa global, que mediante el uso de la tecnología, busca transparentar la gestión de los entes gubernamentales conectando toda la información que se trabaje con el público y también recibir información de estos. Los barrios a tomar en cuenta son 521, sobre los cuales se disponen las 16 mil empresas.

Para observar la influencia que tienen los demás factores sobre las empresas, se utiliza información proporcionada por varios entes gubernamentales que sean de índole espacial. Así en el portal del Gobierno Abierto de Quito también se identifica información de áreas verdes, paradas de buses y Unidades de Policía Comunitaria (UPC) en el DMQ con corte de información del año 2014.

El Sistema Nacional de Información (2021) es el conjunto de varios actores de los cuales se accede, recoge, almacena y transforma sus datos en información accesible para entes gubernamentales y público en general. Este portal almacena información espacial de varios ministerios del país. Los datos que se usan se refieren a la locación de los centros educativos y centros de salud del país con corte al año 2014.

El Instituto Nacional de Estadísticas y Censos (2021) es la entidad encargada de planificar, normar y certificar la producción del Sistema Estadístico Nacional. Por ser la institución encargada de las estadísticas del país, esta posee gran cantidad de información la cual también es de índole espacial. Los datos encontrados son de población del año 2010 y de edificios del año 2014.

2.2 Análisis Espacial

El concepto básico en el que se basa la estadística espacial es aquel de la dependencia o autocorrelación espacial, que analiza los efectos en el cambio de una variable según su localización (Celemín, 2011). Los primeros índices formales para detectar la presencia de autocorrelación espacial se deben a Moran (1948) y Geary (1954). Este concepto se empezó a utilizar en distintas áreas, como: ciencias sociales, econometría, geografía y medicina.

Otro punto a tratar es el relacionado con los clusters o patrones espaciales, que son un conjunto o zona en la cual se albergan objetos que poseen cierta característica en común. Al realizar un análisis de autocorrelación espacial se puede descubrir si una variable proviene de una distribución aleatoria, o si existe una asociación de valores similares en una locación. De este modo, si los valores altos o bajos tienden a agruparse indica una autocorrelación espacial positiva, por otro lado, si los valores obtenidos son distintos entonces la autocorrelación es negativa. Si no se presenta ninguno de los dos casos no existe autocorrelación espacial (Anselin, 1995).

Si se encuentra autocorrelación espacial, se realiza el análisis

de forma local, donde se calcula un indicador para cada zona, tomando en cuenta la información cercana y su alrededor. De esta manera, el mapa se divide en diferentes zonas, aquellas que poseen autocorrelación espacial y otras que no. El indicador de autocorrelación que se utilizará será el Indicador Local de Moran, que es un indicador local de asociación espacial, conocido como Local Indicators of Spatial Association por sus siglas en inglés LISA (Anselin, 1995).

Al utilizar un LISA se identifica los sectores por la densidad que poseen y además si pertenece a un nivel alto o bajo, dependiendo de la variable en cuestión. De este modo, si se toman todas aquellas que están en un nivel alto o bajo, se encuentran patrones espaciales o clusters (Anselin, 1995).

La regresión es el conjunto de técnicas usadas para explorar y cuantificar la relación de dependencia entre una variable cuantitativa llamada dependiente o respuesta y una o más variables independientes llamadas predictoras. El tipo de regresión más conocido es el lineal, ya sea simple o múltiple. Esta regresión es posible aplicarla sobre datos espaciales, pero, si se demuestra que existe una autocorrelación espacial sobre las variables utilizadas en la regresión, se obtienen resultados atípicos, donde este comportamiento se presentaría dentro del término del error. Esto se puede identificar aplicando el índice de Moran sobre el error, encontrando si este tiene autocorrelación espacial. Con lo cual se puede aplicar modelos de regresión espacial.

Los modelos de regresión espacial son variaciones al modelo de regresión lineal que permiten describir la influencia del espacio que puede presentarse dentro de la variable dependiente, independiente o el término del error. Cuando la dependencia espacial se encuentra en la variable dependiente se denomina modelos de retardo espacial, por otro lado, si esta dependencia se presenta en los residuos son llamados modelos de error espacial. También, existen modelos de orden superior que combinan ambos aspectos dentro de la dependencia espacial. Otro modelo es el de Durbin, el cual se utiliza cuando la dependencia espacial se encuentra dentro de la variable dependiente, ocasionada por la influencia de las variables independientes (Anselin, 2003).

Con cualquiera de estos modelos se puede encontrar la influencia de cada variable independiente, al igual que la influencia que posee el espacio sobre la variable dependiente (Anselin, 2003).

Los modelos de regresión espacial no puede ser estimado por mínimos cuadrados. Por este motivo, se utilizan distintos métodos como estimación por maxima verosimilitud o por variables instrumentales. Cualquiera de estos dos métodos tienen sus respectivas hipótesis para estar correctamente especificado y son diferentes entre cada una.

Para este trabajo, se calcula el índice de Moran global y se determina si el espacio es influyente sobre las empresas. Para este proceso se define una matriz de vecinos que ayude a identificarlos, se utiliza la matriz de distancias inversas considerando que los centroides de cada barrio tengan un máximo de 2 kilómetros de separación. A continuación se ajustaría los modelos de la siguiente manera: primero se ajusta el modelo de regresión lineal y se valida que el espacio influye en los residuos. A continuación se ajusta por máxima verosimilitud los tres modelos. Al final se ajusta por variables instrumentales y en caso de detectar heterocedasticidad utilizar ajustes para corregirlo y así obtener un modelo correcto.

3. Materiales y Métodos

En esta sección se explica el fundamento teórico de los indicadores de autocorrelación espacial y los modelos de regresión espacial.

3.1 Índices de Autocorrelación Espacial

Los índices de autocorrelación se dividen entre globales y locales. El global corresponde a calcular el índice utilizando todas las áreas de estudio, mientras que el local evalúa puntualmente un área con sus vecinos. Para poder entender dichos índices, es necesario definir la autocorrelación espacial.

3.1.1 Autocorrelación espacial

De forma general, la autocorrelación espacial se refiere a la relación entre los objetos, actividades o variables de estudio en alguna locación con otros objetos ubicados cerca, buscando una diferencia o similitud. Goodchild (1986) también menciona que son técnicas que relacionan la información de la locación con el atributo en cuestión.

Existen varios métodos para poder estimar la autocorrelación espacial, habitualmente se utiliza el índice de Moran para definir el valor de autocorrelación sobre un conjunto de datos espaciales. Para calcular el índice se debe definir una forma de medir la cercanía dentro de las distintas locaciones, para lo cual se utilizan matrices de contigüidad y de distancias.

3.1.2 Matriz de Contigüidad y Matriz de Distancias

Estas matrices se construyen de la siguiente forma: al tener n locaciones se construye una matriz de tamaño nxn en la cual cada entrada representa si existe contigüidad o la distancia entre las locaciones i j, con i, j entre 1 y n, dependiendo del tipo de matriz. Por la forma de construcción, estas matrices poseen las siguientes propiedades:

- La diagonal principal contiene ceros, esto implica que un área no es vecina consigo misma.
- La matriz es simétrica.
- El número de vecinos de cada área está determinado por los valores distintos de ceros de cada fila de la matriz.

Las matrices de contigüidad tienen distintos tipos, dependiendo de la cantidad de conexiones que posean, sus nombres provienen del ajedrez y los movimientos de las piezas. Es así que, contigüidad tipo Torre son aquellos que únicamente se conectan por el borde. En cambio, aquellos que se conectan únicamente por el vértice, se denominan contigüidad tipo Alfil. Y contigüidad tipo Reina son

aquellos que toman todo tipo de conexiones (Celemín, 2011).

En las matrices de distancia, se considera una distancia máxima entre locaciones de donde se obtiene la información, con el fin de localizar los vecinos. Una variante a esta matriz es utilizar la distancia inversa dentro de cada entrada, para definir mejor el efecto que se quiera tratar en el espacio. Tendrá mayor influencia mientras más lejos se encuentre (matriz de distancias) o mayor influencia mientras más cerca se encuentre (matriz de distancias inversa). De aquí en adelante a las matrices de contigüidad y de distancia serán tratadas como Matrices de Vecinos.

3.1.3 Indicador Global

Indicador de Moran El indicador o índice de Moran es una medida proveniente de los años 1950 (Celemín, 2011), la cual a pesar de ser antigua se sigue manteniendo como una de las más utilizadas para el cálculo de autocorrelación espacial.

La medida I de Moran viene dada por la siguiente expresión:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{1}$$

donde: n es el número de áreas que se consideran; W_{ij} es cada una de las entradas de una matriz de vecinos; x_i el valor de la variable X del área i, y $\bar{x}$ la media aritmética de todos los valores de las áreas. Una interpretación al índice es el coeficiente de pearson utilizando una matriz de pesos.

En párrafos anteriores, se definieron algunos tipos de matrices de vecinos donde cada una puede dar resultados distintos al aplicarlo en el índice. En el trabajo de Rojas (2015), se experimenta con estas matrices dando mayor consistencia con la matriz de distancias inversas en un radio de 2km de interrelación entre sectores.

El dominio del índice es $[-1,+1]$ en donde se considera que si: el indicador es cero, entonces no posee autocorrelación espacial, si el indicador es menor que cero posee autocorrelación negativa y si es mayor que cero posee autocorrelación positiva. Estas interpretaciones de autocorrelación son los siguientes:

- Autocorrelación espacial positiva: los valores de las áreas vecinas son similares. Indica agrupación de las unidades espaciales.
- Autocorrelación espacial negativa: los valores de las áreas vecinas son distintos. Indica una tendencia a la dispersión de las unidades espaciales.
- Sin autocorrelación: no ocurre ninguna de las dos situaciones anteriores. Por lo tanto, los valores de las unidades espaciales vecinas presentan valores producidos de forma aleatoria.

En el caso de que el indicador es cero, se afirma que no posee autocorrelación espacial, y es necesario validar esta afirmación. Para hacerlo se trabaja con pruebas de hipótesis y p-valor; logrando definir si es distinto de cero o no.

Prueba de hipótesis y p-valor La metodología a utilizar se la denomina como Aleatorización y Permutación (Celemín, 2011), que consiste en generar varias muestras donde los datos permutan entre las distintas locaciones, obteniendo el indicador de cada muestra y compararlo con el valor original, con el objetivo de que en cada área pueda ocurrir cualquier valor, dejando que el espacio no tenga influencia sobre este. Con esto la hipótesis nula y alternativa son las siguientes:
H_0 : no autocorrelación espacial
H_1 : autocorrelación espacial

Bajo la hipótesis nula, se obtiene que la esperanza del índice de Moran I es:

$$E(I) = \frac{-1}{(n-1)} \tag{2}$$

siendo n el número de áreas que se consideran. Y su varianza es:

$$Var(I) = \frac{(n[(n^2-3n+3)S_1 - nS_2 + 3S_0^2])}{(n-1)(n-2)(n-3)S_0^2} - \frac{k[(n^2-n)S_1 - 2nS_2 + 6S_0^2]}{(n-1)(n-2)(n-3)S_0^2} - 2E(I) \tag{3}$$

siendo:
$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}$, la suma de todas las entradas de la matriz de vecinos.
$S_1 = \sum_{i=1}^{n}\sum_{j=1}^{n}(W_{ij}+W_{ji})^2$, donde, si la matriz es simétrica se tiene $S_1 = 2S_0$.
$S_2 = \sum_{i=1}^{n}(W_{i.}+W_{.i})^2$, 2 veces la suma de la i-ésima columna y la i-ésima fila.

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}$$

El p-valor desarrollado sobre este procedimiento, depende de, si el valor original pertenece a la muestra obtenida, el cálculo es el siguiente:

$$p = (R+1)/(M+1) \tag{4}$$

donde R es la cantidad de muestras en las cuales su índice es mayor o igual al índice original y M el total de muestras. Si el p-valor es menor al criterio de rechazo, entonces no se puede aceptar la hipótesis nula (Anselin, 2003).

3.1.4 Indicador Local

Los indicadores locales de asociación espacial (LISA), responden a la idea de trabajar en un área específica. El indicador explica los grupos o clusters de valores similares alrededor del área. Las sumas de los indicadores locales de todas las áreas deben describir el indicador global. El indicador local de Moran se obtiene al fragmentar el indicador global a una única área.

Indicador local de Moran

$$I_i = \frac{(x_i-\bar{x})}{v^2}\sum_{j=1}^{n} W_{ij}(x_j-\bar{x}) \tag{5}$$

donde: v^2 es la varianza x_i y sus áreas vecinas; n definido únicamente para todas las áreas vecinas; W_{ij} es cada una de las entradas de una matriz de vecinos; para todo i de 1 hasta n.

Este indicador posee la misma interpretación que el indicador global, por ende, la definición de autocorrelación y la metodología de estimación para el p-valor se mantiene.

Dado que se obtiene un índice sobre cada área, la información resultante es extensa. De modo que usar una herramienta que permita graficar esta información ayuda a definir el comportamiento en cada área. Anselin (2019) presenta el gráfico de dispersión de Moran (ver Figura 1), este se divide en cuatro cuadrantes comenzando por el primero en la parte superior derecha y siguiendo en sentido de las agujas del reloj. En el eje de las x aparecen los valores estandarizados[1] de la variable para cada área, y en el eje y los valores estandarizados del promedio de las unidades vecinas de la misma (Celemín, 2011).

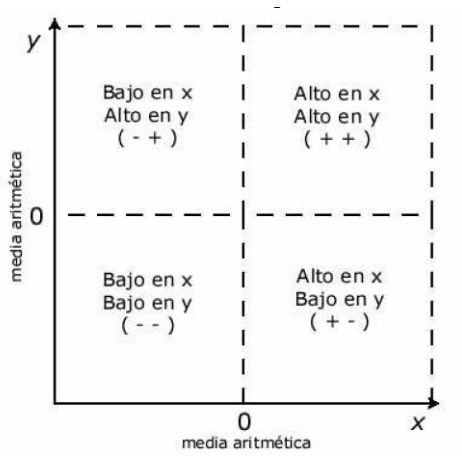

Figura 1. Gráfico de dispersión de Moran
Fuente: Buzai, 2005

Así, en el primer cuadrante se definen las áreas que poseen valores superiores a la media (eje x) y valores superiores a la media en las áreas vecinas (eje y), estos se denominan alto-alto o hotspots. La situación inversa se encuentra en el tercer cuadrante denominándose bajo-bajo o coldspots. En los cuadrantes restantes, se encuentran las áreas donde su valor es distinto al de sus vecinos, ubicando aquí outliers y denominándolos ya sea alto-bajo en el segundo cuadrante y bajo-alto en el cuarto cuadrante (Celemín, 2011).

Los valores que son significativos de acuerdo al p-valor del indicador local, describen las áreas que pertenecen a los clusters alto y bajo, al igual de aquellas no significativas y atípicas.

3.2 Modelos de Regresión Espacial

En esta sección, se describen los modelos de regresión aplicables sobre datos espaciales. Estos se generan a partir del modelo de regresión lineal, para aplicar las respectivas correcciones y la forma de estimación de los parámetros.

[1]Para estandarizar, a cada valor se le resta la media y se le divide para la varianza

3.2.1 Modelo de regresión lineal

Se parte del modelo de regresión lineal, la forma funcional del modelo es:

$$Y = \sum_{i=1}^{N} X_i \beta_i + \varepsilon \tag{6}$$

donde: Y son las observaciones de la variable dependiente, X_i las observaciones de la variable independiente i tomando valores desde 1 hasta N posibles variables, β_i es el coeficiente de la regresión que mide la influencia de la variable i sobre la variable dependiente Y, y finalmente ε es el error que se puede dar por efectos no controladas. Reduciendo, se obtiene:

$$Y = X\beta + \varepsilon \tag{7}$$

En el modelo, los valores dentro del componente del error ε_i son variables independientes e idénticamente distribuidas a una distribución normal con media cero y varianza constante σ^2. También se debe validar que cada observación sea independiente de las otras, al igual que no exista relación entre las variables independientes.

El modelo es utilizado para entender la relación entre la variable dependiente y las variables explicativas, y así entender las posibles causas que ocasionan la variación de Y. Esto se puede obtener estimando los valores de β_i a partir de la información asociada, y también se puede predecir valores de Y a partir de valores de las variables independientes.

La interferencia del espacio dentro de los valores de los datos provoca que los resultados con este modelo contengan inconsistencias, debido a que no se obtiene independencia entre los valores de cada observación. Por eso, se realizan ajustes con distintos modelos, dependiendo del componente afectado por el espacio.

Estos modelos ayudan a entender estas violaciones a los supuestos, siempre y cuando, sean ocasionados por el espacio.

3.2.2 Modelo de retardo espacial

Este tipo de modelos se utiliza cuando la dependencia del espacio se localiza en la variable independiente (Anselin, 2003). Estos modelos son extensión del modelo de regresión lineal en la cual la variable dependiente Y depende de los valores de las áreas vecinas. Este modelo de retardos espaciales es llamado modelo autorregresivo espacial (SAR, por sus siglas en ingles), con forma funcional:

$$Y = \rho WY + X\beta + \varepsilon \tag{8}$$

donde: W representa la matriz de vecinos, y el escalar ρ determina el nivel de relación del área con sus vecinos en función de la matriz W cuyo valor debe ser estimado.

Tomando en cuenta a I_n la matriz identidad de tamaño n, el modelo puede ser reducido de la siguiente forma:

$$Y = (I_n - \rho W)^{-1}(X\beta + \varepsilon) \tag{9}$$

Observando esta forma y al considerar que se mantenga el término del error como variables independientes e idénticamente distribuidas a una distribución normal con media cero y varianza constante σ^2, se puede definir su valor esperado mediante la expresión:

$$E[Y] = (I_n - \rho W)^{-1}(X\beta) \tag{10}$$

El término $(I_n - \rho W)^{-1}$ se denomina multiplicador espacial e indica que el valor esperado de cada observación y_i dependerá de una combinación lineal de valores X tomados por observaciones vecinas, escalado por el parámetro de dependencia ρ.

3.2.3 Modelo de error espacial

Este tipo de modelos se utiliza cuando la dependencia del espacio se localiza en el término del error (Anselin, 2003). Esto puede ser causado por la interferencia de alguna variable que no está controlada en el modelo o por la estructura de los vecinos. Como existen varias razones para presentarse dentro del error, también existen varias formas de corregirlo, donde la más común es representarla a través de sus vecinos de la siguiente manera:

$$\varepsilon = \lambda W\varepsilon + \mu \tag{11}$$

donde λ es el parámetro a estimar, y μ el término del error independiente e idénticamente distribuido a una distribución normal con media cero y varianza constante $\sigma^2 I_n$. Reduciendo esta expresión se obtiene:

$$\varepsilon = (I_n - \lambda W)^{-1}\mu \tag{12}$$

Reemplazando este valor en el modelo de regresión lineal se obtiene:

$$Y = X\beta + (I_n - \lambda W)^{-1}\mu \tag{13}$$

Este modelo es denominado error espacial (SEM).

3.2.4 Modelo de Orden Superior

En este modelo, se combina el modelo de retardo y el de error espacial, es decir, la afectación del espacio se da sobre la variable dependiente Y y el término del error (Anselin, 2003). Este modelo, denominado SARMA, se lo describe de la siguiente manera:

$$Y = \rho WY + X\beta + \varepsilon \tag{14}$$

$$\varepsilon = \lambda W\varepsilon + \mu$$

$$\mu \sim N(0, \sigma^2 I_n)$$

Dentro del modelo, la matriz de vecinos W no necesariamente debe ser la misma para las dos partes, por tanto, se utilizan matrices distintas que logren describir mejor cada efecto.

3.2.5 Prueba para Dependencia Espacial

La presencia de dependencia espacial en un modelo de regresión se puede detectar mediante pruebas de diagnóstico. A continuación, se describe la prueba utilizando el test de Moran, para detectar la dependencia espacial en términos de error para los modelos espaciales descritos en la sección anterior (Anselin, 2003).

La forma de utilizar este estadístico es realizando el contraste a la hipótesis de existencia de autocorrelación espacial con el índice de Moran sobre los residuos que se obtengan del modelo de regresión lineal. El índice se puede reescribir de la siguiente forma:

$$I = \frac{n}{W_0}\frac{\varepsilon' W\varepsilon}{\varepsilon'\varepsilon} \tag{15}$$

con $W_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}$ y $\varepsilon'\varepsilon$ la suma de los cuadrados de los residuos. Sobre este valor se aplica todo lo relacionado al índice global de Moran.

El rechazo de la hipótesis nula sobre la aleatoriedad implica la influencia del espacio dentro del modelo, pero no se especifica el componente que lo causa. Por este motivo, se utilizan los distintos modelos con el fin de obtener el mejor resultado posible (Anselin, 2003).

3.2.6 Estimación

Para estimar los parámetros del modelo de regresión lineal se utilizan mínimos cuadrados ordinarios, los cuales buscan reducir al máximo la suma de los cuadrados de los residuos. Dado que es el método más utilizado para resolver estos modelos, el parámetro se estima de la siguiente manera:

$$\beta = [X'X]^{-1}X'Y \tag{16}$$

Debido a las modificaciones realizadas en el modelo de regresión lineal, el método por mínimos cuadrados o el estimador descrito anteriormente, no son viables a utilizar con las nuevas regresiones descritas.

Para estimar los modelos de regresión espacial se utiliza el método de máxima verosimilitud (MV), el cual se basa en maximizar la probabilidad de distribución conjunta con respecto a los parámetros relevantes. Este método tiene propiedades teóricas asintóticas tales como consistencia, eficiencia o normalidad asintótica, y también se considera robusto para pequeñas variaciones de la suposición de normalidad. También se estima utilizando variables instrumentales o más conocidos como estimación de mínimos cuadrados en dos etapas (Anselin, 2003).

Estimación por Máxima Verosimilitud Partiendo del modelo de regresión de retardo espacial, para simplificar este modelo se escribe $A = I_n - \rho W$ obteniendo:

$$Y = A^{-1}X\beta + A^{-1}\varepsilon \tag{17}$$

y considerando que el término del error sigue la distribución:

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

con σ^2 el parámetro de dispersión que también es desconocido. Con estos puntos descritos, se define que Y sigue la distribución:

$$Y \sim N(A^{-1}X\beta, \sigma^2 A^{-1}(A^{-1})') \tag{18}$$

La función de densidad de una normal $N(\mu, \sigma^2)$ es:

$$f(s) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right) \tag{19}$$

así la función de verosimilitud partiendo de la densidad de Y es la siguiente:

$$\begin{aligned} L(Y|\rho,\beta,\sigma^2) &= \frac{1}{\sqrt{2\pi}}|\sigma^2 A^{-1}(A^{-1})'|^{-1/2} \\ &\exp\left(-\frac{1}{2\sigma^2}(Y-A^{-1}X\beta)'A'A(Y-A^{-1}X\beta)\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}|A|\exp\left(-\frac{1}{2\sigma^2}(AY-X\beta)'(AY-X\beta)\right) \end{aligned} \quad (20)$$

La función de log-verosimilitud corresponde a:

$$\begin{aligned} l(Y|\rho,\beta,\sigma^2) = &-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)+\ln|A| \\ &-\frac{1}{2\sigma^2}(AY-X\beta)'(AY-X\beta) \end{aligned} \quad (21)$$

donde los parámetros a los cuales está atada son ρ,β y σ^2. Así, al realizar las distintas derivadas parciales sobre estos términos se obtiene:

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \frac{1}{\sigma^2}\left[X'(AY-X\beta)\right] \\ \frac{\partial l}{\partial \rho} &= \frac{1}{\sigma^2}\left[Y'W(AY-X\beta)\right]-\frac{1}{|A|}\frac{\partial|A|}{\partial\beta} \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\left[(AY-X\beta)'(AY-X\beta)\right] \end{aligned}$$

Con el sistema de ecuaciones obtenido se encuentra que no es lineal, imposibilitando la obtención de una solución analítica. Para resolverlo se procede a dejar dos parámetros en función del tercero, de este modo, solo se debe maximizar un único término y el resto se calcula sobre este término.

$$\begin{aligned} 0 &= \frac{1}{\sigma^2}\left[X'(AY-X\widehat{\beta})\right] \\ 0 &= X'AY-X'X\widehat{\beta} \\ X'X\widehat{\beta} &= X'AY \\ \widehat{\beta} &= (X'X)^{-1}X'AY \\ \widehat{\beta} &= (X'X)^{-1}X'(I_n-\rho W)Y \\ \widehat{\beta} &= (X'X)^{-1}X'Y-\rho(X'X)^{-1}X'WY \\ \widehat{\beta} &= \widehat{\beta}_Y-\rho\widehat{\beta}_{WY} \end{aligned}$$

El parámetro $\widehat{\beta}$ representa la estimación de β la cual depende de $\widehat{\beta}_Y$ y $\widehat{\beta}_{WY}$ son soluciones de los modelos de regresión de Y y WY. Con este resultado, se puede obtener el error estimado mediante $\widehat{\varepsilon}=AY-X\widehat{\beta}$, y así trabajar con el valor de dispersión σ^2.

$$\begin{aligned} 0 &= -\frac{n}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\left[(AY-X\widehat{\beta})'(AY-X\widehat{\beta})\right] \\ 0 &= -\frac{n}{2\widehat{\sigma}^2}+\frac{1}{2\widehat{\sigma}^4}\left[\widehat{\varepsilon}'\widehat{\varepsilon}\right] \\ \widehat{\sigma}^2 &= \frac{\widehat{\varepsilon}'\widehat{\varepsilon}}{n} \\ \widehat{\sigma}^2 &= \frac{(\widehat{\varepsilon}_Y-\rho\widehat{\varepsilon}_{WY})'(\widehat{\varepsilon}_Y-\rho\widehat{\varepsilon}_{WY})}{n} \end{aligned}$$

donde $\widehat{\varepsilon}_Y$ y $\widehat{\varepsilon}_{WY}$ son los residuos resultantes sobre las regresiones de Y y WY. Aplicando estos ajustes, la función de verosimilitud depende de ρ, donde, al resolverlo se tiene la función de verosimilitud concentrada.

$$\ln L_{con}(\rho) = k+\ln|I_n-\rho W|-\frac{n}{2}\ln\left[(\widehat{\varepsilon}_Y-\rho\widehat{\varepsilon}_{WY})'(\widehat{\varepsilon}_Y-\rho\widehat{\varepsilon}_{WY})\right] \quad (22)$$

donde k es una constante que no depende de ρ.

Por último, se utiliza el mejor método para encontrar el valor que maximice la función de máxima verosimilitud concentrada.

Del valor encontrado se debe descartar que ρ es distinto de cero dado que se volvería al modelo de regresión lineal. Para esto se aplica una prueba de hipótesis, en la cual la hipótesis nula es $\rho = 0$, donde el multiplicador de Lagrange asociado es:

$$LM_{SAR} = \frac{\left(\frac{Y'W\widehat{\varepsilon}}{\widehat{\sigma}^2}\right)^2}{\frac{\widehat{\beta}'X'WMWX\widehat{\beta}}{\widehat{\sigma}^2}} \sim \chi^2_{(1)} \quad (23)$$

con $M = I_n - X(X'X)^{-1}X'$. Anselin (2003)

A continuación, se trabaja con los modelos de error espacial, en el cual sus estimadores se obtienen utilizando el mismo proceso mediante estimadores de maxima verosimilitud. Resultando:

$$\widehat{\beta}_{ERR} = \left[X'(I_n-\lambda W)'(I_n-\lambda W)X\right]^{-1}X'(I_n-\lambda W)'(I_n-\lambda W)Y \quad (24)$$

$$\widehat{\sigma}^2_{ERR} = \frac{(\widehat{\varepsilon}-\lambda W\widehat{\varepsilon})'(\widehat{\varepsilon}-\lambda W\widehat{\varepsilon})}{n} \quad (25)$$

De forma similar, se debe descartar que el valor de λ es distinto de cero, de esta forma, se define la hipótesis nula $\lambda = 0$ con el multiplicador de Lagrange:

$$LM_{ERR} = \frac{1}{2S_0}\left(\frac{\widehat{\varepsilon}'W\widehat{\varepsilon}}{\widehat{\sigma}^2_{ERR}}\right)^2 \sim \chi^2_{(1)} \quad (26)$$

siendo $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$

Por último, se obtiene el modelo de orden superior, en el cual se combinan los resultados obtenidos en los modelo de retardo y error espacial. Lo que resta definir es que los valores de ρ y λ son distintos de cero, fijando la hipótesis nula $\rho = 0, \lambda = 0$ con el múltiplicador de Lagrange:

$$LM_{SARMA} = \frac{\left[\left(\frac{Y'W\widehat{\varepsilon}}{\widehat{\sigma}^2}\right)-\left(\frac{\widehat{\varepsilon}'W\widehat{\varepsilon}}{\widehat{\sigma}^2}\right)\right]^2}{\frac{\widehat{\beta}'X'WMWX\widehat{\beta}}{\widehat{\sigma}^2}} + \frac{1}{2S_0}\left(\frac{\widehat{\varepsilon}'W\widehat{\varepsilon}}{\widehat{\sigma}^2}\right)^2 \sim \chi^2_{(2)} \quad (27)$$

Estimación por Variables Instrumentales Partiendo del modelo de regresión de retardo espacial (8), al realizar el siguiente ajuste $Z = [X\ WY]$ y $\theta = [\beta'\ \rho']'$ se transforma el modelo en:

$$Y = Z\theta + \varepsilon \quad (28)$$

con $\varepsilon \sim N(0, \sigma^2 I_n)$.

Dado que las variables independientes WY están correlacionadas con el término de error, no es posible estimar el modelo mediante métodos habituales, así se busca un instrumento H que no esté correlacionado con el error, y a la vez, este fuertemente correlacionado con WY.

Antes de continuar, se debe definir la convergencia en probabilidad. Sea una sucesión de variables aleatorias X_n con $n \in \mathbb{N}$, converge en probabilidad a una variable aleatoria X si se cumple que: para todo ε mayor a 0 se tiene:

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

Para el caso de las matrices se intercambia el valor absoluto por la norma matricial y se lo representa como $plim_{n\to\infty} X_n = X$.
Utilizando la definición anterior, el instrumento H debe cumplir con:

$$plim_{n\to\infty} n^{-1} H'WY = M_{HWY} \tag{29}$$

y

$$plim_{n\to\infty} n^{-1} H'\varepsilon = 0 \tag{30}$$

siendo M_{HWY} una matriz finita no singular (Kelejian y Prucha, 1998).

Del modelo se tiene por (10) que:

$$E[Y] = (I_n - \rho W)^{-1}(X\beta)$$

expresando el término por la siguiente serie:

$$(I_n - \rho W)^{-1} = I + \rho W + \rho^2 W + \rho^3 W + ...$$

se obtiene que:

$$E[Y] = X\beta + \rho WX\beta + \rho^2 WX\beta + \rho^3 WX\beta + ...$$

y

$$E[WY] = WX\beta + \rho W^2 X\beta + \rho^2 W^2 X\beta + \rho^3 W^2 X\beta + ... \tag{31}$$

Por lo anterior, $E[WY]$ se relaciona linealmente con WX, W^2X, W^3X,..., siendo instrumentos para las variables WY que cumple (29). Habitualmente se utiliza los valores de WX y W^2X como instrumentos (Kelejian y Prucha, 1998).

Al aplicar esta idea en el modelo de retardos espaciales se obtiene que el estimador de variables instrumentales (IV) es:

$$\hat{\theta} = (\hat{Z}'Z)\hat{Z}'Y \tag{32}$$

donde: $\hat{Z} = Q_H Z = [X\ \hat{WY}]$, $\hat{WY} = Q_H WY$ y $Q_H = H(H'H)^{-1}H'$ con los instrumentos $H = [X\ WX]$.

El estimador cumple con:

$$\sqrt{n}(\hat{\theta} - \theta) \to^d N\left(0, \sigma^2 plim_{n\to\infty} n[\hat{Z}'\hat{Z}]^{-1}\right)$$

La estimación IV de estos parámetros puede ser obtenida a través de un proceso de mínimos cuadrados en dos etapas, denominado 2SLS por sus siglas en inglés. Este procedimiento se describe en los siguientes pasos.

1.- Estimar la relación entre los instrumentos $H = [X\ WX]$ y $Z = [X\ WY]$ donde $Z = H\delta + \varepsilon$ y $\delta = (H'H)^{-1}H'Z$, tal que los valores predichos del modelo son: $\hat{Z} = H\hat{\delta} = H(H'H)^{-1}H'Z = Q_H Z$ con $\varepsilon \sim (0, \sigma^2 I)$ para cumplir (4.28).

2.- Estimar $Y = Z\theta + \varepsilon$, utilizando los resultados del paso anterior:

$$\hat{\theta} = (\hat{Z}'Z)\hat{Z}'Y$$

Si se asume que el término del error es independiente pero heterocedástico, se modifica la matriz de varianzas-covarianzas asintótica de $\hat{\theta}$, y se obtiene la siguiente forma:

$$(\hat{Z}'\hat{Z})^{-1}\hat{Z}'\hat{\Sigma}\hat{Z}(\hat{Z}'\hat{Z})^{-1}$$

donde $\hat{\Sigma}$ es una matriz diagonal cuyos elementos i-ésimos son $\hat{\varepsilon}_i^2$, con $\hat{\varepsilon}_i = Y_i - Z_i\hat{\theta}$.

Comparativo de las Estimaciones La diferencia entre estas dos alternativas radica en su funcionamiento bajo los supuestos sobre el término del error.

El estimador de máxima verosimilitud produce estimaciones consistentes en el caso de que los residuos sean independientes y estén idénticamente distribuidos a una normal, en el caso de existir heterocedasticidad el estimador se vuelve inconsistente.

En cambio, el estimador por variables instrumentales tiene la ventaja de que utiliza teoría asintótica sin necesidad del supuesto de normalidad de los residuos, y produce estimaciones consistentes, incluso en el caso de que se detecte heterocedasticidad.

3.2.7 Interpretación de Parámetros Estimados

Una virtud de la econometría espacial es la capacidad de adaptarse a estrategias de modelado extendidas que describen interacciones multiregionales. Sin embargo, este rico conjunto de información también aumenta la dificultad de interpretar las estimaciones resultantes.

Los modelos de regresión espacial aprovechan la complicada estructura de dependencia entre observaciones que representan países, regiones, condados, etc. Por ello, las estimaciones de los parámetros contienen una gran cantidad de información sobre las relaciones entre las observaciones. Un cambio en una sola observación asociada con cualquier variable explicativa afectará a la región misma (un impacto directo) y, potencialmente afectará a todas las demás regiones de manera indirecta (un impacto indirecto).

Los impactos se lo clasifican de la siguiente manera:

1. *Impacto directo promedio*. El impacto de los cambios en la *i*-ésima observación ocasionada por la *j*-ésima variable, se denotará por Xij, sobre *yi* podría resumirse midiendo el promedio de los valores de la diagonal del producto entre el multiplicador

espacial $(I_n - \rho W)^{-1}$ y el vector que contenga el coeficiente asociado a la variable $I_n\beta_j$. Se debe tener presente que, promediar el impacto directo asociado con todas las observaciones i, es similar a las interpretaciones típicas de coeficientes de regresión que representan la respuesta promedio de las variables dependientes a independientes sobre la muestra de observaciones.

2. *Impacto total promedio*. La suma de la *i*-ésima fila del producto entre el multiplicador espacial $(I_n - \rho W)^{-1}$ y el vector que contenga el coeficiente asociado a la variable $I_n\beta_j$, representaría el impacto total en la observación individual *yi* resultante de cambiar la *j*-ésima variable explicativa en la misma cantidad en todas las *n* observaciones. Hay *n* de estas sumas, cada una por cada observación, obteniendo el impacto total del promedio de todas las sumas.

3. *Impacto indirecto promedio*. Es la diferencia entre el impacto total promedio y el impacto directo promedio.

Se define $M(j)_{total}$, $M(j)_{directo}$ y $M(j)_{indirecto}$, que representan los impactos totales promedio, los impactos directos promedio y los impactos indirectos promedio de cambios en la variable j.

$$M(j)_{directo} = n^{-1} tr((I_n - \rho W)^{-1} I_n \beta_j) \tag{33}$$

$$M(j)_{total} = n^{-1} 1_n' \left[(I_n - \rho W)^{-1} I_n \beta_j\right] 1_n \tag{34}$$

$$M(j)_{indirecto} = M(j)_{total} - M(j)_{directo} \tag{35}$$

donde 1_n es el vector de unos con tamaño n.

4. RESULTADOS Y DISCUSIÓN

En esta sección, se resumen los resultados obtenidos con la metodología descrita. Se describe la información que se utiliza para cada paso y los resultados obtenidos.

Como punto de partida, se calcula el índice de Moran global y se determina si el espacio es influyente sobre las empresas. El cálculo del índice de Moran se realiza en el programa GeoDa. Se obtiene que el índice es igual a 0,4261 con un p-valor de 0,001. Esto indica que no se puede aceptar la hipótesis nula de aleatoriedad, lo cual demuestra que el espacio es influyente dentro de la distribución de las empresas en los barrios.

El siguiente paso a realizar es el índice local, de forma similar se obtiene este cálculo en GeoDa. Primero se utiliza el gráfico de dispersión de Moran para calificar a los barrios en los distintos clusters, el cual se presenta en la Figura 2. Definido el cluster al que pertenece cada barrio, se utilizan únicamente aquellos que poseen su índice significativo, y así ver la distribución dentro del mapa.

La cantidad de barrios en cada cluster es la siguiente: en Alto-Alto 48 barrios, en Alto-Bajo 5 barrios, en Bajo-Alto 15 barrios, en Bajo-Bajo 288 barrios y no significativos 165 barrios; la locación se muestra en la Figura 3. De los distintos clusters identificados, se utilizan a los barrios marcados como Alto-Alto, a los cuales se les asocia variables sociodemográficas. De este modo se obtiene que los barrios poseen poblacion entre las 42 y 8 353 personas; el

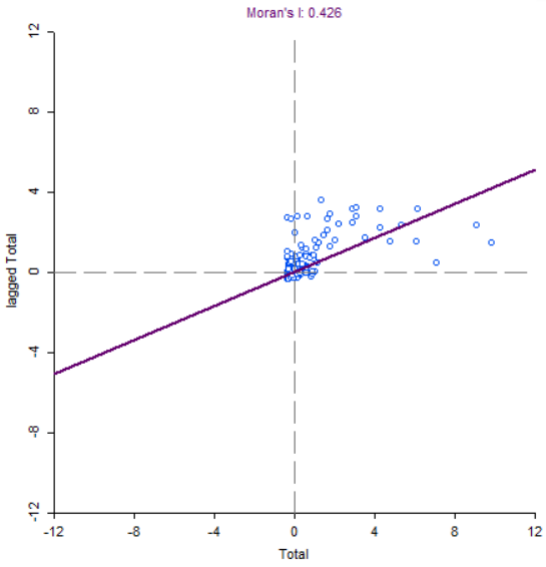

Figura 2. Resultados GeoDa. Gráfico de dispersión de Moran

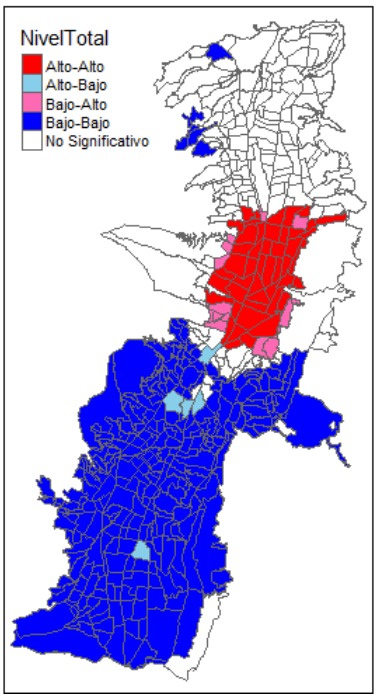

Figura 3. Clusters por total de empresas

área desde los 133 mil al millón 259 mil metros cuadrados; hasta 3 centros educativos; de uno hasta 96 edificios; y hasta 2 unidades de policía comunitaria (UPC).

Se identifican a los barrios que se encuentren dentro de los rangos ya establecidos, por tanto, se obtienen 391 barrios que representan el 75 % del total de la muestra. Esto se debe a que se están tomando todas las empresas sin ningún tipo de filtro.

Por este motivo, se procede a realizar el mismo proceso sobre cada una de las actividades con mayor densidad dentro de Quito, con la diferencia de que se restringe más los datos, donde se toma

como límites el percentil 5 % y 95 %.

Los resultados sobre la cantidad de barrios que fueron definidos con alta densidad al unir toda la información de las actividades económicas se muestran en la Figura 4.

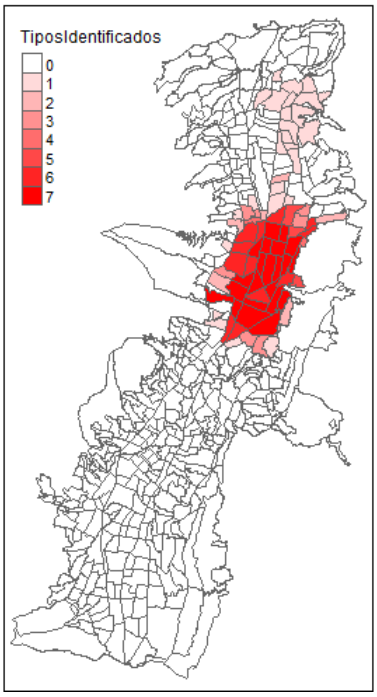

Figura 4. Barrios por actividades con alta densidad

Se considera a los barrios que posean al menos una actividad económica para definir el cluster de barrios con alta densidad de empresas, obteniendo 78 barrios.

Los barrios localizados se ubican en la parte central de la ciudad. Estos parten desde el barrio el Ejido hasta el sector de El Inca. Existe otro foco ubicado al norte de la ciudad, el cual parte desde La Kennedy hasta el terminal de Carcelén. Se muestra la ubicación en la Figura 5.

De igual manera, se trabaja con los barrios que poseen características similares con aquellos con alta densidad. Su distribución se muestra en la Figura 6.

Una vez identificados los barrios con alta densidad, se estudia la relación entre las características sociodemográficas sobre la cantidad de empresas. Para cada una de las variables se analiza la influencia del espacio sobre estas, donde los resultados se resumen en la Tabla 1, los cuales nos indican que las variables poseen autocorrelación espacial debido a que sus valores p son menores al 0.05 y todas son positivas indicando que los valores dentro de los barrios son similares a los de sus vecinos.

Con estas variables y la cantidad de empresas se procede a estimar los modelos descritos anteriormente. Se utiliza el software *R* con la librería *spatialreg* y se ajustan el modelo de regresión lineal; los modelos de regresión espacial, error espacial y de

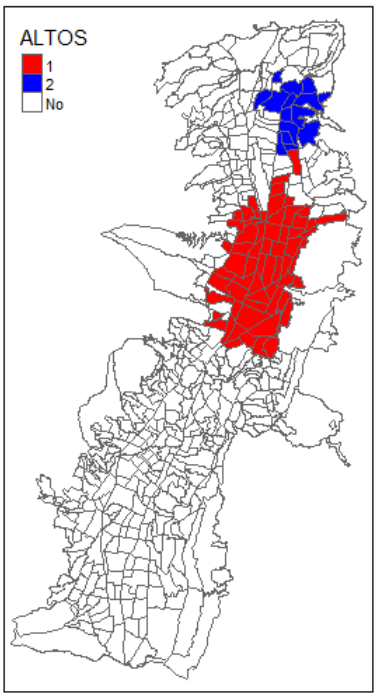

Figura 5. Barrios definidos como altos

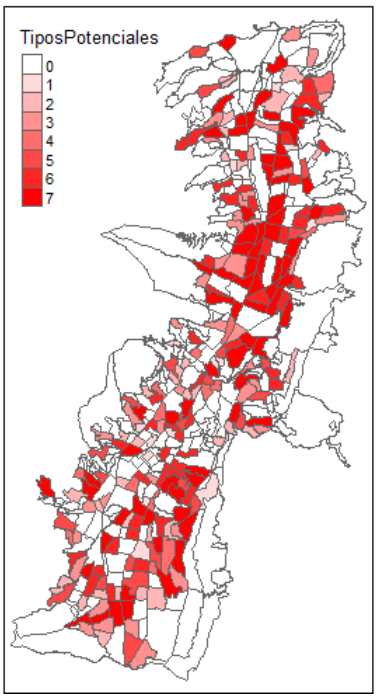

Figura 6. Barrios por actividades con características similares

orden superior por máxima verosimilitud; y el modelo de regresión espacial por variables instrumentales. Para validación de cada modelo se debe tomar en cuenta que los residuos no estén afectados por el espacio, dentro de las estimaciones por máxima verosimilitud los residuos deben ser normales con media cero y varianza constante, por la estimación por variables instrumentales los residuos deben tener media cero y la necesidad de aplicar correcciones frente a la presencia de heterocedasticidad.

Tabla 1. Índices de autocorrelación espacial por variable

Variable	Indicador	Valor p
Población	0,067	0,001
Centros Educativos	0,036	0,009
Centros de Salud	0,045	0,007
Paradas de Buses	0,104	0,001
Edificios	0,177	0,001
UPC	0,051	0,001
Áreas Verdes	0,176	0,001

Tabla 2. Índices de autocorrelación espacial por variable impacto de factores externos sobre el total de empresas

Variable	Directo	Indirecto	Total
Centros Educativos	-7,901	-26,916	-34,817
Edificios	4,148	14,131	18,278
UPC	-13,793	-46,988	-60,781

Con el total de las empresas se trabajó de la siguiente manera: primero se ajustó el modelo de regresión lineal y se valida que el espacio influye en los residuos. A continuación se ajusta por máxima verosimilitud los tres modelos; de los cuales no se puede especificar ninguno dado que no cumplen con la normalidad en sus residuos. Al final se ajusta por variables instrumentales y se detectó heterocedasticidad en los residuos llevando a utilizar el ajuste y así obtener un modelo correcto mostrado en la Figura 7.

Modelo de Regresión Lineal - Total de las Empresas

$$TotalEmpresas \sim Area + Poblacion + CentrosEducativos + Edificios + UPC$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-351.78  -23.08   -5.12    7.12  717.38

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
Area         1.302e-05  3.799e-06   3.428 0.000656 ***
Poblacn     -5.064e-03  1.286e-03  -3.936 9.41e-05 ***
CEdctvs     -9.355e+00  4.111e+00  -2.276 0.023269 *
Edif         5.966e+00  3.522e-01  16.938  < 2e-16 ***
UPC         -1.246e+01  6.163e+00  -2.022 0.043677 *
---

Residual standard error: 72.85 on 516 degrees of freedom
Multiple R-squared:  0.4348,Adjusted R-squared:  0.4294
F-statistic:  79.4 on 5 and 516 DF,  p-value: < 2.2e-16
```

Modelo de Retardos Espaciales (VI) - Total de las Empresas

$$TotalEmpresas \sim CentrosEducativos + Edificios + UPC + 1$$

```
Residuals:
      Min        1Q    Median        3Q       Max
-205.3788  -11.6401    2.1152   11.2195  617.9470

Coefficients:
            Estimate HC0 std. Error z value  Pr(>|z|)
Rho          0.79017        0.11262  7.0163 2.279e-12
(Intercept) -17.39316        4.59781 -3.7829 0.0001550
CEdctvs      -7.30582        3.86766 -1.8890 0.0588985
Edif          3.83538        1.10632  3.4668 0.0005267
UPC         -12.75386        5.00535 -2.5480 0.0108328

Residual variance (sigma squared): 3217.7, (sigma: 56.724)
```

Figura 7. Modelos a todas las Empresas

Estos coeficientes no describen directamente la influencia de las variables sobre el total de las empresas debido a la influencia del espacio. A continuación, en la Tabla 2 se muestran los impactos promedios de las variables.

Se realiza el mismo proceso para todas las actividades obteniendo un modelo correcto para cada una. Se obtiene que el modelo de retardos espaciales estimado por variables instrumentales es el correcto para todas las actividades. Estos se describen a continuación en las Figuras 8 y 9.

Modelo de Retardos Espaciales (VI) - Empresas de Comercio

$$Comercio \sim CentrosEducativos + Edificios + UPC + 1$$

```
Residuals:
      Min        1Q    Median        3Q       Max
-43.422816  -2.758845   0.050823   2.756685 135.582352

Coefficients:
            Estimate HC0 std. Error z value  Pr(>|z|)
Rho          0.80160        0.10331  7.7593 8.438e-15
(Intercept) -3.82655        0.85529 -4.4740 7.678e-06
CEdctvs     -2.04957        0.75134 -2.7279  0.006374
Edif         0.86230        0.19350  4.4563 8.339e-06
UPC         -2.52610        0.99987 -2.5264  0.011522

Residual variance (sigma squared): 138.23, (sigma: 11.757)
```

Modelo de Retardos Espaciales (VI) - Empresas Científico Técnico

$$CTecnico \sim Poblacion + ParadasdeBuses + Edificios + UPC + 1$$

```
Residuals:
      Min        1Q    Median        3Q       Max
-52.49142  -2.89444   0.76013   2.93420 178.83051

Coefficients:
              Estimate HC0 std. Error z value Pr(>|z|)
Rho          0.80982604     0.12997825  6.2305 4.65e-10
(Intercept) -3.41149403     1.07602222 -3.1705 0.001522
Poblacn     -0.00078241     0.00040814 -1.9170 0.055233
PBuses       0.32296556     0.16697490  1.9342 0.053087
Edif         0.76324178     0.29864935  2.5556 0.010599
UPC         -2.92309145     1.35351640 -2.1596 0.030802

Residual variance (sigma squared): 210.13, (sigma: 14.496)
```

Modelo de Retardos Espaciales (VI) - Empresas Administración

$$Administracion \sim Poblacion + ParadasdeBuses + Edificios + UPC + 1$$

```
Residuals:
      Min        1Q    Median        3Q       Max
-26.60923  -1.37029   0.41479   1.44981  70.20100

Coefficients:
              Estimate HC0 std. Error z value  Pr(>|z|)
Rho          0.73216407     0.11307962  6.4748 9.496e-11
(Intercept) -1.68829075     0.52487338 -3.2166  0.001297
Poblacn     -0.00042656     0.00020150 -2.1169  0.034269
PBuses       0.14514609     0.07427475  1.9542  0.050680
Edif         0.41826378     0.15969164  2.6192  0.008814
UPC         -1.33685150     0.63562798 -2.1032  0.035448

Residual variance (sigma squared): 32.32, (sigma: 5.6851)
```

Figura 8. Modelo por Actividad Económica (1)

De estos modelos se obtiene el impacto por cada variable, los cuales se muestran en las Tablas 3 y 4.

Los resultados de estos modelos muestran que los signos de los parámetros son similares a los obtenidos en primera instancia. Así, la cantidad de edificios y las paradas de buses incrementan la cantidad de empresas. Por el contrario, la población, los centros educativos y los UPCs disminuyen la cantidad de empresas. Se debe considerar que en ciertos casos, algunas variables no son signifi-

Modelo de Retardos Espaciales (VI) - Empresas Manufactura

$$Manufactura \sim Poblacion + CentrosEducativos + Edificios + UPC + 1$$

```
Residuals:
      Min        1Q    Median        3Q       Max
-13.518819  -1.284908   0.015475   0.833600  68.109552

Coefficients:
             Estimate HC0 std. Error z value  Pr(>|z|)
Rho          6.8784e-01     1.2467e-01  5.5172 3.445e-08
(Intercept) -9.5429e-01     2.6709e-01 -3.5729  0.000353
Poblacn     -1.5444e-04     6.4975e-05 -2.3770  0.017456
CEdctvs     -6.2957e-01     2.5788e-01 -2.4413  0.014634
Edif         3.2099e-01     3.9995e-02  8.0258 1.110e-15
UPC         -6.1454e-01     2.9195e-01 -2.1049  0.035297

Residual variance (sigma squared): 21.78, (sigma: 4.6669)
```

Modelo de Retardos Espaciales (VI) - Empresas Transporte

$$Transporte \sim Edificios + 1$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.14494  -1.22501  -0.36830   0.52309 108.01010

Coefficients:
             Estimate HC0 std. Error z value  Pr(>|z|)
Rho          0.684470       0.229877  2.9776 0.0029056
(Intercept) -0.795575       0.296854 -2.6800 0.0073617
Edif         0.185807       0.048924  3.7979 0.0001459

Residual variance (sigma squared): 29.497, (sigma: 5.4311)
```

Modelo de Retardos Espaciales (VI) - Empresas Información

$$Informacion \sim Edificios + UPC + 1$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-19.04159  -1.01590   0.24333   0.98758  57.00969

Coefficients:
             Estimate HC0 std. Error z value  Pr(>|z|)
Rho          0.826645       0.146426  5.6455 1.647e-08
(Intercept) -1.316524       0.425845 -3.0916  0.001991
Edif         0.258643       0.087537  2.9547  0.003130
UPC         -1.333262       0.544274 -2.4496  0.014301

Residual variance (sigma squared): 29.993, (sigma: 5.4766)
```

Modelo de Retardos Espaciales (VI) - Empresas Construcción

$$Construccion \sim Poblacion + ParadasdeBuses + Edificios + UPC + 1$$

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.12312  -0.84239   0.12466   0.66450  32.53399

Coefficients:
             Estimate HC0 std. Error z value  Pr(>|z|)
Rho          8.6322e-01     1.2379e-01  6.9733 3.096e-12
(Intercept) -8.0750e-01     1.9500e-01 -4.1410 3.458e-05
Poblacn     -1.4228e-04     7.2856e-05 -1.9530 0.0508250
PBuses       7.4377e-02     3.4280e-02  2.1697 0.0300279
Edif         1.6080e-01     4.5717e-02  3.5172 0.0004361
UPC         -6.2186e-01     2.7170e-01 -2.2888 0.0220904

Residual variance (sigma squared): 10.956, (sigma: 3.31)
```

Figura 9. Modelo por Actividad Económica (2)

cativas dependiendo de la actividad económica.

5. CONCLUSIONES

Se analizan las variables de tipo espacial para comprender su distribución en la ciudad, de lo cual se obtiene que las variables: población, centros educativos, centros de salud, paradas de buses, edificios, unidades de policías comunitarias y áreas verdes son influenciadas por el espacio de forma positiva. La variable que

Tabla 3. Coeficiente del intercepto por actividad económica

Actividad	Intercepto
Comercio	-3,827
Científico Técnico	-3,412
Administrativos	-1,688
Manufacturas	-0,954
Transporte	-0,796
Información	-1,317
Construccion	-0,808

Tabla 4. Impacto total de los factores por actividad económica

	Comercio	C. Tecnico	Administrativo	Manufacturas	Transporte	Información	Construcción
Población		-0,0041	-0,0016	-0,0005			-0,0010
Centros Educativos	-10,3304			-2,0168			
Edificios	4,3462	4,0134	1,5616	1,0283	0,5889	1,4920	1,1756
Paradas de Buses		1,6983	0,5419				0,5438
UPC	-12,7323	-15,3706	-4,9913	-1,9687		-7,6910	-4,5464

explica la cantidad de edificios es la que posee el mayor índice de Moran, el cual al ser positiva nos indica que la cantidad de edificios dentro de un barrio es similar a los de sus vecinos, relacionado por la distancia que los separa. Este fenómeno es el mismo en todas las variables descritas.

Al observar la cantidad de empresas sobre los barrios, se obtiene que son influenciados por el espacio con autocorrelación positiva, es decir, las empresas poseen el mismo efecto que la cantidad de edificios.

Se encontró que en el sector centro norte de la ciudad, partiendo desde el parque El Ejido hasta el sector de El Inca, existe alta densidad de empresas. Esta sección de la ciudad se describe por las actividades de comercio, científico técnicas, administrativas, manufacturas, transporte, información y construcción. Existe otro conjunto de barrios ubicado al norte de la ciudad partiendo desde La Kennedy hasta el terminal de Carcelén, el cual se diferencia del anterior por poseer un incremento en empresas manufactureras.

Al utilizar los modelos de regresión espacial se logró corregir el fenómeno generado por el espacio, en comparación con los resultados de la regresión lineal.

La variable que siempre se toma en cuenta dentro de los modelos es la cantidad de edificios. Estos a pesar de que ocupan determinado espacio en el área de cada barrio, logran ampliarlo al ser construcciones de forma vertical. Es por eso que, en todos los modelos realizados, los edificios resultan ser significativos sobre cada una de las actividades. Donde esta variable siempre posee

signo positivo, indicando mayor influencia sobre las actividades con mayor densidad (comercio y científico técnico).

Otro factor relevante son las paradas de buses. Este tiene asociado un coeficiente con valor positivo, sin embargo no es significativo para todas las actividades. Las paradas de buses tienen relación directa sobre las empresas científico técnicas, administrativas y de construcción.

La población, centros educativos, UPC y el intercepto poseen signos negativos. Todos estos factores tienen en común que cubren grandes áreas como viviendas, centros educativos y unidades de policía comunitaria. También este valor negativo del intercepto indica que, sin ninguno de los factores no se debe tener empresas dentro de los barrios.

En cuanto a las variables área del barrio, centros de salud y áreas verdes, se encuentra que no son significativas para ninguna de las actividades económicas estudiadas.

REFERENCIAS

Asamblea Nacional del Ecuador. (2016, Julio 5). *Ley Orgánica de Ordenamiento Territorial, Uso y Gestión del Suelo.* https://www.sot.gob.ec/sotadmin2/_lib/file/doc/Ley%20Org%C3%A1nica%20de%20Ordenamiento%20Territorial,%20Uso%20y%20Gesti%C3%B3n%20de%20Suelo.pdf

Alcaldia Metropolitana de Quito. (2015, Febrero 22). *Plan Metropolitano de Desarrollo y Ordenamiento Territorial.* https://gobiernoabierto.quito.gob.ec/Archivos/pmdot/PMDOT%202015-2025.pdf

Anselin, L.(2019). *The Moran Scatterplot as an ESDA tool to asses instability in local association.* Spatial Analytical Perspectives on GIS. https://doi.org/10.1201/9780203739051-8

Anselin, L.(1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis*, (27), 93-115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Anselin, L. (2003). *Chapter Fourteen Spatial Econometrics.* In John Wiley & Sons (Ltd.), A Companion to Theoretical Econometrics (pp. 310-330). https://doi.org/10.1002/9780470996249.ch15

Anselin, L. (2003, Junio 15). *Documentación GeoDa.* https://geodacenter.github.io/documentation.html.

Buzai, G.(2005, Septiembre 9). *Los Sistemas de Información Geográfica y sus métodos de análisis en el continuo resolución-integración.* https://www.researchgate.net/publication/299285917_Los_Sistemas_de_Informacion_Geografica_y_sus_metodos_de_analisis_ en_el_continuo_resolucion-integracion

Celemín, J. P. (2011). Autocorrelación Espacial e Indicadores Locales de Asociación Espacial. Importancia, Estructura y Aplicación, (18), 11-31.

Cid, J. C. (2011, Mayo 31). *Aplicación de un modelo de econometría espacial a datos agregados de asistencia escolar en la Argentina.* http://estadisticas.salta.gov.ar/web/archivos/documentos/Aplicacion%20de%20modelo%20de%20econometria%20espacial%20a%20datos%20de%20asistencia%20escolar.pdf

Geary, R.C. (1954). The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, *5*(3), 115- 127+129-146. https://doi.org/10.2307/2986645.

Gobierno Abierto de Quito. (n.d.). *Portal Gobierno Abierto de Quito.* Recuperado Marzo 10, 2021. http://gobiernoabierto.quito.gob.ec/.

Goodchild, M. F. (1986). *Spatial Autocorrelation.* Norwich, United Kingdom: Geo Books https://books.google.com.ec/books?id=2BYnAQAAIAAJ

Instituto Nacional de Estadísticas y Censos (n.d.). *Portal INEC.* Recuperado Marzo 10, 2021. https://www.ecuadorencifras.gob.ec/estadisticas/.

Kelejian, H. H. y Prucha, I. R. (1997). A Generalized Spatial Two Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *The Journal of Real Estate Finance and Economics*, *19*, 99-121. https://doi.org/10.1023/A:1007707430416.

Martori, J. C. y Hoberg, K. (2008). Nuevas Técnicas de Estadística Espacial para la detección de Clusters Residenciales de Población Inmigrante. *Scripta Nova: Revista electrónica de geografía y ciencias sociales, ISSN 1138-9788*, N° 12, 256-265.

Moran, P.A.P. (1948). The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society B (Methodological)*, (10), 243-251. https://doi.org/10.1111/j.2517- 6161.1948.tb00012.x.

Rojas, D. (2015). Localización de centros de empleo y su influencia sobre la distribución de la población en el Distrito Metropolitano de Quito. *Revista de análisis estadístico Analitika*, (9), 55-93. https://www.ecuadorencifras.gob.ec/documentos/web-inec/Revistas/Analitika/Anexos_pdf/Analit_09/3.pdf

Superintendencia de Compañias, Valores y Seguros. (n.d.). *Portal de Información.* Recuperado Marzo 10, 2021. https://www.supercias.gob.ec/portalscvs/.

Sistema Nacional de Información. (2020). *Portal Sistema Nacional de Información.* Recuperado Marzo 10, 2021. https://sni.gob.ec/inicio.

BIOGRAFÍA

Jorge Jácome. Ingeniero Matemático, Mención estadística e Investigación Operativa, de la Escuela Politécnica Nacional (EPN). Graduado del Colegio Nacional Juan Pío Montúfar. Se realizaron pasantías en el Banco Central del Ecuador (BCE) como analista de datos y actualmente, Analista de Datos de Crédito en Banco FINCA. Cuenta con experiencia en implementación de modelos estadísticos y automatización de procesos mediante programación en R con visualización en Power BI.

Miguel Flores. Investigador Postdoctoral de la Universidad Jaume I de Castellón-España en estadística espacio-temporal y Profesor Titular del Departamento de Matemática de la Escuela Politécnica Nacional (EPN). Miembro de grupos de investigación: MODES de Universidad de Coruña; SIGTI y IMPP de EPN. Ingeniero en Estadística Informática de la Escuela Superior Politécnica del Litoral (ESPOL), Magister en Investigación Operativa con mención en Sistemas Logísticos y de Transporte por EPN, Diplomado en Data Mining y Descubrimiento del Conocimiento por la Universidad Iberoamericana Ciudad de México, Ph.D. en Estadística e Investigación de Operaciones y Máster en Técnicas Estadísticas por la Universidad de Coruña.

INDEXACIONES

revistapolitecnica.epn.edu.ec

revistapolitecnica.epn.edu.ec
www.epn.edu.ec

revistapolitecnica.epn.edu.ec
www.epn.edu.ec