EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

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

# Research in Computing Science

Vol. 149 No. 9
September 2020

# Research in Computing Science




***RESEARCH IN COMPUTING SCIENCE***, Año 20, Volumen 149, No. 9, Septiembre del 2020, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: en trámite, otorgado por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 09 de Septiembre de 2020





***RESEARCH IN COMPUTING SCIENCE***, Year 20, Volume 149, No. 9, September 2020, is a monthly publication edited by the National Polytechnic Institute through the Center for Computing Research. Av. Juan de Dios Bátiz S/N, Esq. Miguel Othón de Mendizábal, Nueva Industrial Vallejo, C.P. 07738, Mexico City, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor in charge: Dr. Grigori Sidorov. Reservation of Exclusive Use Rights of Title No. 04-2019-082310242100-203. ISSN: pending, granted by the National Copyright Institute. Responsible for the latest update of this issue: the Computer Research Center, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Last modified on September 9, 2020.





Volume 149(9)

# Advances in Artificial Intelligence

**Oscar Herrera-Alcántara**
**Ma. Lourdes Martínez-Villaseñor (eds.)**

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2020

**ISSN: 1870-4069**


Formerly ISSNs: 1665-9899

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

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



Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

# Table of Contents

Page

ISSN 1870-4069

# Fractal Predictor of the Dynamics of Bitcoin Fluctuations by Using Self-Affine Time Series

Jaime Moreno[1], Oswaldo Morales[1], Ana Lilia Coria[2], Ricardo Tejeida[3], Liliana Chanona[1]

[1] Instituto Politécnico Nacional,
Escuela Superior de Ingeniería Mecánica y Eléctrica,
México

[2] Instituto Politécnico Nacional,
Escuela Superior de Comercio y Administración,
México

[3] Instituto Politécnico Nacional,
Escuela Superior de Turismo,
México

jemoreno@esimez.mx

**Abstract.** The use of cryptocurrency has boomed in recent years, such as Bitcoin, Ethereum or Ripple. It is interesting to have a Bitcoin forecasting tool to try to understand the trends at the global economic level. A virtual currency that can be used as a means of payment just like physical money. Any cryptocurrency uses peer-to-peer technology and is not controlled by any economic or political entity, such as a bank or government. In the year of 2009, Bitcoin was conceived and was priced at 0.39 USD reaching its all-time high in 2017 with a price of 17,549.67 USD, i.e. 45 thousand times more in less than 10 years. This work focuses on predicting the bitcoin-price trending will have in the year 2020 by using a Self-Affine Fractal Analysis as a tool of artificial intelligence. The results that the present work provides in the first 6 months agree with 98% with those actually obtained despite training only with the data from the first days of time series.

**Keywords:** Fractal predictor, Bitcoin, cryptocurrencies, self-affine analysis, fluctuations.

## 1 Introduction

Cryptocurrencies or virtual currencies are offered through the internet globally and are sometimes presented as an alternative to legal tender, although it has very different characteristics:

– It is not mandatory to accept them as a means of paying debts or other obligations.

– Its circulation is very limited.
– Its value fluctuates strongly, so it cannot be considered a good store of value or a stable unit of account.

The strong fluctuations experienced by these cryptocurrencies that seem typical of the classic speculative bubbles are well known. As an example, the average value of bitcoin on the main platforms in which it is traded increased in 2017 from approximately 979.53 USD per unit at the beginning of the year to more than 17,147.04 euros in mid-December. Since then, the trend has been downward. As of February 5, 2018, its price was below 6,902.35 USD, which represents a drop of more than 65% from the December highs. A person who had bought bitcoins in late 2017 and sold them today would suffer very noticeable losses [20].

Additionally, numerous fundraising actions are taking place from investors to finance projects through the so-called initial offers of cryptocurrencies or ICOs. The expression ICO can refer to both the actual issuance of cryptocurrencies and the issue of rights of various kinds, generally called tokens. These assets are put up for sale in exchange for cryptocurrencies such as bitcoins or ethers or official currency (for example, US dollars or USD)[12].

The five main aspects to consider before investing in cryptocurrencies or participating in an ICO:

1. To date, no ICO has been registered, authorized or verified by any supervisory body in Spain. Therefore, there are no cryptocurrencies or tokens issued in ICO whose acquisition or possession worldwide can benefit from any of the guarantees or protections provided in the regulations regarding banking or investment products. Investments in cryptocurrencies or in ICOs outside the regulation are not protected by any mechanism similar to that which protects cash or securities deposited in credit institutions and investment services companies.
2. Before buying this type of digital assets or investing in products related to them, consider all the associated risks and assess whether you have enough information to understand what is being offered. In this type of investment there is a high risk of loss or fraud.
3. Cryptocurrencies lack intrinsic value, becoming highly speculative investments. Furthermore, its strong dependence on poorly consolidated technologies does not exclude the possibility of operational failures and cyber threats that could mean temporary unavailability or, in extreme cases, the total loss of the amounts invested.
4. The absence of markets comparable to organized securities markets subject to regulation can hinder the sale of cryptocurrencies or tokens issued in ICO to obtain conventional cash.
5. In the case of ICOs, the information made available to investors is usually not audited and is often incomplete. Generally, it emphasizes potential benefits, minimizing references to risks.

## 2 Related Work

In recent years a new type of tradable assets appeared, generically known as cryptocurrencies. Among them, the most widespread is Bitcoin. Bariviera et al in [6] compared Bitcoin and standard currencies dynamics, analyzing their returns at different time scales. They investigated the long memory in return time series from 2011 to 2017, using transaction data from one Bitcoin platform. In addition, they computed the Hurst exponent by means of the Detrended Fluctuation Analysis method, measuring long range dependence. They found changes in the Hurst eponent values over the first years of the studied period, tending to stabilize in the last part of the the period. In the aftermath they claim that the Bitcoin market can be described as a self-similar process, displaying persistent beavior until 2014.

Caporale et al in [7] studied the persistence behavior in the cryptocurrency market, applying the R/S analysis and fractional integration long-memory methods and taking as inputs the four main cryptocurrencies (BitCoin, LiteCoin, Ripple, Dash) over the sample period 2013–2017. According to their outcomes, they found that the cryptocurrency market displays persistence or positive correlations between its past and future values. Hence, they claim for a market inefficiency because they did not found random walk behavior (market efficiency) in the cryptocurrency market.

Owing ryptocurrencies have acquired a great development and valorization, Costa et al in [9] analyzed the following four cryptocurrencies, based on their market capitalization and data availability: Bitcoin, Ethereum, Ripple, and Litecoin, using detrended fluctuation analysis and detrended cross-correlation analysis and the respective correlation coefficient. Bitcoin and Ripple seemed to behave as efficient financial assets, while Ethereum and Litecoin displayed some persistence. When authors correlated Bitcoin with the other three cryptocurrencies, at short time scales all the cryptocurrencies had been correlated with Bitcoin, although Ripple had the highest correlations. On the other hand, at higher time scales, Ripple was the only cryptocurrency with significant correlation.

Quintino et al in [19] determined the persistence exhibited by the Bitcoin measured by the Hurst exponent from the Brazilian market daily prices from 9 April 2017 to 30 June 2018, and comparing them with Bitcoin in USD. They used Detrended Fluctuation Analysis, for the period. The analyzed the prices of Bitcoins yielded from negotiations made by two Brazilian financial institutions: Foxbit and Mercado. The authors found that Mercado and Foxbit returns followed Bitcoin dynamics and showed persistent behavior, although the persistence was higher for the Brazilian Bitcoin.

On the other hand, Haque et al in [18] studied forecasts of Bitcoin price using the autoregressive integrated moving average (ARIMA) and neural network autoregression (NNAR) models. The authors forecasted next-day Bitcoin price both with and without re-estimation of the forecast model for each step by using the static forecast approach. For cross-validation of forecast outcomes, they took into account a first training-sample where NNAR performed better

than ARIMA, while ARIMA outperformed NNAR in the second training-sample. Moreover, ARIMA with model re-estimation at each step outperformed NNAR in the two test-sample forecast periods. Hence, they claim the superiority of forecast results of ARIMA model over NNAR in the test-sample periods, and therefore the ARIMA superiority of volatile Bitcoin price prediction.

Deniz and Stengos in [11] analyzed the behaviour of Bitcoin returns and those of several other cryptocurrencies in the pre and post period of the introduction of the Bitcoin futures market, using the principal component-guided sparse regression (PC-LASSO) model. The authors observed the Search intensity as the most important variable for Bitcoin for all periods, whereas for the other cryptocurrencies there were other variables that seemed more important in the pre period, while search intensity still stood out in the post period Furthermore, GARCH analyses suggested that search intensity increased the volatility of Bitcoin returns more in the post period than it does in the pre period. Therefore, the authors asserted that the top five cryptocurrencies were substitutes before the launch of Bitcoin futures.

As we observed, the Bitcoin price-returns dynamics has been characterized applying fractal analysis tools such as the Detrended Fluctuation Analysis and R/S analysis. On the other hand, the Bitcoin price has also has been forecasted by applying econometrics (GARCH) and artificial intelligence (neural network autoregression) tools. In this work, we chacracterized the Bitcoin price fluctuations by calculating the Hurst exponent (H) by estimating the structure function of the time series of Bitcoin price fluctuations in different time interval of the sample. Aditionally, we forecasted the Bitcoin prices applying an artificial intelligence tool. Hence the following sections of this work are the following. In Section 3 the Theoretical Framework is exposed with main features of BItcoin and Fractal Theory. while in Section 4, main Proposal is defined along with methodology as a set of ordered steps. Finally, In Section 5 shows the experimentation methodology as well as the most important results that prove the value of this work.

## 3 Theoretical Framework

### 3.1 BitCoin

Bitcoin is a virtual, independent and decentralized currency, since it is not controlled by any State, financial institution, bank or company. It is an intangible currency, although it can be used as a means of payment just like physical money. Virtual currencies constitute a heterogeneous set of innovative payment instruments that, by definition, lack a physical support to back them up.

The term bitcoin has its origin in 2009, when it was created by Satoshi Nakamoto (pseudonym of its author or authors) [13], who created it with the aim of being used to make purchases only through the Internet. Bitcoin was born with high ambitions: to provide citizens with a means of payment that enables the execution of fast, low-cost transfers of value and which, in addition, does

not can be controlled or manipulated by governments, central banks or financial entities.

Virtual currency uses cryptography to control its creation. The system is programmed to generate a fixed number of bitcoins per unit of time through computers called miners. Currently, that number is fixed at 25 bitcoins every ten minutes, although it is programmed so that it is halved every 4 years. Thus, starting in 2017, 12.55 bitcoins will be issued every ten minutes. Production will continue until 2140, when the limit of 21 million units in circulation is reached [12].

To make use of this virtual currency, it will need to download software to the computer or our mobile that will serve as a *virtual wallet* and will generate a bitcoin address, which can be used to send and receive money from other users. In addition, the sending of bitcoins is instantaneous and all operations can be monitored in real time. Transactions with this currency involve a transfer of value between two bitcoin addresses, public although anonymous. To guarantee security, transactions are secured using a series of key cryptographies, since each account has a public and a private key. As in other virtual currencies, bitcoin also has a number of risks that must be highlighted to know exactly the magnitude of this currency. To identify them, we again resorted to the report of the General Directorate of Operations, Markets and Payment Systems of the Payment Systems Department of the Wold Bank, which groups them into:

- Financing of illegal activities and/or money laundering. Due to the decentralized nature of the scheme, transfers take place directly between the payer and the beneficiary, without the need for an intermediary or administrator. This implies a difficulty in identifying and early warning of possible suspicious behavior of illegal activities.
- The fact that organized crime networks are making widespread use of emerging electronic payment systems can create a negative reputation for digital payment methods.
- Despite the fact that, in principle, any computer can actively participate in the process of creating new bitcoin units, the high computational capacity required implies that, in practice, this activity is dominated by a small group of actors. Possible fraudulent transactions. To the extent that the protocols on which bitcoin is based are open software developments, the implementation of its different versions does not have to occur uniformly among all users.
- Impact on price stability and financial stability, since private trading platforms where Bitcoins can be exchanged for legal tender currencies are marked by the high volatility of prices due to speculative movements.
- From the point of view of fraud, bitcoin presents a significant weakness compared to other payment methods in the online world, such as cards.

### 3.2 Fractal Theory

*Fractals are mathematical objects that generalize Euclidean geometric objects to non-whole dimensions and allow us to delve into the study of complex sys-*

*tems, disorder and chaos* [15]. Fractals refer to any class of phenomena that possess scaling that exhibits dilatation symmetry, or scale invariance, often characterized by the appearance of a power-law. Invariant scale systems are usually characterized by non-integer dimensions. Benoit Mandelbrot developed the fractal geometry to unify a number of previous studies on irregular shapes and natural processes. Hence, fractal geometry is a mathematical tool for dealing with complex systems that do not have a characteristic scale of length, or scale invariance [15].

Mandelbrot also focused on a particular set of such objects and forms where a part of the object is identical to a larger piece, i.e. self-similar objects. According to [15], there are deterministic fractals when a small piece of a fractal is separated and isotropically magnified to the size of the original, both of them look the same. By magnifying isotropically, all the directions have been rescaled by the same factor.

On the other hand, there are systems that are invariant only under *anisotropic* magnifications, which are called *elf-affine fractals*s. If a self-affine curve or time series is invariant in scale under the transformation $x \to bx$, $y \to ay$, it is observed in Equation 1:

$$F(bx) = aF(x) \equiv b^H F(x), \tag{1}$$

where $H = \frac{\log a}{\log b}$ is the exponent of Hurst [15].

Time series are sets of data or records of any observable variable under study. These records are separated by a same time interval, such as seconds, minutes, hours, weeks, months, years, etc. Time series reflect the behavior of a complex system over time.

Kantz and Schreiber [14] proposed an approach to study such systems from their time series fluctuations, in order to characterize their dynamics by means of scaling laws, which are valid over a wide range of time scales and that they are a property of fractals. When carrying out a fractal characterization of time series fluctuations generated by some complex system, what is sought is to find persistent behavior, since this will allow to make probabilistic predictions about the future states of the system based on the value of the scaling exponents obtained for this behavior.

After calculating the $H$ or roughness exponent from Equation 1, the values $H < 1/2$ indicate long-term anti-correlation (or anti-persistence) behavior: if the values of the observable variable increase, most likely the next value is less than the last, and vice versa; values $H > 1/2$ indicate positive (or persistent) long-term correlated behavior: if the values of the observable variable increase, it is most likely that the next value is greater than the last, and vice versa. Finally, for values of $H = 1/2$ do not exist correlations, i.e. a totally random behavior.

# 4 Proposal

## 4.1 Theoretical Definition

Statistically, the fluctuation or volatility of financial time series $p(t)$ are characterized by their standard deviations $v(t,\tau)$ for a sampling time interval $\tau$ considered [8], exhibiting power law correlations, so these complex systems may not respond immediately to a quantity of information that flows towards them, but react gradually in a certain period of time [14,2,4,8]. The analysis of scaling or fractal properties of fluctuations has offered relevant information about the underlying processes responsible for the observed macroscopic behavior of complex systems [1,2,4,5,8].

Moreover, in this paper is studied the long-term correlations displayed by the time series fluctuations by applying their structure function, defined by Equation 2 as follows:

$$\sigma(\tau,\delta_t) = \overline{\left\langle \left[\nu(t+\delta_t,\tau) - \nu(t,\tau)\right]^2 \right\rangle}^{\frac{1}{2}}, \tag{2}$$

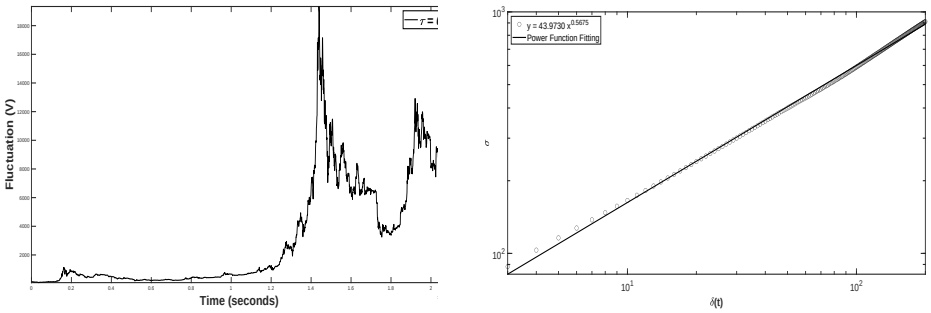
where the upper bar denotes average over all times $t$ in the time series of length $T-\tau$ with $T$ as the length of the original time series $p(t)$ and triangular parentheses denote average over different realizations of the time window of size $\delta_t$ [14]. The structure function of the fluctuations exhibits the power law behavior $\sigma \propto (\delta_t)^H$ with $H$ as the local or roughness exponent, even though the time series fluctuations $\nu(t,\tau)$ exhibit apparently randomness [14,2,4,8]. The scaling behavior $\sigma \propto (\delta_t)^H$ characterizes the correlations in the time series fluctuations treated as a growing interface in a dimension $(1+1)$ [17,10,21].

Accordingly in this paper the structure function was applied to study the dynamics of time series fluctuations associated with the Bitcoin price, by analyzing the behavior of standard deviations for different sampling time intervals. Hence, the time series of standard deviations were treated as interfaces in motion, where the considered sampling time interval $\tau$ plays the role of time variable and the physical time $t$ plays the role of the spatial variable [3].

## 4.2 Methodology

Within artificial intelligence there are systems that think rationally, these try to emulate the logical logical thinking of humans, that is, it investigates how to make machines perceive, reason and act accordingly. The Proposed system tries to reason the average fluctuation to propose future fluctuations and therefore the estimation of the price trend [22].

To study the dynamics of time series fluctuations, in this paper the time series of standard deviations $\nu(t,\tau)$ of the original series $p(t)$ from the open price (op) and close price (cp) of Bitcoin. For this study the length of each original financial series was $T_{op} = 2410$ and $T_{cp} = 2410$ (usd-dollar versus time). In addition, the sampling rate $\triangle t = 1$ business day, with ranges $\tau_m \leq \tau \leq \tau_m$ and rates $(\delta_t)$ from samples of time intervals $(\delta_t)$ from 3 to 200 standard deviations.

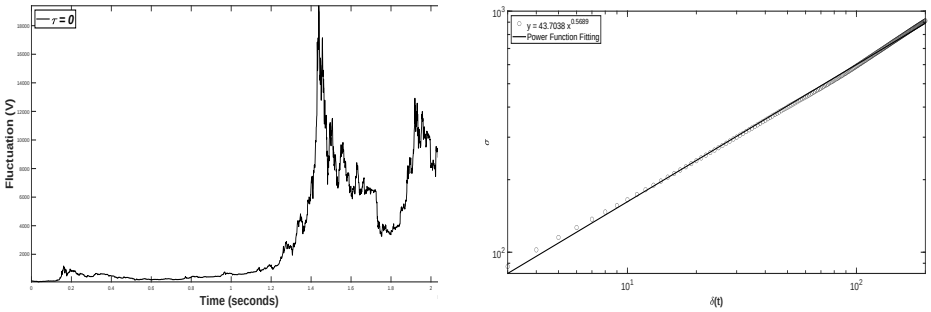



(a) Open (b) Open Fluctuation

**Fig. 1.** (**a**) Original time series of Bitcoin open-price, $p(t)$ (USD), with $T_{op} = 2410$. (**b**) Dynamic scaling of the structure function for the volatility time series with $H_{op} = 0.5675$.

In Figure 1(a) is shown the graph of the original time series, $p(t)$ to the Bitcoin open-price, and in Figure 1(b) is shown the structure function of the the time series of standard deviations $\nu(t, \tau)$ with a value or the Hurst exponent (or slope) equals $H_{op} = 0.5675$.

On the other hand, in Figure 2(a) is shown the graph of the original time series, $p(t)$ to the Bitcoin close-price, and in Figure 2(b) is shown the structure function of the the time series of standard deviations $\nu(t, \tau)$ with a value or the Hurst exponent (or slope) equals $H_{cp} = 0.5689$.




(a) Close (b) Close Fluctuation

**Fig. 2.** (**a**) Original time series of Bitcoin close-price, $p(t)$ (USD), with $T_{op} = 2410$. (**b**) Dynamic scaling of the structure function for the volatility time series with $H_{cp} = 0.5689$.

Algorithm to determine fractal behavior of the volatility time series is the following steps:

1. Collect time series with at least 2000 data of past tendency about Bitcoin open-price and Bitcoin close-price and train the system with the H findings found.

2. Construct one hundred ninety-eight time series of standard deviations (fluctuations or volatility) for every original time series, applying the equation of the standard deviation, for different sample intervals: $3 \leq \tau \leq 200$. Then, construct 198 time series $\nu(t, \tau)$ for both samples.
3. Determine the type of correlations displayed by the time series of fluctuations $\nu(t, \tau)$, applying the Equation 2 of the structure function for predicting the future tendencies.
4. If the fluctuation structure function, obtained in Step 3, exhibits a power law behavior $\sigma \propto (\delta_t)^H$, then, from the same equation, obtain the dynamic exponent of roughness or local $H$, in order to determine if the system displays positive correlations (persistence) in the long-term and to establish at what time the fluctuations move from a power-law behavior to one of saturation (horizontal curve or zero slope). This is done with the purpose of establishing if the behavior of the Bitcoin price fluctuations over the time can be described in the Family-Vicsek fractal model $w(L, t) \sim L^H f\left(\frac{t}{L^{\frac{H}{\beta}}}\right)$, which represents the dynamic scaling behavior of self-affine surfaces in motion.
5. Values of the dynamic scaling exponent $H$ are obtained for both samples. If, based on the values of the exponents $H$, the fluctuation behavior is not fitted to the behavior described by the Family-Vicsek model, look for another model that explains and predicts the behavior of fluctuations for both the Bitcoin open-price and the Bitcoin close-price.
6. Finally, trend findings are projected and the future trend of Bitcoin prices is proposed through a power law in a logarithmic space.

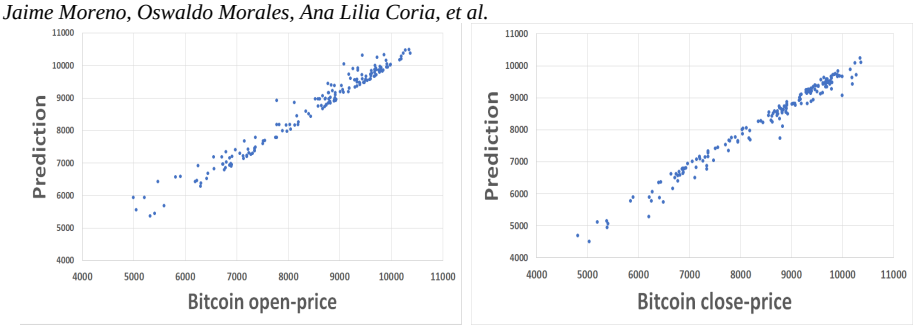
## 5 Experimental Results

### 5.1 Results

The database of the opening and closing prices of the bitcoin-USD price was downloaded from the Markets Insider site, from January 27, 2013 to June 20, 2020 [16]. Then, The original series was splitted into two parts from May 27, 2013 to December 31, 2019, with the aim of predicting the first half of 2020.

Quantitatively, the self-affinity of the time series of Bitcoin price fluctuations was characterized by the scaling behavior $\sigma \propto (\delta_t)^H$, as shown in Figures 2 and 1. The structure function displays a power law $\sigma \propto (\delta_t)^H$ with $H(\tau) = const$ within a range of intervals $\delta_t$. In the Figures 1(b) and 2(b) the graphs of the dynamic scaling of the structure function for the Bitcoin open-price fluctuations and the Bitcoin open-price fluctuations, respectively, we observe the following values: $H_{op} = 0.5672$ and $H_{cp} = 0.5689$. It means that fluctuations for both Bitcoin open-price and Bitcoin close-price display and persistent behavior. Therefore, the dynamics of both Bitcoin prices are fitted to the power-law behavior ranging much more time scales (scale-invariance), and hence claiming a better approximation to the F-V model.

Figures 3(a) and 3(b) show the scatter plot of the actual price of the bitcoin versus the predicted one, respectively. The correlation of these data show us that




(a) Open (b) Close

**Fig. 3.** Actual Bitcoin price vs Predicted Bitcoin price, $p(t)$ (USD), with $T_{op} = 172$. (**b**) Open and (**a**) Close.

the prediction of the opening of the bitcoin price has an effectiveness of 98.48%, while for the close of the trading day of 98.74%.

### 5.2 Discussion

According to the values of $H$ for the Bitcoin open-price and the Bitcoin close-price, it could be pointed out the existence of a dynamic scaling behavior similar to the dynamic scaling of Family-Vicsek $(F - V)$ for the roughness kinetics of a moving interface [17], see Figures 1(b) and 2(b). The scaling relation $\sigma \propto (\delta_t)^H$ implies that the structure function of the time series of fluctuations displays the dynamic scaling behavior $\sigma(\tau, \delta_t) \propto \tau^{\beta} f\left(\frac{\delta_t}{\tau^{\frac{\beta}{H}}}\right)$, where the scaling function behaves like $f(u) \propto u^H$ if $u < 1$ and like $f(u) \propto$ const if $u \gg 1$. That is, the dynamic scaling relations of $F-V$, expressed in power laws, with critical scaling exponents that reflect the scale invariance of the time series fluctuations of both Bitcoin prices. Finally, it should be noted that, because the values of the scaling exponents $H_{op}$ and $H_{cp}$ are greater than 0.5, in the critical area the fluctuations of Bitcoin prices display positive long-term correlations, i.e. persistent behavior.

From Figure 4, the power law increment in the width transitates to a saturation regime (horizontal region) during which the width reaches a saturation value, $w_{sat}$. As $L$ grows, $w_{sat}$ increases as well, and the dependency likewise follows a power law, $w_{sat}(L) \sim L^H$ with $[t \ll t_x]$. The exponent $H$, *the roughness exponent*, characterizes the degree of roughness of the saturated interface.

For small $u$, the scaling function is increased as a power law. In this regime we have $f(u) \sim u^H$ with $[u \ll u_x]$.

As $t \to \infty$, the width saturates. Saturation is reached for $t \ll t_x$, that is $u \gg 1$. In this limit $f(u) =$ constant with $[u \gg 1]$.

The saturation time $t_x$, with the saturation width, $w_{sat}$, increases with the size of the system, which suggests that the saturation phenomenon constitutes a finite size effect. This leads us to affirm that you can predict what the fluctuations will be and therefore the opening and closing price of bitcoin with respect to the USD, our results indicate that there is a correlation greater than 98%.

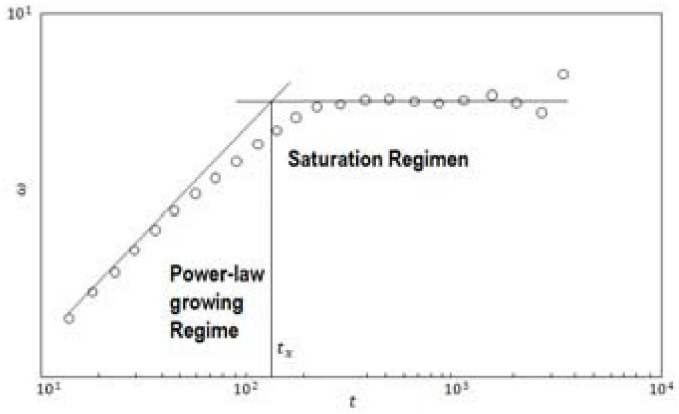



**Fig. 4.** Growth of the interface width time for the BD model for a horizontal-sized system [5].

## 6 Conclusions

The dynamic fluctuations of Bitcoin open-price and Bitcoin close-price exhibited persistent behavior. Hence it can be modeled by the family-Vicsek model.

Moreover, the dynamic scaling of $\sigma(\tau, \delta_t) \propto \tau^{\beta} f\left(\frac{\delta_t}{\tau^{\frac{\beta}{H}}}\right)$, $f(u) \propto u^H$ if $u < 1$ and $f(u) \propto$ const if $u \gg 1$ allowed to treat the dynamics of fluctuations of both Bitcoin prices as a kinetic roughening of a moving interface. The findings of this work point out to the existence of a dynamic scaling behavior similar to the dynamic scaling of Family-Vicsek for the kinetic roughening of a moving interface. Therefore, the kinetic roughening theory tools can be used to characterize and model the time series fluctuations of Bitcoin price.

This work is a novel algorithm that can predict what the fluctuations will be and therefore the opening and closing price of bitcoin with respect to the USD can be estimated. Our results indicate that there is a correlation greater than 98 %, which means that it has high reliability.

## References

1. Ashkenazy, Y., Ivanov, P. C., Havlin, S., Peng, C.-K., Goldberger, A., Stanley, H.: Correlaciones de magnitud y signo en fluctuaciones de latido del corazón. Phys. Rev. Lett., vol. 86, no. 9, pp. 1900–1903 (2001)
2. Balankin, A., Matamoros, O., Gálvez, E., Pérez, A.: Crossover from antipersistent to persistent behavior in time series possessing the generalyzed dynamic scaling law. Physical Review E, vol. 69, pp. 036121 (2004)
3. Balankin, A. S.: Dynamic scaling approach to study time series fluctuations. Physical . Rev. E, vol. 76, pp. 056120 (2007)

4. Balankin, A. S., García Paredes, R., Susarrey, O., Morales, D., Castrejon, F.: Kinetic roughening and pinning of two coupled interfaces in disordered media. Physical Review Letter, vol. 96, no. 5-10, pp. 101–104 (2006)
5. Barabasi, A. L., Stanley, H.: Fractal Concepts in Surface Growth. Cambridge University Press (1995)
6. Bariviera, A. F., Basgall, M., Hasperué, W., Naiouf, M.: Some stylized facts of the bitcoin market. Physica A: Statistical Mechanics and its Applications, vol. 484 (2017)
7. Caporale, G. M., Gil-Alana, L., Plastun, A.: Persistence in the cryptocurrency market. Research in International Business and Finance, vol. 46, pp. 141–148 (2018)
8. Constantin, M., Sarma, S. D.: Volatility, persistence, and survival in financial markets. Phys. Rev. E, vol. 72, no. 5, pp. 106–116 (2005)
9. Costa, N., Silva, C., Ferreira, P.: Long-range behaviour and correlation in dfa and dcca analysis of cryptocurrencies, vol. 7 (2019)
10. De Queiroz, S.: Roughness of time series in a critical interface model. Physical . Rev. E, vol. 72, no. 6, pp. 104–110 (2005)
11. Deniz, P., Stengos, T.: Cryptocurrency returns before and after the introduction of bitcoin futures. Journal of Risk and Financial Management, vol. 13, no. 6 (2020) doi: 10.3390/jrfm13060116
12. Fraser, J. G., Bouridane, A.: Have the security flaws surrounding bitcoin effected the currency's value? In: 2017 Seventh International Conference on Emerging Security Technologies (EST). pp. 50–55 (2017)
13. Ghimire, S., Selvaraj, H.: A survey on bitcoin cryptocurrency and its mining. In: 2018 26th International Conference on Systems Engineering (ICSEng). pp. 1–6 (2018)
14. Kantz, H., Schreiber, T.: Nonlinear Time Series Analysis. Cambridge University Press, 2 edn. (2003)
15. Mandelbrot, B.: The Fractal Geometry of Nature. W.H.Freeman & Co Ltd (1982)
16. Markets Insider: https://markets.businessinsider.com/currencies/btc-usd,
17. Meakin, P.: Fractals, Scaling and Growth Far from Equilibrium. Cambridge University Press (1998)
18. Munim, Z. H., Shakil, M. H., Alon, I.: Next-day bitcoin price forecast. Journal of Risk and Financial Management, vol. 12, no. 2 (2019)
19. Quintino, D., Campoli, J., Burnquist, H., Ferreira, P.: Efficiency of the brazilian bitcoin: A dfa approach. International Journal of Financial Studies, vol. 8, no. 2 (2020)
20. Rahouti, M., Xiong, K., Ghani, N.: Bitcoin concepts, threats, and machine-learning security solutions. IEEE Access, vol. 6, pp. 67189–67205 (2018)
21. Ramasco, J., López, J. M., Rodríguez, M. A.: Generic dynamic scaling in kinetic roughening. Phys. Rev. Lett., vol. 84, no. 10, pp. 2199–2202 (2000)
22. Schwartz, T.: Expert focus-estimating manpower requirements for expert system projects. IEEE Expert, vol. 3, no. 2, pp. 12–15 (1988)

ISSN 1870-4069

# A Statistical Prediction of COVID-19 Behavior in Mexico

Seyed Habib Hosseini Saravani, Sara Besharati,
Ponciano Jorge Escamilla-Ambrosio, Alexander Gelbukh

Instituto Politécnico Nacional,
Centro de investigación en Computación,
Mexico

{hosseinihaamed, besharatisara62}@gmail.com,
pescamilla@cic.ipn.mx, gelbukh@gelbukh.com

**Abstract.** This research uses a new method to predict COVID-19 pandemic behavior in Mexico. The advantage of this model over other statistical models is that it is less sensitive to the sudden changes in the behavior of the pandemic because this model uses the data of the countries in which the pandemic is in its advanced stages. This model predicts two scenarios for the future of the pandemic in Mexico: (a) the pandemic decreases steadily after its first peak until it ends, (b) the pandemic shows a second (or more) peak. In this research, based on the four most frequent patterns existing in the data of different countries, the data of daily infected cases in four countries – Italy, Spain, Iran, and France – is normalized with the data of Mexico to predict the pandemic behavior in Mexico. The new data gives us an estimate about the future of the pandemic in Mexico since we fit the characteristics of the data of Mexico to the data of the four countries chosen for this research and generate a new data for Mexico that follows the patterns occurred in those countries.

**Keywords:** Pandemic, COVID-19, prediction, statistical model.

## 1 Introduction

History has witnessed many fatal epidemics that ravaged peoples in different countries, wiping out many families and devastating many businesses. The infectious disease SARS-CoV-2 or COVID-19, starting in late 2019 from Wuhan, China, is another global disease outbreak infecting millions of people and causing thousands of deaths from different age groups. During COVID 19 crises, apart from professions in medical and paramedical areas, who are in the front line of battling the virus, researchers from other disciplines like computer science also tried to attribute to the improvement of the situation by providing statistical and probabilistic models and predictions about the behavior of the pandemic in the future days to come. If these models and predictions are accurate enough, they can save the lives of many people, prevent the resources of the countries from being wasted, and help other scientists and decision makers to take more effective decisions about the situation.

Statistical and probabilistic models are based on the data provided by the countries after the first number of infected cases and deaths were seen in those countries. Among all of the statistical methods, Gaussian based models, which describe a symmetric distribution where most of the observations cluster around the central peak and the probability for values further away from the mean scatter equally in both directions, are widely used to explain the behavior or statistical distribution of natural including pandemics. However, the evolution of pandemics like COVID-19 is not completely random and follows a life cycle pattern from the outbreak to the acceleration phase, inflection point, de-acceleration phase and eventual stop or ending [1]. Consequently, what has happened in the countries that experienced the situation first and are in a much more advanced situation can give us much more important information than the statistical predictions because it can help us understand the life cycle pattern of the pandemic. Therefore, we can use this vital information and fit the general patterns existing in it to the data of the countries that are in a much less advanced stages to predict the future of the pandemic in them.

In this research, we use the data of the daily number of cases and also the daily number of deaths in four countries that witnessed two different scenarios to predict each of these two scenarios for Mexico. The advantage of this model over the other statistical models is that it is less sensitive to the sudden changes in the behavior of the pandemic. However, it is worth mentioning that predicting the behavior of a virus in a society depends on several variables including the characteristics and the uniqueness of the virus in a certain society, the decisions that the authorities of that society have made to control the pandemic, the cultural characteristics of a society, and so on.

In the following, in Section 2, we will have a review on the previous related works; in Section 3, we explain the methodology; in Section 4, we compare the results obtained from the models used in this research; and in Section 5, we summarize our findings and talk about future work.

## 2 Related Work

After the outbreak of the coronavirus pandemic from Wuhan, China, many researchers tried to help to overcome this situation by predicting the pandemic behavior for those countries that have not experienced the peak yet. Thus, there have been many approaches toward the prediction of the behavior of the COVID-19 pandemic since its very beginning. Lou [2] made a research to predict the end of COVID-19 in the world. That research uses the context-specific and explainable SIR model in their predictions. To estimate the pandemic life cycle, they used a kind of daily updated COVID-19 data to regress the SIR models. Their predictions provide three alternative estimates of end dates in the order of conservativeness: (a) the date to reach the last expected case; (b) the date to reach 99% of the total expected cases; (c) the date to reach 97% of the total expected cases. For example, based on their prediction, Singapore was expected to bend the curve around May 5, and to through 97% of the cycle in the country around June 4; whereas, Italy and the United States were predicted to end 97% of their pandemics on May 7 and May 11 respectively. The theoretical ends for Singapore, United States and Italy all fall in August. According to that research, In Mexico, COVID-19 was expected to end 97% on 12-Jun-20, 99% on 25-June-2020, and 100% on 4-Sep-2020.

Barmparis et al. [3] estimated the infection horizon of COVID-19 in eight countries with a data-driven approach. Their quantitative approach is based on a Gaussian spreading hypothesis that is shown to arise as a result of imposed measures in a simple dynamical infection model. That research assumes the evolution will be similarly qualitatively Gaussian, although clearly with differences. They selected eight countries Greece, Netherlands, Spain, Italy, Spain, France, UK, and USA – and use data reported on April 4, 2020 for the task of prediction. Based on that research, COVID-19 pandemic horizon in Spain, Italy, Spain and USA was expected to be on 08-May-2020, 08-May-2020, 02-May-2020, and 10-May-2020 respectively.

Barbero [4] performed a Gaussian-based research to predict the end of the pandemic in Argentina and other Southern Hemisphere countries. However, they also reported their prediction about some other countries all over the world. Based on that research, the end of the pandemic in Italy and Spain was expected to be on 16-May-2020 and 11-May-2020 respectively. Also, that research predicted the end of the pandemic in Mexico to be on 25-May-2020.

As it can be seen, curve-fitting models are one of the most common methods used to predict COVID-19 pandemic behavior. Courtney [5] has made a research in the Technological University of Dublin, which presents a Simple Curve Approximation Tool (SCAT). This tool allows the user to approximate and draw the curve and allows testing of assumptions, trajectories and the wildly varying figures reported in the media. SCAT creates meaningful comparisons and understandable visualizations for COVID- 19 and other diseases.

Zhigljavsky et al. [6] showed in their research that standard SIR-type models are not accurate enough and are also not stochastic; therefore, they should be used with extreme caution. They provide a flexible modelling approach that increases the accuracy. They analyzed different scenarios relevant to the COVID-19 situation in the UK and present a stochastic model that captures the inherently probabilistic nature of contagion between population members. The computational nature of their model means that spatial constraints, like communities and regions, the susceptibility of different age groups and other factors such as medical pre-histories can be incorporated with ease. They analyze different possible scenarios of the COVID-19 situation in the UK.

Most of the predictions made by the researches above were either far from what happed later or not accurate enough. The main reason why it happens seems to be the nature of a pandemic's behavior, which cannot be known enough until it happens. Because of that, the method we used in this research is based on the behavior of the pandemic in the countries where the pandemic is in a much more advanced situation. In this way, we expect that the same kind of behavior will occur in of countries that were affected by the pandemic later than those countries.

## 3 Methodology

In this research, the European Centre for Disease Prevention and Control data [7] is used. This data, which gets updated every day, gives the information about the number of cases and also the number of deaths in all of the countries in the world that are affected by coronavirus.

The data we used in this research shows that some countries have passed their peak(s) and are now experiencing a negative slope (less daily cases) while, in some other countries, the number of new cases is still increasing. However, a few types of patterns are being repeated in most of the countries. Therefore, we chose four countries – Italy, Spain, Iran, and France – based on three criteria to predict the future of behavior in Mexico. Our criteria for choosing those countries where: (a) the number of affected cases in the countries must be close to that of Mexico; (b) the countries must pass their peaks; (c) two out of the four countries should experience a second (or more) peak. Thus, all of the countries we chose fulfilled criterion (a), which means that they can be put in the same group of countries based on the number of affected cases they have; considering criterion (b), we chose Italy and Spain because they were close to their ends; and based on criterion (c), we chose Iran – because they were experiencing quite a big second peak after their first peak –, and France – because they were facing small fluctuations (small peaks) close to the end of the pandemic.

Choosing these four different countries based on the patterns they experienced, we created our models to predict two major scenarios that might be witnessed in Mexico. Scenario one is what happened in Italy and Spain, where the number of cases decreased steadily after their first peak, and they did not experience a second peak, and the second scenario arises when the country faces a second (or more) peak after the first peak, so the patterns that occurred in Iran and France has the information that we need to predict the second scenario for Mexico.

The algorithm used in our models for all of the countries studied in this research is the same. First, we calculate the slope of the line with which the data of Mexico grew to the point it was by the time this research was being done. Then we obtain the values of the slopes of the lines with which the data of the other four countries grew to their first peaks, and also the slopes of the lines with which the data of these four countries decreased from their first peaks to the point where they were when the research was being done. Then we generated a new slope using the slopes we calculated. The equations for the calculation of the slopes and also generating the new slope are:

$$m = \frac{y_2 - y_1}{x_2 - x_1},$$

$$m' = \frac{|m_0|}{\frac{(|m_1| + |m_2|)}{2}},$$

where $m$ is the general equation for calculating the slopes, $m_0$ refers to the slope of the data of Mexico, and $m_1$, and $m_2$ are the positive and the negative slopes of the data of the country that is being normalized with the data of Mexico respectively. Also, $m'$ is the new slope or the ratio that is multiplied by each number of daily cases of the data of the four countries that are being normalized with the data of Mexico. The equation for the final normalization of the data of the four countries is:

$$c' = c_i \times m'.$$

In equation above, $c_i$ is the number of daily cases in the country which is being normalized with the data of Mexico, and $c'$ stands for the normalized number of daily cases which can predict the behavior of the pandemic in Mexico.

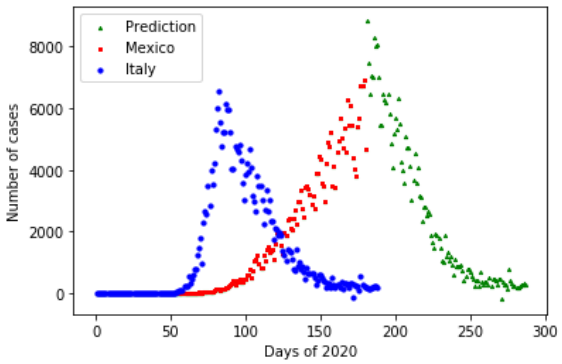



**Fig. 1.** Prediction of the pandemic in Mexico based on the data of Italy.

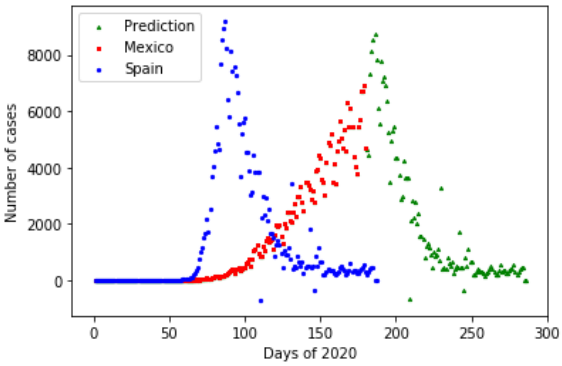



**Fig. 2.** Prediction of the pandemic in Mexico based on the data of Spain.

**Table 1.** Results obtained from the models based on the data of Italy and Spain.

| | **Model based on Italy** | **Model based on Spain** |
|---|---|---|
| Pandemic Peak | 200$^{th}$ day of 2020 | 190$^{th}$ day of 2020 |
| Pandemic End | 300$^{th}$ day of 2020 | 290$^{th}$ day of 2020 |

## 4 Experimental Results

As it is shown in Figures 1 and 2, using our model, we normalized the data of Italy and Spain with the data of Mexico to predict the peak and the end of the pandemic in Mexico.

As Table 1 shows, the data normalized with the data of Italy predicted that the peak in Mexico occurs in mid-July (200$^{th}$ day of 2020), when there will be about 8600 cases per day and the end of the pandemic in Mexico is almost in late October (the 300$^{th}$ day of 2020). Also, the model which was based on the data of Spain says that, at the peak of the pandemic, Mexico experiences almost 8400 cases in the second week of July (190$^{th}$ day of 2020), and the pandemic ends in Mexico in mid-October (290$^{th}$ day 2020).

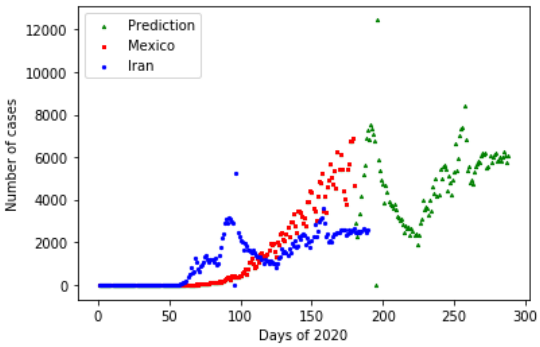



**Fig. 3.** Prediction of the pandemic in Mexico based on the data of Iran.

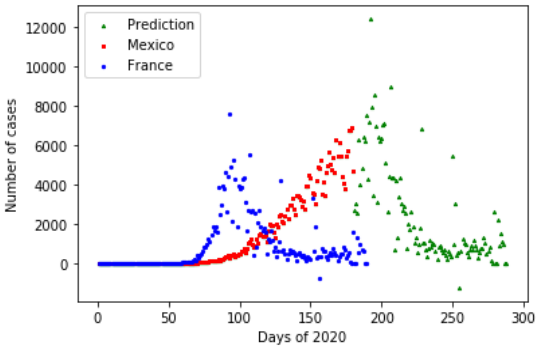



**Fig. 4.** Prediction of the pandemic in Mexico based on the data of France.

**Table 2.** Results obtained from the models based on the data of Iran and France.

| | **Model based on Iran** | **Model based on France** |
|---|---|---|
| First Peak | 200$^{th}$ day of 2020 | 200$^{th}$ day of 2020 |
| Second Peak | 225$^{th}$ day of 2020 | 280$^{th}$ day of 2020 |

About the second scenario, when the data faces a second (or more) peak, as Figure 3 shows, the pandemic behavior in Iran was in a way that they faced a second peak when the number of affected cases was decreasing gradually. In case of such a scenario, our model estimated that Mexico faces its first peak around in mid-July (200$^{th}$ day of 2020), and the number of cases decreases after that. However, after mid-August (225$^{th}$ day of 2020), again the number of cases increases until Mexico experiences its second peak around mid-September (260$^{th}$ day of 2020).

Moreover, the model based on the data of France predicted that the peak of the pandemic in Mexico is around mid-July (200$^{th}$ day of 2020) when Mexico experiences about 8500 affected cases per day, and the pandemic gets close to its end in late October (the 300$^{th}$ day of 2020); however, based on this model, we will see a second (or more) small peak between 250$^{th}$ and 300$^{th}$ days of the year.

## 5 Conclusions

The models we developed in this research predicted the peak and the end of the pandemic in Mexico based on the shape of the distribution of the data in Italy and Spain, where the pandemic is very close to its complete end. Moreover, using the data of two countries that experienced a second peak – Iran and France –, our model predicted two scenarios for facing a second (or more) peak in Mexico. The advantage of our model over other statistical models is that it is less sensitive to the sudden changes in the behavior of the pandemic because our predictions are based on the data of the countries where different possible changes have already occurred.

In spite of the fact that statistical models can give us a really good estimate about the future, it is naïve if we think that how COVID-19 behaves in a society like Italy will be repeated exactly in the same way in other societies. As mentioned before, there are always many variables that can change the behavior of a pandemic. For example, the climate, whether a society is in quarantine or not, the dates that the government announced the nationwide quarantine, if the people of a certain society will welcome the new situation and follow the quarantine rules, and also many other unknown variables can change the behavior of a pandemic.

For future work, we aim to use pattern recognition methods that can learn the patterns existing in the data of different countries to predict which patterns a country might follow in the future.

## References

1. Al-Nahhal, I., Dobre, O.A., Basar, E., Moloney, C., Ikki, S.: A fast, accurate, and separable method for fitting a gaussian function. IEEE Signal Processing Magazine, 6, pp. 157–163 (2019)
2. Luo, J.: When will COVID-19 end? Data-driven prediction. Singapore University of Technology and Design (2020)
3. Barmparis, G.D., Tsironis, G.P.: Estimating the infection horizon of COVID-19 in eight countries with a data-driven approach. Chaos, Solitons & Fractals, 135, pp. 109842 (2020)
4. Barbero, C.A: A statistical forecast of low mortality and morbidity due to COVID-19, in Argentina and other Southern Hemisphere countries. MedRxiv (2020)
5. Courtney, J.: COVID-19: Tracking the pandemic with a simple curve approximation tool (SCAT). MedRxiv (2020)
6. Zhigljavsky, A., Whitaker, R., Fesenko, I., Kremnizer, K., Noonan, J., Harper, P., Gillard, J., Woolley, T., Gartner, D., Grimsley, J., de Arruda, E.: Generic probabilistic modelling and non-homogeneity issues for the UK epidemic of COVID-19. MedRxiv (2020)
7. European Centre for Disease Prevention and Control: Download historical data (to 14 December 2020) on the daily number of new reported COVID-19 cases and deaths worldwide. IOP Publishing Physics Web (2020)

ISSN 1870-4069

# Data Augmentation vs Regularization for Time Series Forecasting

Juan J. Flores, Miguel A. Reynoso, Josué D. González, Felix Calderon

Universidad Michoacana de San Nicolás de Hidalgo,
Facultad de Ingeniería Eléctrica,
División de Estudios de Posgrado,
Mexico

{juan.flores, 0850750b, jdgonzalez, felix.calderon}@umich.mx

**Abstract.** This article presents a study on the introduction of noise to the training data of an Artificial Neural Network that models the forecasting process for time series. The introduction of noise can act as a form of regularization, improving the network capacity to generalize on test data. When using an Artificial Neural Network to perform time series forecasting, we need to find out the best way to introduce noise to the data, preserving its temporal relation, while being able to populate the training set with a large enough number of noisy data points as to achieve the desired regularization effect. First we convert the time series to a design matrix, by determining the reconstruction dimension of the underlying process that produced it ($m$), together with a subsampling period ($\tau$); sweeping the time series forms all possible delay vectors. The resulting design matrix maps the forecasting problem into a regression problem. Noise is introduced to the design matrix, populating it with noisy delay vectors; noise is generated following a normal distribution $N(0, \sigma)$. Neural Networks are designed to produce regressors that solve the forecasting problem for a set of time series. The forecasting accuracy of the noise-regularized models is compared against models regularized by Early Stopping, second order Bayesian Regularization and both. Noise regularization produced better results in the vast majority of the cases.

**Keywords:** Time series, forecasting, regularization, data augmentation, noise.

## 1 Introduction

A time series is a time-ordered sequence of observations values of a variable made at equally spaced time intervals, represented as a set of discrete values [1]. Time series are found in many fields, such as economics, sociology, meteorology, medicine, seismology, oceanography, geomorphology, astronomy, etc. [2]. Time series analysis helps to detect regularities in the observations of a variable, detect regularities in data, determine a suitable model, and/or exploit all information included in this variable to better predict future developments [3].

Time series forecasting is currently a very important research area, due to the importance of prediction in many fields. Applications range from natural (wind speed, ambient temperature, solar irradiance, etc.) to anthropic phenomena (stock price, electric energy consumption, etc.). Time series forecasting is an area in which past observations of the same variable are collected and analyzed to develop a model that describes the underlying relationship. The model is then used to extrapolate the time series into future scenarios [4].

Over the past several decades, many studies have been conducted to develop innovative forecasting approaches and improve their accuracy. In general, these models can be categorized into three types: statistical models, artificial intelligence models and hybrid models [5].

Statisticians and econometricians tend to rely on autoregressive integrated moving average (ARIMA) and derived or related models, while the artificial intelligence community mainly looks at neural networks, either using multilayer perceptrons or recurrent networks [6].

After fitting a time series model, one can evaluate it with forecast fit measures. The researcher may subtract the forecast value from the observed value of the data at that point in time and obtain a measure of error or bias. The statistics used to describe this error are similar to the univariate statistics just mentioned, except that the forecast is often substituted for the average value of the series. The chosen fitness function to evaluate the amount of this forecast error is the symmetric mean absolute percentage error (SMAPE), defined as shown by equation (1). This version of SMAPE ranges from 0 to 100%:

$$SMAPE = \frac{100}{n} \sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{y_t + \hat{y}_t}. \tag{1}$$

In equation (1), $n$ is the length of the time series, $y_t$ represents the real value, and $\hat{y}_t$ the forecast value.

We will address several research questions in this article: what is the relation between regularization and the introduction of noise in the training process of an ANN?, how to introduce noise to a time series, while preserving the relation of its measurements with time?, how much noise and how many noisy points must be introduced in the ANN training set?, and finally, is the introduction of noise to the training data any better or at least comparable to the process of Bayesian Regularization?

## 2 Forecasting as Regression

When using artificial neural networks (ANN) to solve the problem of time series forecasting there are several possibilities. One of them is to use Recurrent Neural Networks, which maintain a state of the sequence presented to the network and takes it into consideration for the following computation. GRU and LSTM are examples of such networks [7]. Another approach is the Dynamic Multi-layer Perceptrons [8], which use tapped delay lines to constantly provide part of the

time series history to an MLP. Another resource and the one we are using in this implementation is to convert the time series to a database or design matrix, that constitutes the examples to feed the ANN (training, validation, and test sets).

One problem present with any of these approaches is to determine how much of the history of the time series can be fed as input to the ANN. Let us say that we determined that a certain interval of the past is important to the ANN to produce an accurate forecast. The following problem to solve is if all of those measurements are important, or only a subset of them. Flores et al. [9] solved this problem using Evolutionary Computation.

Kantz and Screiber [10] solve the first part of this problem using a method called false nearest neighbors. This method allows determines the dimension, $m$, of the underlying dynamic system that produced the time series data. Another method based on mutual information determines a subsampling period, $\tau$, such that the amount of information provided to the ANN is maximized.

Let $\mathbb{S} = [s_1, s_2, \ldots, s_t, \ldots, s_N]$ be a time series, where $s_t$ is the value of variable $s$ at time $t$. It is desired to obtain the forecast of the point at $t + \Delta$ as a function of the observations available in $\mathbb{S}$.

By using a $\tau$ delay and an embedding dimension $m$, it is possible to build delay vectors of the form $[s_{t-(m-1)\tau}, s_{t-(m-2)\tau}, \ldots, s_{t-\tau}, s_t]$, which constitute a reconstruction of the phase portrait of the underlying dynamic system. We then append the value to be forecasted, $s_{t+\Delta}$ to each of those vectors. The resulting matrix is the design matrix, which represents the set of examples to be presented to the ANN for training, validating, and testing.

This transformation, from the time series to a design matrix, maps the forecasting problem to a regression problem. The first part of each example, $[s_{t-(m-1)\tau}, s_{t-(m-2)\tau}, \ldots, s_{t-\tau}, s_t]$, represents the independent variables and the second part, $s_{t+\Delta}$, the dependent variable in a regression problem. We already know how to solve this kind of problem using an MLP.

## 3 Neural Networks

An artificial neural networks is essentially a collection of nonlinear transfer functions that relate some output variable(s) of interest to some input variables, which may themselves be functions of even deeper explanatory variables [11].

The ANNs approach has been suggested as an alternative technique to time series forecasting [12, 13]; this approach has gained immense popularity in the last few years. ANNs try to recognize regularities and patterns in the input data, learn from experience, and then provide generalized results based on previous knowledge.

The main ANN used in forecasting problems is Multi-Layer Perceptrons (MLP). This model is characterized by a network of three types of layers (input, hidden, and output). These networks consist of neurons arranged in layers in which every neuron is connected to all neurons of the next layer by introducing the input in a feed-forward manner, which propagates through the hidden layer and the output layer [14].

The signals propagate from node to node and are modified by weights associated with each connection. The receiving node sums the weighted inputs from all of the nodes connected to it from the previous layer. The output of this node is then computed as the function of its input called the "activation function". The data moves forward from node to node with multiple weighted summations occurring before reaching the output layer [15].

Neural networks have been mathematically shown to be universal approximators of functions thus are inherently nonlinear and estimate well non-linear functions [16]. The universal approximator property suggests that neural networks can effectively model seasonality [17].

## 4 Generalization Problems with ANN

Overfitting is produced when a model fits the data too closely, so it is unlikely to generate good classification or prediction on previously unseen patterns. When that happens, the model's loss in training data may become very small, but the model loses its generalization capabilities. In the extreme, the loss tends to zero, in which case we can say the model memorizes the data. This is a problem found in most of the machine learning techniques [18].

Neural networks use a great number of parameters (weights and biases) to model phenomena, that is what makes neural networks so easy to overfit. An example of overfitting is shown in figure 1, the black curve follows perfectly all the data (red dots), but it is unlikely to have a good performance on new data. On the other hand, the blue curve is the best model to fit the data, since it is not too dependent on the data.

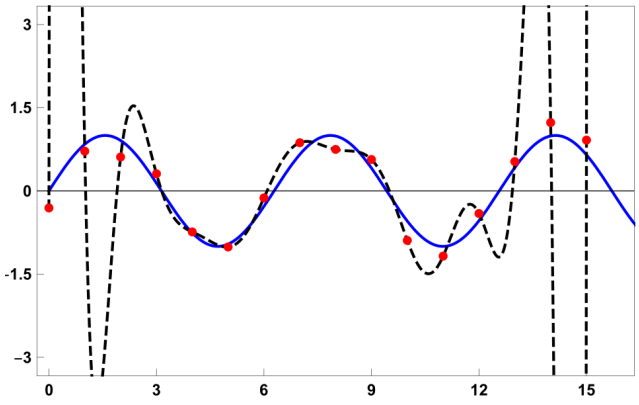



**Fig. 1.** Overfitting

Contrary to what happens with overfitting, underfitting is present when the model is incapable of capturing the underlying dynamics of the data.

Underfitting can happen for many reasons, one reason is when there is not enough data to build an accurate model, another reason is when models are not complex enough to reproduce the dynamics exhibited by the data.

Figure 2 shows the data to be fit by the model. The black curve is unable to follow the data, and, again, it is unlikely to have a good performance on new data.

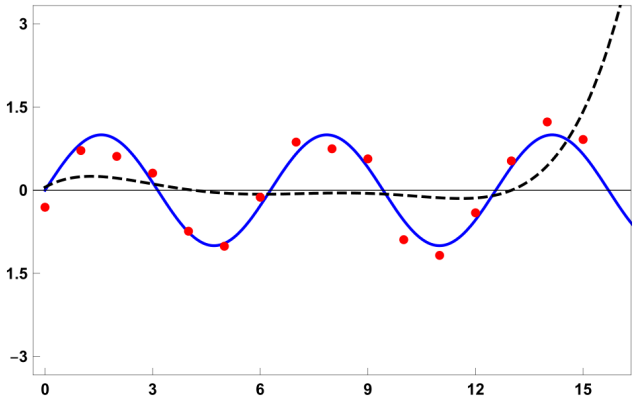



**Fig. 2.** Underfitting.

There are several strategies to avoid overfitting. One of them is to stop the learning process when the generalization starts to decrease, *early stopping*. Another strategy is to penalize complex model, *regularization*. One third strategy is to obtain more training data, *data augmentation (DA)*.

**Early Stopping** When the model performance, as measured by the loss function on the validation set, deteriorates in an arbitrary number of iterations, we can infer that the model's generalization capability is decreasing (Figure 3). That is, the model is being to overtrain; the training algorithm recognizes that condition and the training process stops. The number of iterations the network "waits" for the loss function to recover and continue improving is called **patience**.

**Regularization.** Since a larger number of parameters cause overfitting, a natural approach is to constrain the model to use fewer parameters, the fewer degrees of freedom it has, the harder it will be for it to overfit the data [19].

To create less complex models when the number of features in the data set is large, it is helpful to use some of the regularization techniques to avoid overfitting. One technique is called Ridge Regression; this technique regularizes the model by adding a *regularization term* equal to $\lambda \sum_{i=1}^{n} \beta_i^2$ to the cost function. This forces the learning algorithm to fit the data and keep the model's

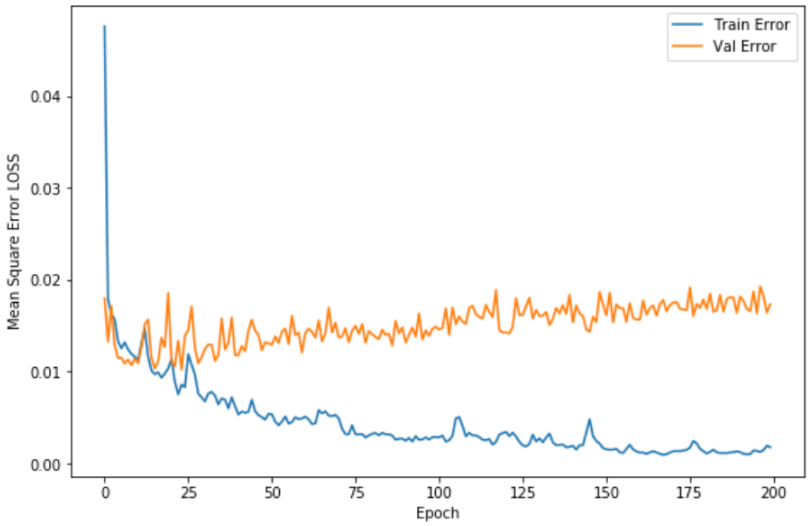



**Fig. 3.** Overfitting detection: while error in training set tends to decrease, it tends to increase in validation set.

weights as small as possible. The parameter $\lambda$ controls the entropy of the model distribution. As $\lambda \to \infty$, the model distribution tends a uniform distribution and all weights end up very close to zero and the prediction is a line going through the data's mean [20].

**Data augmentation.** is a set of techniques whose objective is to acquire synthetic samples from the training set [21]. DA is used for two reasons, first of all, in many cases the training set is limited, and it is difficult to obtain more samples; using DA we can produce additional synthetic samples. The second is to gain better generalization properties of the model. For example on computer vision the samples of the training set are images, we can obtain synthetic samples from the training set by applying transformations to the images, e.g., rotations, translations, etc. [3]. This technique provides a larger data set and makes the vision system more robust to changes in images.

There are other methods to provide additional samples to the training set. Examples of those techniques are adding noise to the samples, randomly erasing parts of a data sample, etc. Currently, generative adversarial networks (GANs) are attracting attention from research [21, 22].

In GANs, a generator captures the data probability distribution to randomly generate new samples with the same distribution; finally a discriminator tries to learn which samples are real and which are fake.

The DA techniques that give good results in computer vision, give poor results (or are not suitable) for time series analysis. Therefore, great care must

be taken when choosing the DA technique; it is important to have a large number of samples as the generate are of good forecasting accuracy [23].

The data augmentation technique used in this paper is noise injection. This technique consists on generating synthetic samples by taking each sample of the training set and adding Gaussian noise [24]. From each sample of the training set, $S$ synthetic samples will be created. To generate a synthetic sample, a certain amount of Gaussian noise is added to the original sample. Therefore, the training set has $N \cdot S$ synthetic samples

The noise to be injected into a sample (a delay vector) must be normally distributed. The amount of noise added when creating the synthetic samples is given by an SNR value, given in decibels [25]. A standard deviation, $\sigma$, must be calculated from this value [26]; this standard deviation is used to generate the Gaussian noise that is added to the original samples while creating the synthetic samples.

To determine the value of $\sigma$, for a given SNR, we first compute the power of the time series:

$$E_{ts} = \frac{1}{L} \sum_{i=0}^{L-1} |s_i|^2 , \tag{2}$$

where $L$ is the length of the time series. Expressing the decibel SNR to a linear scale:

$$SNR_{lin} = 10^{SNR_{dB}/10}, \tag{3}$$

allows us to finally determine $\sigma$:

$$\sigma = \sqrt{\frac{E_{ts}}{SNR_{lin}}}. \tag{4}$$

A sample delay vector is shown in Figure 4, plotted in red, as well as four synthetic samples in black. The synthetic samples were created from the original sample contained in the training set. We can observe that although all the samples are different, they maintain the same qualitative form.

## 5 Tests Description

Seven-time series were used in the experiments, each of them was used to train a neural network. For each time series, we created one multilayer perceptron as forecasting tool, figure 5 shows one example of MLP. We considered in their design their modeling capabilities; i.e., they do not underfit the data.

For each time series, four tests were performed using the original training set:

**Test 1** Train a neural network without regularization nor early-stopping (this test serves as a reference point for the following).
**Test 2** Train the neuronal networks using early-stopping.
**Test 3** Training the neuronal networks using regularization L2.

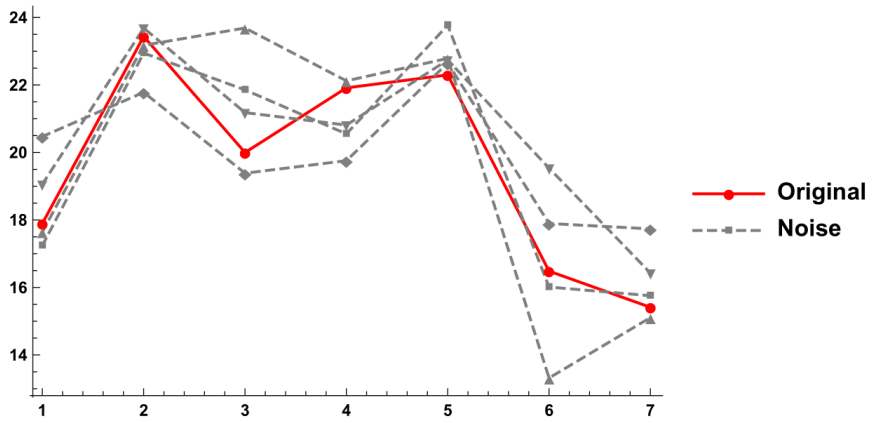



**Fig. 4.** Data augmentation.

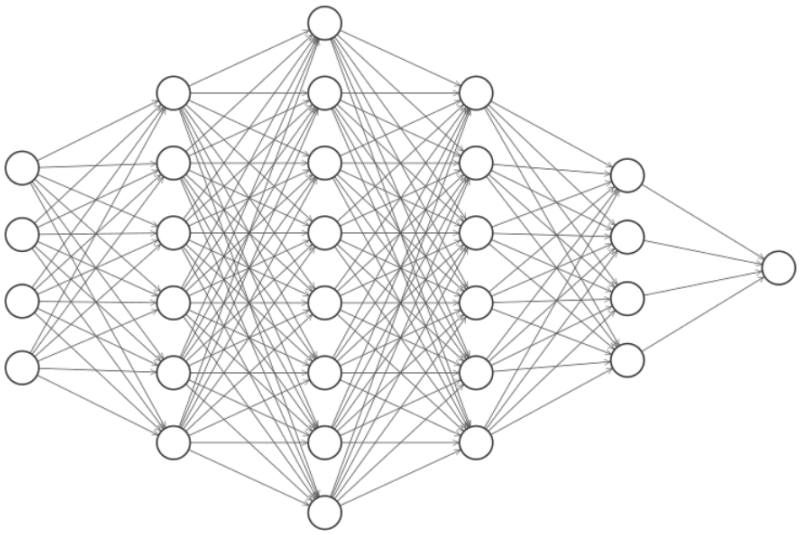


**Fig. 5.** Five hidden layers MLP.

**Test 4** Train the neuronal network using early-stopping and regularization L2.

We repeated those four tests augmenting the training set with synthetic samples.

In the comparison of the results, we used the SMAPE of the test set against the prediction of the neural network. This error function ranges from 0 to 100%, providing a way to compare results in a time series of different scales. The difference between the test lies in the way of training the neural networks:

Eight-time series were used in the experiments. Those time series are:

1. Ambient temperature: This dataset provides samples every 10 minutes from Morelia, Michoacán. This time series is stationary. The current day temperature is quite similar to the previous day, thus exhibiting daily seasonality.
2. Ambient temperature: subsampled hourly.
3. Air passengers: The air passengers dataset provides a monthly total of US airline passengers from 1949 to 1960.
4. Solar irradiance: Just like the temperature dataset, solar irradiance provides samples every 10 minutes and also shows daily seasonality and stationary.
5. Solar irradiance: subsampled hourly.
6. Sunspots: The sunspots dataset provides a monthly mean of the seen sunspots from 1749 to 2017.
7. Electric power distribution: Electric power distribution dataset provides samples every hour.

Figure 6 shows the time series used for the experiments.

## 6 Results

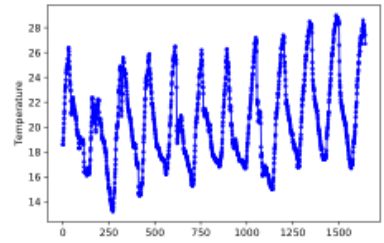
Table 1 shows the SMAPE values obtained from the tests. The first column shows the time series, the second column shows the results using the original data, and the third column shows the results using the original data plus the synthetic samples.

In addition, the second and third columns are subdivided into four columns: simple (results from the test without early-stopping nor regularization), early stopping (ES), regularization ($\lambda_2$), early stopping, and regularization (ES+$\lambda_2$). The results marked in red are the lowest error we find, and the green ones are the lowest error on the other category.

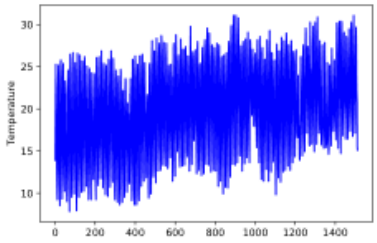
**Table 1.** Summary using SMAPE.

| Time Series | Original | | | | Data Augmentation | | | |
|---|---|---|---|---|---|---|---|---|
| | Simple | ES | $\lambda_2$ | ES+$\lambda_2$ | Simple | ES | $\lambda_2$ | ES+$\lambda_2$ |
| Temperature | 1.16 | 1.16 | 0.81 | 0.99 | 0.69 | 0.71 | 0.67 | 0.61 |
| Temperature (subsampled) | 2.21 | 2.46 | 2.22 | 2.23 | 2.20 | 2.24 | 2.30 | 2.21 |
| Airpassengers | 0.45 | 0.45 | 0.40 | 0.45 | 0.33 | 0.39 | 0.32 | 0.32 |
| Sunspots | 23.24 | 23.19 | 23.15 | 23.58 | 22.68 | 22.67 | 22.79 | 23.80 |
| Solar irradiance | 57.65 | 59.23 | 57.60 | 59.17 | 56.89 | 59.40 | 57.13 | 57.34 |
| Solar irradiance (subsampled) | 59.60 | 58.78 | 59.42 | 59.51 | 58.85 | 58.354 | 58.75 | 58.48 |
| Distribution data | 12.09 | 12.67 | 12.07 | 11.99 | 11.94 | 11.90 | 11.98 | 12.42 |

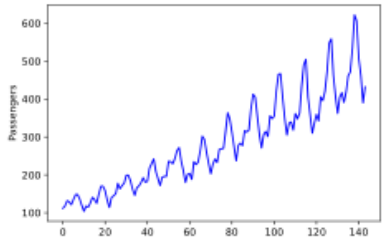
By observing Table 1, it can be noted that better results are obtained in most cases by adding synthetic samples to the training set. However, some considerations should be taken into account when using this Data Augmentation technique, such as the fact that two additional hyper-parameters are added to the



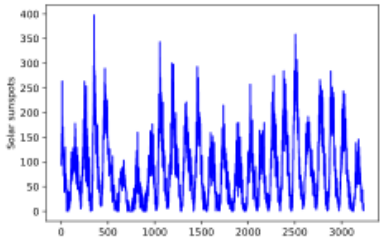

(a) Temperature.



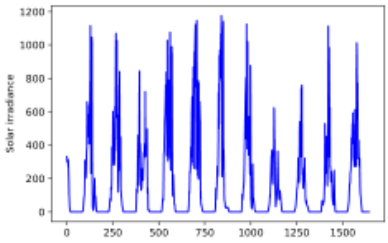

(b) Temperature subsampled.



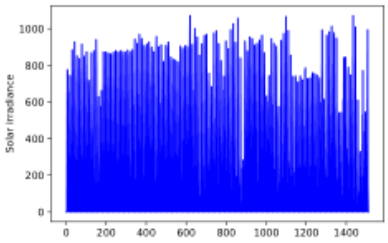

(c) Air passengers.



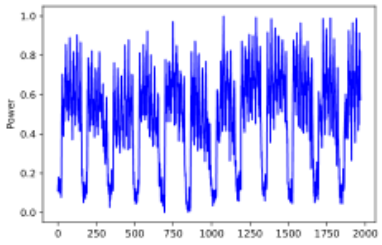

(d) Solar sunspots.




(e) Solar irradiance.

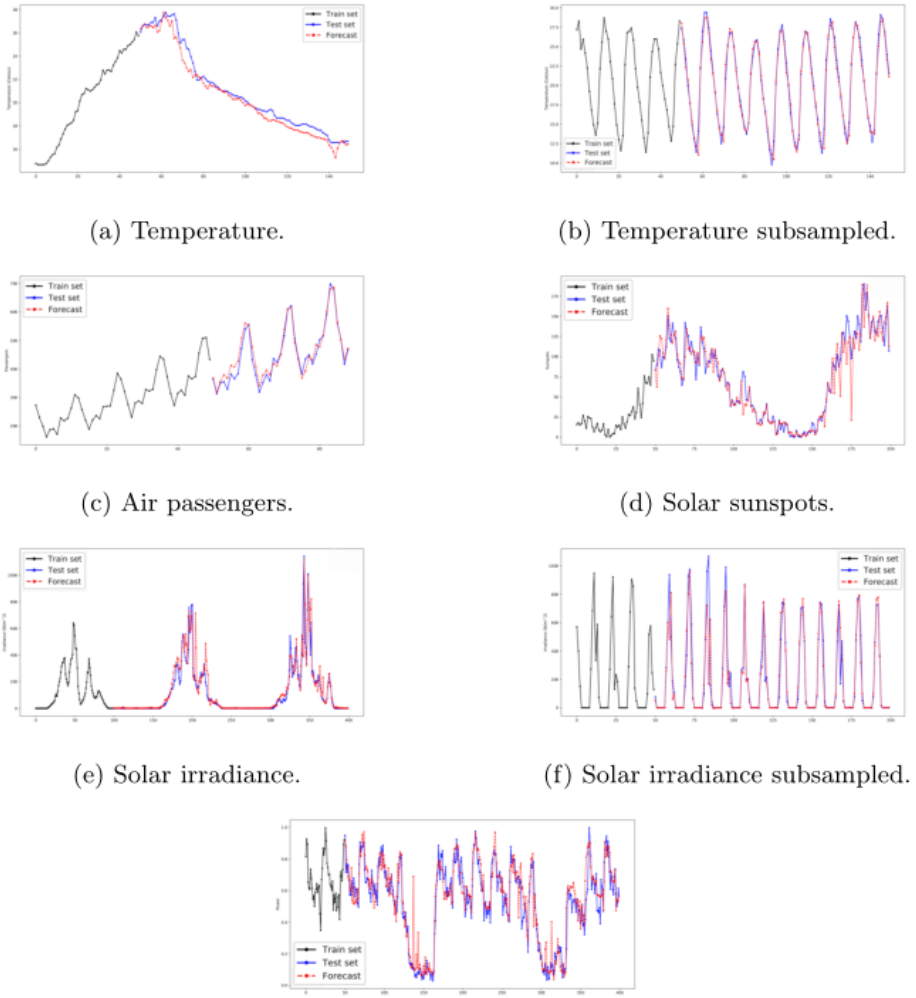



(f) Solar irradiance subsampled.




(g) Electric distribution.

**Fig. 6.** Plots of the time series used for testing.




(a) Temperature.

(b) Temperature subsampled.

(c) Air passengers.

(d) Solar sunspots.

(e) Solar irradiance.

(f) Solar irradiance subsampled.

(g) Electric distribution.

**Fig. 7.** Forecast plots. The black line is the train set, the blue curve is the test set and the red-dashed line is the forecast.

training process. These two hyper-parameters are the magnitude of the noise to be added and the number of noisy samples used in the creation of the augmented data sets.

The value of SNR varies depending on the time series. When considering the number of synthetic samples, something similar happens since each time series yields good results with a different number of samples. Besides, the extra

computational effort that is made by increasing the number of samples in the training set must be considered.

## 7 Conclusions

Regularization is a filter used to reduce signal noise; with the appropriate parameters, it is possible to have a noise-free signal with entropy reduction. In this case, a regularized model reduces the entropy of the model distribution, thus reducing overfitting.

We propose one method to introduce noise to the forecaster ANN training set. First, we convert the time series to a design matrix, by determining the reconstruction dimension of the underlying the process that produced it ($m$), together with a subsampling period ($\tau$); sweeping the time series forms all possible delay vectors. The resulting design matrix maps the forecasting problem into a regression problem. Noise is then introduced to the design matrix, populating it with noisy delay vectors; noise is generated following a normal distribution $N(0, \sigma)$.

How much noise and how many noisy points we need to introduce to the training set are questions that were solved empirically. On one extreme, too little a $\sigma$ introduces no noise; on the other end, too large a $\sigma$ destroys the information the training set conveys; we performed several experiments and determined that a good SNR for the introduced noise was 30 $dB$. From there, we determined the appropriate value for $\sigma$. About the second question, if we introduce 0 noisy points, the training set remains unchanged.

Adding more noisy points improves the network generalization capabilities up to a certain point, where it stabilizes asymptotically to an error limit. Increasing the size of the training set is not free, it consumes time and memory resources in the training process. Executing different experiments, we determined that adding 3 noisy points to the training set achieves a good equilibrium between generalization and training time.

Finally, Table 1 shows that the introduction of noise in the training set outperforms other regularization techniques. A set of experiments were designed to compare the performance of DA against models regularized by Early Stopping, second-order Bayesian Regularization, and both. Noise regularization produced better results in the vast majority of the cases.

## References

1. Palit, A., Popovic, D.: Computational intelligence in time series forecasting: Theory and engineering applications. Advances in Industrial Control, Springer London (2006)
2. Granger, C., Newbold, P., Shell, K.: Forecasting economic time series. Elsevier Science (2014)
3. Kirchgässner, G.,Wolters, J.: Introduction to modern time series analysis. Springer (2008)

4. Zhang, G.P.: Time series forecasting using a hybrid arima and neural network model. Neurocomputing, 50, pp. 159–175 (2003)
5. Xu Shuojiang, C.H.K., Tiantian, Z.: Forecasting the demand of the aviation industry using hybrid time series sarima-svr approach. Transportation Research Part E: Logistics and Transportation Review, 122(2) (2019)
6. Lemke, C., Gabrys, B.: Meta-learning for time series forecasting and forecast combination. Neurocomputing, 73(10-12), pp. 2006–2016 (2010)
7. Ayoubi, M.S.M., Sinsel, S.: Dynamic neural units for nonlinear dynamic systems identification. Lecture Notes in Computer Science, 930, pp. 1045–1051 (1995)
8. Brezak, D., Bacek, T., Majetic, D., Kazac, J., Novakovic, B.: A comparison of feed-forward and recurrent neural networks in time series forecasting. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–6 (2012)
9. Flores, J., Graff, M., Rodriguez, H.: Evolutive design of arma and ann models for time series forecasting. Renewable Energy, 44, pp. 225–230 (2012)
10. Kantz, H., Schreiber, T.: Nonlinear time series analysis. 7, Cambridge University Press (2004)
11. Kamstra, M., Donaldson, R.G.: Forecast combining with neural networks. Journal of Forecasting, 15 (1996)
12. Khashei, M., Reza-Hejazi,S., Bijari, M.: A new hybrid artificial neural networks and fuzzy regression model for time series forecasting. Fuzzy Sets and Systems, 159 (2008)
13. de Oliveira, K.A., Vannucci, A., da Silva, E.C.: Using artificial neural networks to forecast chaotic time series. Physica A: Statistical Mechanics and its Applications, 284 (2000)
14. Shirvany, Y., Hayati, M., Moradianc, R.: Multilayer perceptron neural networks with novel unsupervised training method for numerical solution of the partial differential equations. Applied Soft Computing, 9 (2009)
15. Pijanowski, B.C., Brown, D.G., Shellito, B.A., Manik, G.A.: Using neural networks and gis to forecast land use changes: A land transformation model. Computers, Environment and Urban Systems, 26 (2002)
16. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems (MCSS), 2 (1989)
17. Nelson, M., Hill, T., Remus, W., O'Connor, M.: Time series forecasting using neural networks: Should the data be deseasonalized first?. Journal of forecasting, 18 (1999)
18. Berzal, F.: Redes neuronales and deep learning. Fernando Berzal (2018)
19. Aggarwal, C.: Neural networks and deep learning: A textbook. Springer International Publishing (2018)
20. Géron, A.: Hands-on machine learning with scikit-learn and tensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media (2017)
21. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning (2017)
22. Zhang, X., Wang, Z., Liu, D., Ling, Q.: Dada: Deep adversarial data augmentation for extremely low data regime classification. In: ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2807–2811 (2019)
23. Hassan, I.F., Germain, F., Jonathan, W., Lhassane, I., Pierre-A., M.: Data augmentation using synthetic data for timeseries classification with deep residual networks (2018)

24. Brown, W.M., Gedeon, T.D., Groves, D.I.: Use of noise to augment training data: A neural network method of mineral–potential mapping in regions of limited known deposit examples. Natural Resources Research, 12(2), pp. 141–152 (2003)
25. Velázquez,V.M.T.: Modelado del error de predicción en series de tiempo basado en la calidad de sus datos. Master's thesis, Universidad Michoacana de San Nicolás de Hidalgo, 8 (2019)
26. Viswanathan, M.: How to generate awgn noise in matlab/octave (without using in-built awgn function) (2015)

ISSN 1870-4069

# A Comparative Study on the Effectiveness of Naïve Bayes Classifiers in Spam Filtering

Sara Besharati, Seyed Habib Hosseini Saravani, Alexander Gelbukh

Instituto Politécnico Nacional,
Centro de investigación en Computación,
Mexico

{besharatisara62, hosseinihaamed}@gmail.com,
gelbukh@gelbukh.com

**Abstract.** Naïve Bayes classifiers are among the very effective machine learning classifiers being used for the task of spam filtering. However, there are different kinds of Naïve Bayes classifiers based on their training algorithms and also their attitudes toward the data distribution and representation. In this research, conducting a comparative study on the effectiveness of multinomial, Bernoulli, and Gaussian Naïve Bayes classifiers in spam filtering, we show that multinomial and Bernoulli Naïve Bayes algorithms are more effective than Gaussian Naïve Bayes algorithm for the task of spam filtering. However, based on F1 Score, multinomial Naïve Bayes has the best performance among all the three models.

**Keywords:** Email, spam, Naïve bayes, multinomial, Bernoulli, gaussian.

## 1 Introduction

The tendency to use email technology as a kind of medium has increased to a large extent since internet became an inevitable aspect of everyday life of the majority of people in the world. Consequently, little by little, the ones who wanted to advertise their products and services, started making use of this phenomenon (email) to absorb their customers. Today, we receive many unwanted emails that occupy the space of our inboxes, and we have to spend some of our time removing them. Thus, the need for having a system that can filter the emails in order to detect the unwelcome mass emails being sent to the users (spam emails) has become more essential than before. Spam emails target their users mainly to advertise their contents; nevertheless, in some cases, they have destructive and even criminal intentions. Thus, the spam filtering systems should be strong and accurate enough to distinguish between spam emails and non-spam ones effectively.

Among all the machine learning method being used for the task of spam detection, Naïve Bayes classifiers are very famous. In this research, we compare the performance of three main Naïve Bayes algorithms – multinomial, Bernoulli, and Gaussian Naïve Bayes classifiers – that can be used for task of spam filtering. We compare the effectiveness of these classifiers in distinguishing spam emails from non-spam emails and introduce the most effective ones for this task.

In the following paragraphs, we review the related work in Section 2; we explain the algorithms of the three Naïve Bayes methods in Section 3; we compare the results obtained from the models developed in this research in Section 4; and we summarize our findings and talk about future work in Section 5.

## 2 Related Work

Schneider [1] applied multivariate Bernoulli Naïve Bayes and multinomial Naïve Bayes classifiers to the task of spam detection and reported that the multinomial models achieved higher accuracy than the Bernoulli model did. The multinomial Naïve Bayes and the Bernoulli model in that research were reported to have the accuracy 98.86% and 98.00% respectively, which is quite high.

McCue [2] compared the accuracy of Support Vector Machine (SVM) and Bayesian classifiers for the task of spam detection. The data used in [2] consisted of a matrix with 2000 email rows, each with 2000 feature columns, resulting in a $2000 \times 2001$ matrix. Thus, each email in this dataset has 2000 features, each of which is a binary of a word's existence within that email. The results of that paper showed that, despite the simplicity of Naïve Bayes algorithm, it can give a better prediction results on the testing set used in that research, coming in at a respectable 97.8% in comparison with the best accuracy obtained from SVM classifiers, which was 96.6%. Also, that paper reported that Gaussian Naïve Bayes algorithm cannot work with spam detection because of its 70% accuracy.

Méndez et al. [3] applied four different Naïve Bayes classifiers to the spam classification task. They presented a comparative study for the impact of five feature selection methods when using four variants of the original Naïve Bayes algorithm working as spam filter. The feature selection methods studied were Information Gain (IG), Odds ratio (OR), Document Frequency (DF) ($d$) $\chi 2$ statistic, and Mutual Information (MI). Moreover, we have analyzed the following Naïve Bayes alternatives: ($i$) Multivariate Bernoulli, ($ii$) Multinomial Naïve Bayes, ($iii$) Multivariate Gaussian, and ($iv$) Flexible Bayes. That research reported that, considering DF as the feature extraction, first Bernoulli and then multinomial Naïve Bayes models had the best performance, but the OR method presents a high performance level when it is used with Gaussian-based Naïve Bayes algorithms.

Almeida et al. [4] performed a comparison of performance achieved by four Naïve Bayes anti-spam filters – multinomial term frequency Naïve Bayes, multinomial Boolean Naïve Bayes, multivariate Bernoulli Naïve Bayes, Flexible Bayes – to classify messages as legitimate or spam. Among all the classifiers they used, multivariate Bernoulli Naïve Bayes achieved the best performance having the accuracy 98.90%. The next best accuracy, which was 97.47%, was obtained by the multinomial Boolean Naïve Bayes in that research.

In this research, doing a comparative research, we focus on the aspects which were not paid enough attention in the previous researches. We analyze the effectiveness of multinomial, Bernoulli, and Gaussian Naïve Bayes classifiers in filtering spam emails, and we show that each of these three classifiers can be effective for a specific task in spam filtering.

**Table 1.** Number of positive (spam) and negative (non-spam) emails in the dataset.

| | |
|---|---|
| **Positive** | 1368 |
| **Negative** | 4360 |

# 3 Methodology

We use a Kaggle dataset named 'Spam filter' [5], which consists of 5528 emails -- 4260 non-spam and 1368 spam emails (Table 1). We randomly chose 15% of the data and allocated it for the test dataset, and the rest of the data was used as the training dataset. Also, for the implementation of the codes and classification of the data, we use Scikit-learn [6], which is a free software machine learning library for the Python programming language.

## 3.1 Naïve Bayes

Naïve Bayes algorithm is a kind of a frequently used supervised learning method that examines all its training input and applies Bayes theorem with the "naïve" assumption of conditional independence between features given the value of the class variable [6]. Equation 1 below shows Bayes theorem, where $c$ stands for class variable and $x_1$ through $x_n$ are dependent feature vectors:

$$P(C \mid x_1, \cdots x_n) = \frac{P(C)P(x_1, \ldots, x_n|C)}{p(x_1, \ldots, x_n)}. \quad (1)$$

There are different kinds of Naïve Bayes classifiers based on their training and classification algorithms and their attitude toward the data distribution. In the following of this section, we will review the training algorithm of multinomial, Bernoulli, and Gaussian Naïve Bayes classifiers.

## 3.2 Multinomial Naïve Bayes

This algorithm is implemented to the data that is multinomially distributed and is one of the Bayes variants that is usually used in text classification [6]. The distribution is parametrized by vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each class y, where $n$ is the number of features – the size of the vocabulary – and $\theta_{yi}$ is the probability $P(x_i|y)$ of feature $i$ appearing in a sample belonging to class $y$. The parameters $\theta_y$ is estimated by a smoothed version of maximum likelihood, i.e. relative frequency counting:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} . \quad (2)$$

In the equation above, $N_{yi} = \sum_{x \in T} x_i$ is the number of times feature $i$ appears in a sample of class y in the training set T, and $N_y = \sum_{i=1}^{n} N_{yi}$ is the total count of all features for class y. The smoothing priors $\alpha \geq 0$ accounts for features not present in the learning

samples and prevents zero probabilities in further computations. Setting $\alpha = 1$ is called Laplace smoothing, while $\alpha < 1$ is called Lidstone smoothing [6].

### 3.3 Bernoulli Naïve Bayes

When the data is distributed according to multivariate Bernoulli distributions, where there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, Boolean) variable, Bernoulli Naïve Bayes can be used for the classification of the data [6]. In this kind of classification, this class requires samples to be represented as binary-valued feature vectors. Bernoulli Naïve Bayes makes the decision based on Equation 3, where P(I | y) refers to the probability of finding a term *i* in a given message belonging to class *y*:

$$P(x_i / y) = P(i / y)\, x_i + (1 - P(i / y))\, (1 - x_i). \quad (3)$$

This learning differs from multinomial Naïve Bayes rule in that, unlike the multinomial variant, it does not simply ignore feature *i* if it does not occur in class *y*. In the case of text classification, word occurrence vectors may be used to train the model. Bernoulli might perform better on some datasets, especially those with shorter documents [6].

### 3.4 Gaussian Naïve Bayes

In comparison with multinomial Naïve Bayes, Gaussian Naïve Bayes classifier assumes that the distribution associated to each term is a Gaussian distribution for each class *y*, and considers that the values of the attributes are independent in each class. Gaussian Naïve Bayes classifier uses continuous features by representing the frequency of the terms in an input [6, 2]. The likelihood of the features in this kind of classification is assumed to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right). \quad (4)$$

In Equation 4, $\mu_y$ and $\sigma_y$ represent the mean and the standard deviation of the appearance frequency of the terms in the inputs belonging to class *y* [6, 2].

### 3.5 Evaluation

Since accuracy is the most intuitive metric that can simply give us a ratio of correctly predicted observation to the total observations, the first evaluation metric we used was accuracy. In addition, in order to have the ratio of correctly predicted positive observations to the total predicted positive observations and also the ratio of correctly predicted positive observations to the all observations in actual class, we used the metrics precision and recall respectively. Finally, to take both false positives and false negatives into account, we used F1 Score, which is the weighted average of Precision and Recall:

**Table 2.** Comparison on different Naïve Bayes models used in this research.

| Models | Precision | Recall | F1 Score |
|---|---|---|---|
| Multinomial Naïve Bayes | 99.51 | 96.72 | 98.10 |
| Bernoulli Naïve Bayes | 95.67 | 98.02 | 96.83 |
| Gaussian Naïve Bayes | 76.92 | 85.10 | 80.80 |

**Table 3.** Comparison of our models – Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), Gaussian Naïve Bayes (GNB) – with the models in previous works.

| Models | Accuracy (%) |
|---|---|
| MNB | 99.06 |
| BNB | 98.48 |
| GNB | 91.16 |
| MNB - Schneider | 98.86 |
| BNB - Schneider | 98.00 |
| MNB - Almeida | 97.47 |
| BNB - Almeida | 98.90 |
| GNB - McCue | 70.00 |

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}}, \tag{5}$$

$$\text{Precision} = \frac{\text{TP}}{TP+FP}, \tag{6}$$

$$\text{Recall} = \frac{TP}{TP+FN}, \tag{7}$$

$$\text{F1 Score} = \frac{2\times(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}. \tag{8}$$

Equations 5, 6, 7, and 8 show the formulae for the calculation of accuracy, precision, recall, and F1 Score respectively, where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) refer to the correctly predicted spam emails, correctly predicted non-spam emails, incorrectly predicted spam emails, and incorrectly predicted non-spam emails.

## 4 Experimental Results

As Tables 2 and 3 shows, the multinomial and the Bernoulli Naïve Bayes have had a very good performance in classification of the emails. In this research, the multinomial models had the best performance reaching the accuracy 99.06%, and after that the

**Table 4.** Confusion Matrix of multinomial NB model.

| | **Positive** | **Negative** |
|---|---|---|
| Positive | 207 | 7 |
| Negative | 1 | 645 |

**Table 5.** Confusion Matrix of Bernoulli NB model.

| | **Positive** | **Negative** |
|---|---|---|
| Positive | 199 | 4 |
| Negative | 9 | 648 |

**Table 6.** Confusion Matrix of Gaussian NB model.

| | **Positive** | **Negative** |
|---|---|---|
| Positive | 160 | 28 |
| Negative | 48 | 624 |

Bernoulli model had the accuracy 98.48%. Also, it can be seen in Tables 2 and 3 that Gaussian Naïve Bayes model cannot compete with the multinomial and the Bernoulli Naïve Bayes models in spam filtering having the accuracy 91.16%.

As the confusion matrices of the three models (Tables 4, 5 and 6) show, the multinomial Naïve Bayes model has the best precision among all the three models, detecting the most number of spam emails correctly. However, when it comes to recall, the Bernoulli Naïve Bayes has the best performance, which means that this model sends the least number of spams to your inbox, but, comparing with multinomial model, it has more mistakes in spamming your non-spam emails. Figure 1 below shows the differences between the models considering different metrics for the evaluation of the models.

## 5 Conclusions and Future Work

The results obtained from the models developed in this research showed that, among the three well-known types of Naïve Bayes classifiers, the multinomial and the Bernoulli classifiers are more appropriate for the task of spam filtering. Both multinomial and Bernoulli algorithms had a good performance in detecting the spam emails, and their results were very close to each other; however, the multinomial model was stronger than the Bernoulli model considering precision as the evaluation metric, while the Bernoulli model had a better recall than the multinomial Naïve Bayes model. Nevertheless, it must be borne in mind that non-spam emails are more important than spam emails for the users, and a spam filtering system should make the less possible mistakes in classifying non-spam emails. As a result, if we consider precision as the evaluation metric, our results show that the multinomial Naïve Bayes model has the best performance among all the models we developed in this research.

For the future work, working on a selective Bayes classifier that can do an effective feature extraction for the task of spam filtering is aimed.

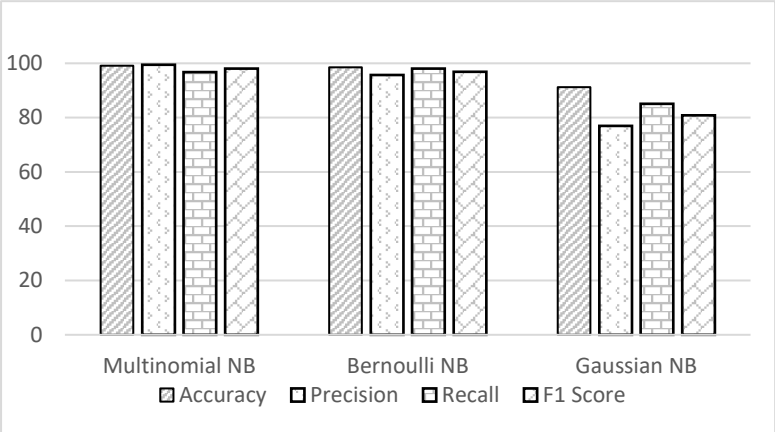



**Fig. 1.** Comparison of the results of different Naïve Bayes models used in this research.

## References

1. Schneider, K.M.: A comparison of event models for Naïve Bayes anti-spam e-mail filtering. In: Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 1, pp. 307–314 (2003)
2. McCue, R.: A comparison of the accuracy of support vector machine and Naïve Bayes algorithms in spam classification. University of California at Santa Cruz (2009)
3. Méndez, J.R., Cid, I., González-Peña, D., Rocha, M., Fernandez-Riverola, F.: A comparative impact study of attribute selection techniques on Naïve Bayes spam filters. In: ICDM´08 Industrial Conference on Data Mining, pp. 213–227 (2008)
4. Almeida, T.A., Yamakami, A., Almeida, J.: Probabilistic anti-spam filtering with dimensionality reduction. In: SAC´10 Proceedings of the ACM Symposium on Applied Computing, pp. 1802–1806 (2010)
5. Karthickveerakumar: Spam filter. 1 (2017)
6. Scikit-learn: 1.9. Naïve Bayes (2020)

ISSN 1870-4069

# Semantic Improvements to MPEG-7 Descriptors for Content-Based Image Retrieval

K. Salvador Aguilar Domínuez

Tecnológico Nacional de México / Centro Nacional de Investigación y Desarrollo Tecnológico,
Departamento de Ciencias de la Computación,
Mexico

kevin.aguilar17ca@cenidet.edu.mx

**Abstract.** This abstract shows a brief description of the principal parts of research project called: "Improvements to Recommended Descriptors by the MPEG-7 standard for content-based image retrieval", who is carried out for obtain the degree of doctor, in the department of computational sciences in CENIDET.

**Keywords:** MPEG-7, CBIR, semantic, descriptors.

## 1 Introduction

The file multimedia storage is increasing; therefore, the need arises to efficient filter, search and identify similar visual information [2]. The Moving Picture Experts Group (MPEG) developed the MPEG-7 standard (Multimedia content description interface [2]) to solve this problem and its applications are varied, like [3, 8]. One of the applications that are being given to the visual descriptors of this standard is in Content Based Image Retrieval (CBIR) systems [4, 6], but the descriptors present some problems in image recovery among millions of pictures [5, 9, 10].

Attempts have been made to propose modifications to these descriptors [10, 11], however, the methods are very mathematical and still present problems in recovery: the results only are similar in one of their characteristics such as color or shape [5]. These descriptors could be improved by introducing semantics to avoid those errors that are only similar in some characteristics and do not consider the semantic content of the image.

## 2 Previous Works in the Area

Currently, several MPEG-7 standard descriptors are still being implemented for images [12], as well as for video [13] and audio [14]. Similarly, the CBIR systems are still applied to solve many problems and work has been done to improve them as shown in [15], where is proposed a novel CBIR technique based on the visual words fusion of Speeded-Up Robust Features (SURF) and Fast REtina Keypoint (FREAK) feature descriptors. Other example is [16], this work focuses on a uniform partitioning scheme

which is applied in the Hue, Saturation and Value (HSV) color space to ex-tract Dominant Color Descriptor (DCD) features; in the proposed CBIR scheme, the DCD features were initially extracted as the color features, and then an appropriate similarity measure was applied. Among other works that solve other problems [17–21]. Modifications and proposals of different descriptors have been made to solve some problems as mentioned in [22], which proposed a novel feature descriptor named ‘‘Correlated MicroStructure Descriptor (CMSD)’’ for image retrieval; this descriptor represents high level semantics by identifying microstructures via establishing correlations between texture orientation, color, and intensity features. Other work is [23], where propose an accurate edge detector using Richer Convolutional Features (RCF), that encapsulates all convolutional features into more discriminative representation, which makes good usage of rich feature hierarchies, and is tractable to training via backpropagation. Among others works [24–26]. One example of semantic used is the semantic Web [27].

## 3 Research Objectives

General objective: Investigate, propose and evaluate an improved CBIR descriptor that overcomes one deficient or some deficiencies found in current ones, by means of a semantic descriptor.

Specific objectives: Study the current descriptors; Detect deficiencies; Propose improvements (semantic descriptor); and Apply it to a domain accepted by the international community as a test platform.

## 4 Methodology

We are considering the methodology in two parts: a) the collection of data and b) the process.

**a) Data collection**. It will seek to obtain collections of problematic digital images, that is, those where other descriptors have problems and fail; for example, extreme cases where, if semantics are not used, the descriptor would fail, such as the retrieval of images of specialized materials or dangerous activities, in these the semantic description is required not only form, texture and color.

In this sense we plan to have collections created by ourselves, as well as those commonly used in specialized literature; This may mean having synthetic digital image collections (artificially created to evaluate a certain aspect), as well as real-world digital images. With this, we hope to cover the entire possible spectrum of practical interest.

**b) Process**. We assume two basic stages: i) learning, where the proposed algorithm-descriptor is trained and its parameters are automatically tuned for a set of particular images; At the same time, an expert human has the possibility of making an additional subsequent adjustment, to include the human expert, and ii) recognition, where new images are fed to observe the behavior of the previously created model. At the heart of the model, there will be not only color or shape descriptors, but semantic descriptors,

similar to how Frames [28] or other forms of representation of semantic knowledge [29, 30].

## 5 Preliminary Results

It has been done the experimentation with the descriptors using the datasets in the literature whit the evaluation metric ANMRR (Averaged Normalized Modified Retrieval Rate), which is the recovery rate proposed by MPEG-7 [31]. The characterization time of the entire dataset is also measured in the case of Edge Histogram Descriptor (EHD) and the similarity calculation time in Color Layout Descriptor (CLD). similarly, we have presented proposals for the calculation of similarity using the CLD descriptor. This preliminary results were supervised by M. Lux, professor at Klagenfurt University, Austria [32] and Dr M. Mejía-Lavalle, researcher at TecNM/ CENIDET, Mexico.

## Referencias

1. Baresi, L., Colazzo, S., Mainetti, L., Morasca, S.: W2000: A modelling notation for complex web applications. In: Mendes, E., Mosley, N. (Eds): Web Engineering, Springer, pp. 335–364 (2006)
2. Koch, N., Knapp, A., Zhang, G., Baumeister, H.: UML-based web engineering: An approach based on Standards. In: Rossi, G., Pastor, O., Schwabe, D., Olsina, L. (Eds.): Web engineering: Modelling and implementing web applications, Springer, pp. 157–191 (2008)
3. Sikora, T.: The MPEG-7 visual standard for content description - An overview. In: IEEE Transactions on Circuits and Systems for Video Technology, 11(6), pp. 696–702 (2001)
4. Pattanaik, S., Bhalke, D.G.: Efficient content based image retrieval system using Mpeg-7 Features. In: International Journal of Computer Applications, 53(5), pp. 19– 24 (2012)
5. Mejía-Lavalle, M., Pérez-Lara, C., Ruíz-Ascencio, J.: The MPEG-7 visual descriptors: A basic survey. In: International Conference on Mechatronics, Electronics and Automotive Engineering, pp. 115–120 (2013)
6. Hyun, J., Kim, H.K., Oh, W.G.: Study on performance of MPEG-7 visual descriptors for deformable object retrieval. In: 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV) (2015)
7. Tyagi, V.: MPEG-7: Multimedia content description standard. Content-Based Image Retrieval, pp. 85–100 (2017)
8. Vertan, C., Badea, M.S., Florea, C., Florea, L., Bădoiu, S.: MPEG-7 visual descriptors se-lection for burn characterization by multidimensional scaling match. E-Health Bioeng, In: E-Health and Bioengineering Conference (EHB), pp. 253–256 (2017)
9. Zinzuvadia, K.M., Tanawala, P.B.A., Brahmbhatt, P.K.N.: Feature based image retrieval using classification and relevance feedback technique. In: International

Journal Advance Engineering and Research Development, 2(5), pp. 508–513 (2015)
10. Reta, C., Solis-Moreno, I., Cantoral-Ceballos, J.A., Alvarez-Vargas, R., Townend, P.: Improving content-based image retrieval for heterogeneous datasets using histogram-based descriptors. Multimedia Tools and Applications, 77(7), pp. 8163–8193 (2018)
11. Nazir, A., Ashraf, R., Hamdani, T., Ali, N.: Content based image retrieval system by using HSV color histogram, discrete wavelet transform and edge histogram descriptor. In: iCoMET International Conference on Computing, Mathematics and Engineering Technologies: Invent, Innovate and Integrate for Socioeconomic Development, pp. 1–6 (2018)
12. Georgescu, F.A., Răducanu, D., Datcu, M.: New MPEG-7 scalable color descriptor based on polar coordinates for multispectral earth observation image analysis. In: IEEE Geoscience Remote Sensing Letters, 14(7), pp. 987–991 (2017)
13. Ling-Yu Duan, Vijay Chandrasekhar, Shiqi Wang, Yihang Lou, Jie Lin, Yan Bai, Tiejun Huang, Alex Chichung Kot, Wen Gao: Compact descriptors for video analysis: The emerging MPEG standard. IEEE Computer Vision and Pattern Recognition (2017)
14. Lee, J., Lee, J.S.: Music popularity: Metrics, characteristics, and audio-based prediction. In: IEEE Transactions on Multimedia, 20(11), pp. 3173–3182 (2018)
15. Jabeen, S., Mehmood, Z., Mahmood, T., Saba, T., Rehman, A., Mahmood, M.T.: An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model. PLoS One, 13(4), pp. 1–24 (2018)
16. Fadaei, S., Amirfattahi, R., Ahmadzadeh, M.R.: A new content-based image retrieval system based on optimised integration of DCD, wavelet and curvelet features. IET Image Process, 11(2), pp. 89–98 (2017)
17. Xia, Z., Zhu, Y., Sun, X., Qin, Z., Ren, K.: Towards privacy-preserving content-based image retrieval in cloud computing. In: IEEE Transactions on Cloud Computing, 6(1), pp. 276–286 (2018)
18. Huang, F., Jin, C., Zhang, Y., Weng, K., Zhang, T., Fan, W.: Sketch-based image retrieval with deep visual semantic descriptor. Pattern Recognition, 76, pp. 537–548 (2018)
19. Wu, M., Xiao, W., Hong, Z.: Similar image retrieval in large-scale trademark databases based on regional and boundary fusion feature. PLoS One, 13(11), pp. 1–25 (2018)
20. Dash, J.K.: Content-based image retrieval using fuzzy class membership and rules based on classifier confidence. IET Image Process, 9(9), pp. 836–848 (2015)
21. Mohamadzadeh, S., Farsi, H.: Content-based image retrieval system via sparse representation. IET Computer Vision, 10(1), pp. 95–102 (2016)
22. Dawood, H., Alkinani, M.H., Raza, A., Dawood, H., Mehboob, R., Shabbir, S.: Correlated microstructure descriptor for image retrieval. In: IEEE Access, 7, pp. 55206–55228 (2019)
23. Liu, Y., Cheng, M.M, Hu, X., Wang, K., Bai, X.: Richer convolutional features for edge detection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8), pp. 1939–1946 (2019)

24. Zhang, J., Xia, Y., Xie, Y., Fulham, M., Feng, D.D.: Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features. In: IEEE Journal of Biomedical and Health Informatics, 22(5), pp. 1521–1530 (2018)
25. Napoletano, P.: Visual descriptors for content-based retrieval of remote-sensing images. In: International Journal of Remote Sensing, 39(5), pp. 1343–1376 (2018)
26. Fernández, L., Payá, L., Reinoso, O., Jiménez, L.M., Ballesta, M.: A study of visual descriptors for outdoor navigation using Google street view images. Journal of Sensors (2016)
27. Villagrán, A., Hernandez-Gress, N., Gonzalez-Mendoza, M.: Introducción a las tecnologías de la web semántica. Research in Computing Science (2012)
28. Minsky, M.: A framework for representing knowledge (1974)
29. Douglas, B.L.: CYC: A large-scale investment in knowledge infrastructure. In: Communications of ACM, 38(11), pp. 33–38 (1995)
30. Chen, L., Lambon-Ralph, M.A., Rogers, T.T.: A unified model of human semantic knowledge and its disorders. Nature Human Behaviour, 1(3), pp. 39 (2017)
31. Manjunath, B.S., Sikora, T., Salembier, P.: Introduction to MPEG-7 : Multimedia content description interface. John Wiley & Sons (2002)
32. Lux, M., Marques, O.: Visual information retrieval using java and lire. Morgan & Claypool Publishers (2013)

ISSN 1870-4069

# Detecting Mental Disorders in Social Media Using a Multichannel Representation

Mario Ezra Aragón Saenzpardo[1],
Adrián Pastor López Monroy[2], Manuel Montes y Gómez[1]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica, Mexico

[2] Centro de Investigación en Matemáticas, Mexico

pastor.lopez@cimat.mx, {mearagon, mmontesg}@inaoep.mx

**Abstract.** Common mental disorders such as depression, anorexia, dementia, post-traumatic stress disorder (PTSD) and schizophrenia affect millions of people around the world. In this work, to detect mental disorders in social media, we propose: 1) different representations from the information shared by the users. For example, topic information, phonetic or writing style, and emotion information. 2) A model that automatically creates a representation combining the previous representations. With these, the model can learn to represent social media documents (a.k.a. posts) by using the combination of these different types of information. The generated representations (individual and combined) will be evaluated in different tasks related to mental disorders, for example, depression detection, anorexia detection and post-traumatic stress disorder (PTSD). As preliminary results; we design a new representation considering emotions as information called Bag of Sub-Emotion(BoSE), which represents social media documents by a set of fine-grained emotions automatically generated using a lexical resource of emotions and sub-word embeddings. We evaluated this first representation in depression and anorexia detection. The results are encouraging; the usage of fine-grained emotions improved the results from traditional representations and a representation based on the core emotions and obtained competitive results in comparison to state of the art approaches. We also present results from a representation inspired by the emotional changes of a user, this representation combined with BoSE obtain better results than using them separately.

**Keywords:** Mental disorders, natural language processing, machine learning, deep learning.

## 1 Introduccion

**Motivation:** Currently, millions of people around the world are affected by different mental disorders that interfere in their thinking and behavior, damaging their daily life [1, 2]. Timely detection of mental disorders is important to help people before the illness gets worse, minimizing disabilities and returning them to their normal life. The stigma related to mental disorders creates barriers to improve the resources that help the detection of these problems. The most popular way for people to share information is using social media platforms, and people tend to share topics related to work issues and personal matters. People with mental disorders tend to share more about their concerns looking for some advice, support or just because they want to relieve suffering. This creates an excellent opportunity to automatically detect users that have a mental disorder and refer them as soon as possible to seek professional help.

**Previous Works in the area:** To accomplish the detection of mental disorders, automated analysis of social media is made using predictive models that uses features or variables that are extracted from the data posted by the users in their social media accounts. For example, one of the most commonly used features are word frequencies [6]. Other works used the LIWC dictionary [5], to represent the posts by means of a histogram of psycho-linguistic categories [7], or consider a representation based on the polarity information of the posts [8].

**Hypothesis:** People that present some mental disorder tend to express differently than healthy people; their topics of interests, writing style, relation with others and even their activity hours have different behavior. The hypothesis is that learning to combine different channels of information, could give a broader view that helps to detect signs of mental disorders and obtain better classification results that using single information.

**Objective:** Design a method applying traditional NLP techniques combined with deep learning techniques to automatically learn a multichannel representation using the information generated by the users in social media platforms. Then use this representation for the detection of mental disorders and improve the results obtained by traditional and state of the art approaches.

**Specific Objectives:**

1. Design methods that learn new representations of the different channels in the post history of the users: the context, the style of the author, the emotions used and phonetic information.
2. Design a model that automatically combines the different information channels and focuses on the critical parts of the data.
3. Develop a method to incorporate the importance of temporal information presented in the sequences of the posts.
4. Evaluate the utility of our proposed method in different tasks related to mental disorders.

## 2 Methodology

This section briefly describes the methodology to reach the proposed objectives.

1. **Identify and Obtaining datasets related to mental disorders.** We plan to obtain datasets like depression detection, anorexia detection, PTSD detection. For example, datasets from **CLEF eRisk** or **CLPSYCH**. [1]
2. **Develop methods that extract information in different channels.** In this step, it is necessary the analysis of different kinds of information presented in the posts to extract and create separate channels. For example, could be the topics that are contained, the emotions presented or the style of the author for writing or expressing
3. **Develop a model to create a representation that combines the different channels automatically.** This step involves the development of a model that automatically combines the different channels obtained in the step before, and creates a new representation. To overcome this problem using models inspired in Deep Neural Networks that learn to combine and or give importance to a different type of information.
4. **Design an approach that effectively incorporates sequential information in the representation.** Due to the nature of the information that is created involving the sequencing of actions, where a user writes a post one after another. For example, deep learning models like Recurrent Neural Networks that take time and sequence into account.

## 3 State of the Research

This section presents the preliminary work that has been done until now that supports our hypothesis and research proposal.

1. Our first experimental approach consists of the usage of the emotions channel (part of the second step in the methodology); it is proposed a new representation called Bag of Sub-Emotions (BoSE). This channel represents social media documents using a set of fine-grained emotions that are automatically generated using lexical resources based on emotions and sub-word embeddings. To evaluate this representation, two different tasks were used: depression and anorexia detection. The results are promising; the usage of these fine-grained emotions improved the results from a representation based on traditional methods and based on the core emotions. The results obtained are also competitive in comparison to state of the art approaches (Table 1).
2. Temporal analysis for the emotion channel. A first exploration of the temporal information that is presented in the emotion channel. With this analysis we can appreciate the behaviour of the emotions through time of people with a mental disorder.

[1] https://early.irlab.org/ and http://clpsych.org/

**Table 1.** F1 results over the positive class against baseline methods

| Method | Dep'17 | Dep'18 | Anor'18 |
|---|---|---|---|
| BoW-unigrams | 0.60 | 0.58 | 0.69 |
| BoE-unigrams | 0.57 | 0.60 | 0.50 |
| BoSE-unigrams | 0.61 | 0.61 | **0.82** |
| BoW-ngrams | 0.59 | 0.60 | 0.69 |
| BoE-ngrams | 0.61 | 0.58 | 0.58 |
| BoSE-ngrams | **0.64** | **0.63** | 0.81 |
| delta-BoSE | - | 0.53 | 0.79 |
| Late Fusion | - | **0.64** | **0.84** |

3. An early and late fusion of the temporal features with the original BoSE (part of the third step in the methodology). This approach obtains a little increase in the results that using the information separated.
4. An approach inspired in the modeling of fine-grained emotions expressed by the users and deep learning architectures with attention mechanisms for the detection of depression and anorexia.

## 4 Conclusions

In this document we present the research proposal during the Ph.D. program and part of the work that has been made. The main interest of this research is the detection of mental disorders in users through the post in different social media platforms. The work will focus on the detection of these users improving the state of the art results, using a new multichannel representation that exploits traditional natural language process methods combined with deep learning methods. The emotional channel present useful information that helps the detection of mental disorders. BoSE obtained better results than the proposed baselines and also improved the results of only using broad emotions. Incorporating temporal analysis over the emotion channel and combine it with the previous representation demonstrate that helps the detection of users that presents signs of mental disorders. It is worth mentioning the simplicity and interpretability of the representation, creates a more straightforward analysis of the results.

## References

1. Kessler, R., Bromet, E., Jonge, P., Shahly, V., Marsha.: The burden of depressive illness. Public Health Perspectives on Depressive Disorders (2017)
2. Mathers, C., Loncar, D.: Projections of global mortality and burden of disease from 2002 to 2030. PLOS Medicine, Public Library of Science (2006)

3. Aragón, M.E., López-Monroy, A.P., González-Gurrola, L.C., Montes-y-Gómez, M.: Detecting depression in social media using fine-grained emotions. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1 (2019)
4. Aragón, M.E., López-Monroy, A.P., Montes-y-Gómez, M.: INAOE-CIMAT at eRisk 2019: Detecting Signs of Anorexia using Fine-Grained Emotions. Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland. (September, 2019).
5. Tausczik, YR., Pennebaker, JW.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology (2010)
6. Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., Ohsaki, H.: Recognizing depression from twitter activity. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (2015)
7. Schwartz, HA., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., Kosinski, M., Ungar, L.: Towards assessing changes in degree of depression through facebook. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (2014)
8. Coopersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In: Workshop on Computational Linguistics and Clinical Psychology (2014)

ISSN 1870-4069

# Development of a Metaheuristic to Obtain Frequent Similar Patterns

Gretel Bernal-Baró, Ansel Y. Rodríguez González,
Rosa M. Valdovinos Rosas

Universidad Autónoma del Estado de México,
Facultad de Ingeniería,
Mexico

gretelbernalbaro@gmail.com

**Abstract.** Most frequent pattern mining algorithms assume that two sub-descriptions of instances are similar if and only if they are equals. However, other similarity functions are used in the soft sciences. In fact, some algorithms find patterns using similarity functions other than equality, but those algorithms are shown difficulties for mining datasets that involve a large variety of attribute values. Although these algorithms find a complete set of Frequents Similars Patterns (FSPs), the number of patterns mined is often too large, requiring high analysis and processing costs. The research proposed here is focused on developing a metaheuristic for obtaining a representative subset of FSP, which describes the whole dataset, from large mixed datasets in a shorter time than the traditional algorithms used for mining FSPs.

**Keywords:** Frequent patterns, similarity, mixed data, data mining.

## 1 Introduction

A frequent pattern is a combination of feature values of an instance in the dataset that appears with a frequency not less than a user-specified frequency threshold [1]. Most current algorithms for frequent patterns mining assume that two instances are similar when they are equal. However, in many real-world problems, two instances can be considered similar even though they are not identical. In these problems, the concept of similarity between instances descriptions is used to compare instances and count how many times an instance appears in a dataset. When similarity functions different than equality are used, new patterns emerge called Frequent Similar Patterns (FSP). Thus, a FSP is a combination of feature values of the instance in the dataset, such that the accumulation of frequency of its similar patterns is not less than a user-specified frequency threshold. [4].

In the literature several algorithms are proposed for finding FSPs (ObjectMiner [2], STreeDC-Miner [4] and STree NDC-Miner [4]), but these algorithms' behavior is affected negatively when large volume or high-dimensional datasets are used. When the similarity is the criteria

considered to obtain the frequency calculation, more operations need to be done, and as a consequence, the runtime required by these algorithms is increased. Furthermore, although new patterns can be found using these algorithms, the number of mined patterns is often too large, making additional analysis and processing costs. The research here presented proposes to develop a metaheuristic to mine a subset of representative FSPs, which explores the search space more efficiently than the existing algorithms used in mining FSPs. Specifically, the research is focused on: To develop a mechanism for the efficient retrieval of instance sub-descriptions and their frequencies. To propose a quality measure of a FSP that allows obtaining a representative subset of FSPs that describes the dataset. If is required, to adapt the existing theory about the mining frequent patterns to mining FSPs. Finally, to develop a metaheuristic for mining FSPs in mixed datasets.

### 1.1 Motivation and Justification

Several studies demonstrated that, when the similarity concept is included in the frequency calculation, it is possible to discover hidden patterns in the traditional frequent patterns context [2]. Similarly, other studies validate that when using the FSPs in tasks such as classification, the classifier obtains higher precision when new objects are classified. In fact, it is unnecessary to extract the entire set of FSPs; for example, a reduced set of frequent similar patterns without information loss, named Closed Frequent Similar Patterns, can be extracted [3]. In addition, some important difficulties of the FSP mining algorithms are: high computational cost required for finding the solutions, prohibitive in datasets with high dimensionality (more than hundreds of attributes), and too many FSPs obtained from the mining process, resulting in analysis and processing costly.

## 2 Previous Works in the Area

In the state of art, there are several algoritms for obtain frequent pattenrs, however, related with FSP are too few. The referent research could be the CFSP-Miner algorithm [3] which discover a subset of FSPs, but the number of FSPs mined and the runtime can be high too. On the other hand, ObjectMiner [2] was the first algorithm that uses similarity functions for mining frequent patterns, is inspired in the Apriori algorithm, and includes a pruning method to diminish the search space.

STreeDC-Miner [4] works by following a depth-first search strategy, using a tree structure called STree. In the STree each path from the leaf to the root represents a sub-description, where the same sub-descriptions are grouped. Each leaf stores the repetitions of the pattern and also holds the similarity among this pattern and its similar patterns. STreeNDC-Miner [4], like STreeDC-Miner, assumes an order among the features that describe the instances. But, unlike ObjectMiner and STreeDC-Miner, it does not include a method for pruning the search space. However, the computational effort for searching all FSPs is reduced using a top-down strategy and the STree data structure.

## 3 Hypothesis

With the development of a metaheuristic for mining a subset of FSPs, in mixed datasets, a subset of FSPs will be obtained with quality greater than or equal to 75%, in less runtime than the current algorithms that mine FSPs.

## 4 Methodology

To meet the stated objective, it is proposed to follow the following methodology:

1. To obtain from the repository  textit UCI machine learning repository [1] datasets with dimensionality and size required for validate of the proposed metaheuristic.
2. To develop a metaheuristic for mining FSPs.
   - To develop a mechanism for the efficient retrieval of instance sub-descriptions.
   - To define a new fitness function to measure the quality of a FSP.
   - To implement the different metaheuristics used in traditional frequent pattern mining and develop a new metaheuristic to mine a subset of FSPs.
3. To perform tests and compare the results.
   - To evaluate the performance of the proposed metaheuristic according the run-time required and quality of the FSPs obtained.
   - To use the FSPs mined for classification purposes and to evaluate the subsets quality, based on the classifier accuracy.

## 5 State of the Research

The research is at the beginning of the second year. During the first year, a data structure was developed to represent the dataset efficiently. For that, the dataset is divided into several sub-trees, called FP-Similar-Tree. Each FP-Similar-Tree is associated with a FSP and represents a compact structure that stores quantitative information about the FSPs presents in the dataset. Several experiments were carried out using ten datasets from the UCI machine learning repository to evaluate the FP-Similar-Tree behavior. The number of instances in the selected datasets varies between 4,000 and 1,000,000 to assess their performance in different scenarios. To analyze the FP-Similar-Tree improvements, the number of mined patterns and the required run-time was measured concerning the STreeDC-Miner algorithm. This algorithm was selected for being one that shows the best performance FPS mining nowadays.

The results obtained show that both algorithms found the same number of patterns in the datasets tested (ten datasets), except in two. In these two datasets, the proposed algorithm, FP-Similar-Tree, mined all FSPs while

[1] https://archive.ics.uci.edu/ml/index.php

the STreeDCMiner algorithm did not work due to memory requirements. FP-Similar-Tree algorithm got the best results in terms of runtime in the other eight of the ten datasets tested, achieving an improvement of up to 57 %. Therefore, it can be assured that the proposed data structure is an excellent alternative to extract a subset of frequent similar patterns.

## 6 Conclusions

Nowadays, the mining of FSP is strongly attracting attention as an alternative solution in the development of descriptive strategies. The main problem identified in the existing methods is difficult for dealing with high dimensionality data and the large number of mined patterns. About it, the first development made in the research proposed was to build a data structure capable of representing the dataset more efficiently, the FP-Similar-Tree. The structure proposed was compared with the STreeDC-Miner algorithm to analyze its effectiveness. The preliminary results allow us to show that the FP-Similar-Tree mine a whole set of FSP in mixed data collections in lower run-time than the time reached by one of the most competitive algorithm, STreeDC-Miner. The main improvements of the FP-Similar-Tree are: Reduction the number of comparisons made between the sub-descriptions of the dataset and fewer accesses to the dataset because only is required single access. As future work, we plan to develop a quality measure that allows the generation of a subset of representative FSPs. In addition to developing a metaheuristic that uses the data structure and the quality measure proposed.

## References

1. Baró, G. B., Martínez-Trinidad, J. F., Rosas, R. M. V., Ochoa, J. A. C., González, A. Y. R., Cortés, M. S. L.: A pso-based algorithm for mining association rules using a guided exploration strategy. Pattern Recognition Letters, vol. 138, pp. 8–15 (2020)
2. Danger, R., Ruíz-Shulcloper, J., Llavori, R. B.: Objectminer: A new approach for mining complex objects. In: ICEIS (2). pp. 42–47. Citeseer (2004)
3. Rodríguez-González, A. Y., Lezama, F., Iglesias-Alvarez, C. A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., de Cote, E. M.: Closed frequent similar pattern mining: Reducing the number of frequent similar patterns without information loss. Expert Systems with Applications, vol. 96, pp. 271 – 283 (2018) doi: https://doi.org/10.1016/j.eswa.2017.12.018
4. Rodríguez-González, A. Y., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., Ruiz-Shulcloper, J.: Mining frequent patterns and association rules using similarities. Expert Systems with Applications, vol. 40, no. 17, pp. 6823 – 6836 (2013) doi: https://doi.org/10.1016/j.eswa.2013.06.041

ISSN 1870-4069

# Graphs to Face the Class Imbalance Problem in Big Data

A. Guzmán-Ponce[1,2], R.M. Valdovinos-Rosas[1], J.S. Sánchez-Garreta[2], J.R. Marcial-Romero[1]

[1] Universidad Autónoma del Estado de México, Facultad de Ingeniería, Mexico

[2] Universitat Jaume I, Castelló de la Plana, Institute of New Imaging Technologies, Department of Computer Languages and Systems, Spain

aguzmanp643@alumno.uaemex.mx, {rvaldovinosr,jrmarcialr}@uaemex.mx,sanchez@uji.es

**Abstract.** Class imbalance is one of the data complexities widely studied in the field of Data science. The class imbalance problem occurs when one class is strongly over-represented in comparison with the other classes, biasing the learning towards the most represented class. Due to the large volume of data that needs to be processed in Big Data context, is imperative to clean the data for diminish the volume and improve the results. In this way, graph theory is becoming a popular technique in data science to given solutions in real-world problems by transforming them in terms of vertices and edges. In order to face the class imbalance problem, we proposed a graph-based under-sampling method. This method was experimentally validated by using a collection of two-class imbalanced big data sets. The experimental results show a competitive in the classification performance measured by the geometric mean when we compare to several state-of-the-art methods.

**Keywords:** Big data, graph theory, class imbalance, under-sampling.

## 1 Introduction

The notion of Big Data is a consequence of the fast generating data, thus brings challenges for the class imbalance problem due to the Big Data characteristics (Volume, Velocity, Variety, Veracity and Value). In this sense, the class imbalance problem in Big Data became a challenging gap to develop strategies in order to give a solution without reducing the classifier performance.

In a binary data set is said to be imbalanced when one of the classes has a lower number of instances than the other, called minority class or positive class ($C^+$), while the class with a high number of instances is named as majority class or negative class ($C^-$) [4].

In Big Data context, the most common strategy proposed to deal the class imbalance problem have been the resampling techniques, which are divided into two solve direction: over-sampling consist on increasing the number of instances in the $C^+$ and the under-sampling, which remove instances from the $C^-$. This research is focused on the under-sampling approach because recent studies on Big Data [4] have shown that under-sampling methods produce better results than over-sampling methods. Moreover, much data is not necessary, so reducing the data size in Big Data sets became a need.

Fostered by the fast proliferation of the use of graph-theory in data science, where the aim is extracting knowledge from graph topologies, for instance, the clustering communities or feature selection proposed both of them consider weighted graphs [3,6]. Faced with this reality, we introduce the use of graph theory for facing the class imbalance problem in Big Data. The main contribution of this research is: *The use of graph theory to obtain an induced subgraph, which allows getting the borderline of the negative class in Big Data sets to face the class imbalance problem.*

## 2 Related Work

Most of the existing methods in Big Data have been developed with MapReduce, which is composed of two functions that let parallelize processing data: the Map function split the data into different subsets of data, and the Reduce function which fuse the local outputs into a single final result [5].

Some researches have been conducted through scaling well-know resampling methods [5]. The classic over-sampling method ROS that replicates randomly instances from $C^+$ until has the same size of $C^+$, and the classic under-sampling method RUS that removes randomly instances from $C^-$ until has the same size of $C^+$. Del Río et al. [5] scaled the above algorithms mentioned through MapReduce. In ROS, the Map function randomly balances the positive class by replicate instances, while the Reduce function took all results generated by each Map function and randomly took instances to balance the data set. In the RUS algorithm, the Map function groups all the instances by class, while Reduce collects and balances by removing randomly negative class instances.

In Big Data there are two proposals that use the SMOTE algorithm (Synthetic Minority Oversampling TEchnique) [2]: SMOTE-MR [2] uses the same partition that an instance belongs for computing the $k$-neighbors from an instance from the positive class, and SMOTE-BD [1] which principal difference consists of computing the $k$ nearest neighbors based on the *kNN-IS*.

## 3 Hypothesis or Research Objectives

With the use of graph theory to obtain an induced subgraph can be getting the borderline of the negative class in Big Data sets to face the class imbalance problem in terms of performance of a classifier.

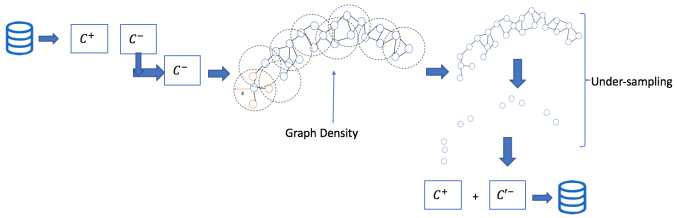



**Fig. 1.** The steps of Graph-based under-sampling method.

## 4 Methodology

To achieve the aim of this research, the methodology followed is described as follows:

- Acquisition of data: For carried out the experiments, we use 12 two-class imbalanced Big Data sets taken from the UCI Machine Learning repository, that consider different imbalance ratio (i.e the ratio of the number of instances from the $C^-$ to the number of instances from the $C^+$). The biggest repository has 4954752 instances and 28 features.
- Data manipulation: In this phase the under-sampling proposal was developed. For that, we obtain an induced subgraph, which allows getting the borderline of the $C^-$ (Figure 1).
  In general, the proposal consists on create a weighted graph based on the density of the data set. Given $v \in C^-$ a vertex, an edge is built according to *eps-neighborhood of an instance* $p$ such that $N_{eps}(p) = \{q \in C^- | dist(p,q) \leq eps\}$, where $dist$ is the Euclidean distance. Thus an edge is built from $p$ to each instance in $N_{eps}(p)$ and their corresponding weighted is given by the dist, this procedure uses MapReduce by using the same partition that an instance belongs.
- Analysis of data: In order to compare the performance behavior of under-sampling proposal on Big Data, we carried out the experimental study with rebalancing methods: RUS, ROS, SMOTE-MR, and SMOTE-BD.

## 5 State of the Research and Preliminary Results

For each Big Data set, Table 1 reports the geometric mean (averaged across the 5 fold cross-validation) of the Decision Tree classifier from the MLlib Spark API, along with the Friedman's average rank for each algorithm.

According to Friedman's average rank, the best resampling method for the Big Data sets used is ROS, this suggests that duplicate instances extend the sample space, however, this increases the data size. Notwithstanding, as we can see, the top three methods are composed by under-sampling techniques such as RUS and our proposal. Thus, these results show that the proposed method provides competitive advantages with respect to other state-of-the-art methods.

**Table 1.** mean results obtained per data set, the best performance value is stressed in bold.

| Data set | Baseline | ROS | RUS | SMOTE-MR | SMOTE-DB | GraphDensity |
|---|---|---|---|---|---|---|
| poker1 | 33.7 | **53.3** | 52.9 | 42.5 | 33.7 | 51.8 |
| SEA | 82.0 | 82.9 | 82.9 | 82.9 | 82.0 | **83.0** |
| Agrawal | 94.4 | **95.1** | 95.0 | 94.4 | 94.6 | 94.8 |
| MiniBooNE | 85.2 | 87.9 | 88.0 | 87.9 | **88.1** | 84.0 |
| Susy | 68.8 | **76.8** | 76.7 | 76.3 | 76.3 | 76.7 |
| Click | 16.2 | 62.1 | **62.2** | 56.2 | 56.8 | 61.9 |
| poker0 | 17.0 | **58.1** | 56.2 | 59.7 | 53.9 | 56.2 |
| HEPMASS | 71.9 | **83.3** | **83.3** | 66.8 | 65.5 | 83.0 |
| HIGGS | 11.6 | 66.0 | 65.8 | 64.4 | 64.9 | **66.1** |
| Covtype | 72.8 | **93.2** | **93.2** | 92.6 | 93.0 | **93.2** |
| Credit | 87.7 | 91.9 | 91.3 | **93.0** | 92.7 | 90.0 |
| RLCP | 10.3 | **93.2** | **93.2** | 93.0 | 93.1 | **93.2** |
| Avg. rank | 5.63 | **1.92** | 2.33 | 3.96 | 4.08 | 3.08 |

## 6 Conclusions

In this short paper, a graph-based method to face the class imbalance problem in Big Data is proposed. The obtained results show the potential of graph theory to face this problem. So, the future work is focusing on improving the performance of this proposal through the study of other intrinsic data characteristics. In the same way, an experimental comparison of the behaviour of this and another resampling methods will be addressed for concluding the PhD Thesis.

## References

1. Basgall, M., Hasperué, W., Naiouf, M., Fernández, A., Herrera, F.: An Analysis of Local and Global Solutions to Address Big data Imbalanced Classification: A Case Study with SMOTE Preprocessing. In: Naiouf, M., Chichizola, F., Rucci, E. (eds) Cloud Computing and Big Data. pp. 75–85. Springer International Publishing, La Plata, Argentina (2019)
2. Basgall, M. J., Hasperué, W., Naiouf, M., Fernández, A., Herrera, F.: SMOTE-BD: An Exact and Scalable Oversampling Method for Imbalanced Classification in Big Data. Journal of Computer Science and Technology, vol. 18, no. 3, pp. 23–28 (2018)
3. Newman, M. E. J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E, vol. 69, pp. 026113 (Feb 2004)
4. Rendón, E., Alejo, R., Castorena, C., Isidro-Ortega, F. J., Granda-Gutiérrez, E.: Data Sampling Methods to Deal With the Big Data Multi-Class Imbalance Problem. Applied Sciences, vol. 10, no. 4, pp. 1276 (2020)
5. del Río, S., López, V., Benítez, J. M., Herrera, F.: On the use of MapReduce for imbalanced Big Data using Random Forest. Information Sciences, vol. 285, pp. 112–137 (2014)
6. Zhang, Z., Hancock, E. R.: A graph-based approach to feature selection. In: Jiang, X., Ferrer, M., Torsello, A. (eds) Graph-Based Representations in Pattern Recognition. pp. 205–214. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)

ISSN 1870-4069

# Data-Driven Methodology for Candidate Well Selection and Ranking

Héctor De Paz Galicia[1], Eduardo Roldán Reyes[1], Leonid Sheremetov[2]

[1] Tecnológico Nacional de México,
Instituto Tecnológico de Orizaba, Veracruz,
Mexico

[2] Instituto Mexicano del Petróleo,
Mexico

hdepazg@ito-depi.edu.mx

**Abstract.** This research work proposes a new data-driven methodology that helps at different stages of wells productivity analysis to increase the production in oil and gas fields. The first two stage is aimed in developing a data-driven methodology for the identification, selection and ranking of candidate wells for productivity analysis based on the historical oilfield data. The methodology integrates several supervised and unsupervised machine learning techniques, multi-criteria decision analysis and case-based reasoning.

**Keywords:** Candidate-wells selection and ranking, data mining, TOPSIS, genetic algorithm, NARX neural network, CBR.

## 1 Introduction

Wells productivity analysis (WPA) has been seen as a main strategy for increasing production in oil and gas fields [1]. Production analysis of historical data evidence that different workovers within the same oilfield produce very non-uniformly, and up to 30% may not be producing at all due to a combination of reservoir heterogeneity, geomechanical, workover design and deployment factors. Diverse studies on WPA, underline the effectiveness of the analysis and the quality of the results, which are based on three critical points: the choice of the variables involved in the process, the candidate-wells selection (CWS) and the workover design.

WPA begins with the selection and prioritization of wells that have the best characteristics to increase the probability of success. However, the main difficulties that arise during this analysis, is the large number of wells to study, the necessity to examine the high volumes of data to identify opportunities for production of wells operating, which require considerable effort and a significant investment of resources using conventional numerical analysis methods [2]. Therefore, a data-driven methodology that streamlines these information processing operations is a time challenge.

## 2 Related Work

Recently many efforts have been focused on different processes of well productivity. Analyzing candidate well selection (CWS) for one of such processes, hydraulic fracturing (HF), Zoveidavianpoor and Gharibi classified the methods into conventional and advanced [3]. Unfortunately, conventional approaches to CWS are limited by the experience, expertise and preconceived notions of the specialist who is conducting the evaluation. Another limitation of these techniques is the time required to perform the analysis. Advanced methodologies represent an alternative approach to the selection and ranking of wells. Particularly, by the opportunity to examine the data and their relationships in different ways, maximizing the potential of the data. A portfolio of advanced algorithms includes analytic hierarchy process (AHP), decision trees, random forest, ANNs, linear regression and support vector machines (SVM) [4, 5]. This classification can be extended to other well productivity processes like the stimulation treatment and workovers, where gradient boosting and probabilistic expert system technique have been used [6, 7]. Although data mining and AI methods have been reported for CWS, these have only been focused on particular processes like HF, workovers, stimulation, or refracturing. An integrated data-driven methodology for WPA have not been reported in the literature so far.

## 3 Methodology

The proposed data-driven CWSR methodology follows the main stages of a process of the productivity analysis, which begins with the acquisition of information and ends with the proposed solution to increase wells' production:

1. Data acquisition, pre-processing, and transformation.
2. CWS for productivity analysis.
3. OPS analysis and diagnostics.
4. Optimization, selection and ranking of workover proposals.
5. Conceptualization and knowledge base.

The first stage can be considered as a typical stage of any data analytics project. Supervised and unsupervised machine learning techniques are used for data pre-processing and candidate well selection, as well as estimation of the type of intervention (1, 2). A forecasting model based on Nonlinear Autoregressive network with eXogenous inputs (NARX) is capable of predicting the response that the oil production will have, in the following three months after the treatment (3). A Multi-Criteria Decision Analysis (MCDA) method called a Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is applied for selected wells' ranking (4). The measures borrowed from information retrieval, cumulative gain and discounted cumulative gain (DCG), are applied under the premise that the first well from the list of candidates has the highest probability of success.

A genetic algorithm is proposed to optimize the weights applied in TOPSIS. Finally, a case-based reasoning (CBR) will allow for a weighted evaluation of the workover proposals based on previous experiences and best practices (5).

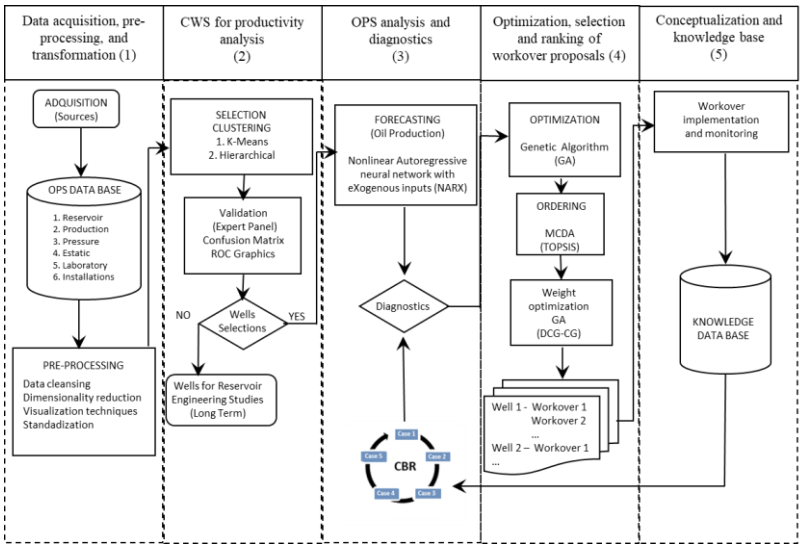



**Fig. 1.** Proposed methodology for the selection and ranking of wells.

## 4 Results and Discussion

To test the proposed methodology, a mature field located in the southeast of Mexico is considered. For the case study, 50 wells producing from the Mesozoic formation were selected with a total of 700 workovers and the dimensions of the original data sample with 38 variables and 11381 observations. The first four stages of the methodology have been developed. At the first stage, the α-trimmed mean ($\alpha = 0.2$) technique was used for the outliers' detection. Data transformation included both the definition of additional characteristics to be taken into consideration, dimensionality reduction with PCA, variables' selection and scaling. After dimensionality reduction, nine variables were selected.

The second stage deals with CWS, where both K-means and hierarchical clustering algorithms were used. In order to validate the obtained results, the candidate wells' list generated by the experts' panel was used as the reference solution. As a result of CWS, 30 wells were selected as candidates for further analysis and ranking and 20 as non-candidates with the values of F1 Score 0.95 and 0.933 for the hierarchical clustering and K-Means respectively. At the third stage, wells' models were developed for well productivity analysis based on available historical data in order to predict expected outcome of the workover for each well from the list. The forecast tool helped to identify the future event by providing the conditions (estimated variables) to carry out the diagnosis of the well. Time series forecasting techniques were used at this stage [8]. Finally, for wells' ranking, multi-objective data analysis (TOPSIS) was integrated with the evolutionary algorithm to optimize the parameters of the developed models. With the use of GA with the DCG-based fitness for the optimization of the weights, a clear correlation with experts' panel is obtained ($R^2 = 0.99$), not the case with the CG-based fitness. The differences in the ranking found by TOPSIS and the experts' panel can be

attributed to the subjectivity of the experts in assigning the weights on the selected criteria; in the same way, the results of the relative proximity indicate that, for values that are very close or equal, the allocation of the ranking of the wells becomes indistinct.

## 5 Conclusions and Future Work

The selection and ranking of wells should be considered as a fundamental part in the methodology of well productivity. The application of the hybrid approach based on the integration of data analytics, MCDA, and evolutionary optimization to support decision making, is an alternative that offers efficient solutions to solve this problem. The approach allows to obtain acceptable results in a short time, reducing the analysis time for the CWSR by 78% in comparison with conventional methods. At the moment of writing, the experiments on productivity analysis and forecasting are in their final steps. Meanwhile, the CBR model, which will permit to capitalize knowledge by reusing past experiences, has been under development. To increase its efficiency, the proposal of a similarity measure relying on the spherical indexing algorithm is envisaged. This will reduce the adaptation effort since this task is a common issue to be solved.

## References

1. Hendraningrat, L., A Salam, I.H., Mohsin, N., Wah, T.S., Rinadi, M., Sedaralit, M.: Enhancing and accelerating reservoir modelling and history matching of an offshore mature oilfield with geological and production system complexity using data-driven approach. In: Proceeding of the International Petroleum Technology Conference (2020)
2. Enciso-Paredes, E.J., de Antuñano-Muñoz, Y., Pérez-Herrera, R., Perera-Pérez L.M.: Metodología para identificar oportunidades de producción en campos maduros. Ingeniería Petrolera, 59(2), pp. 18 (2019)
3. Zoveidavianpoor, M., Abdoullatif, G.: Applications of type-2 fuzzy logic system: handling the uncertainty associated with candidate-well selection for hydraulic fracturing. Neural Computing and Applications, 27(7), pp. 1831–51 (2016)
4. Temizel, C., Purwar, S., Abdullayev, A., Urrutia, K., Tiwari, A.: Efficient use of data analytics in optimization of hydraulic fracturing in unconventional reservoirs. In: Proceeding of the Abu Dhabi International Petroleum Exhibition and Conference (2016)
5. Wang, Y., Salehi, S.: Refracture candidate selection using hybrid simulation with neural network and data analysis techniques. Journal of Petroleum Science and Engineering, 123, pp. 138–46 (2014)
6. Makhotin, I., Koroteev, D., Burnaev, E.: Gradient boosting to boost the efficiency of hydraulic fracturing. Journal of Petroleum Exploration and Production Technology, 9(3), pp. 1919–25 (2019)
7. Zangl, G., Neuhofer, L., Zabel, D., Tippel, P., Pantazescu, C.I., Krcmarik, V., Krenn, L., Hachmöller, B.: Smart and automated workover candidate selection. In: Proceeding SPE Intelligent Energy International Conference and Exhibition (2016)

8. Sheremetov, L., Cosultchi, A., Martínez-Muñoz, J., Gonzalez-Sánchez, A., Jiménez-Aquino, M.A.: Data-driven forecasting of naturally fractured reservoirs based on nonlinear autoregressive neural networks with exogenous input. Journal of Petroleum Science and Engineering, 123, pp. 106–19 (2014)

ISSN 1870-4069

# Incremental Learning Models for Identifying Imagined Words in Continuous EEG Signals

Tonatiuh Hernández-del-Toro, Carlos Alberto Reyes-García

Instituto Nacional de Astrofísica Óptica y Electrónica,
Puebla, Mexico

{tonahdz, kargaxxi}@inaoep.mx

**Abstract.** In the task of building a fully asynchronous Brain Computer Interface that takes imagined words in continuous EEG signals as input, it is needed to be able to detect when an imagined word starts and finishes. In this work, two algorithms are studied. The first one to detect the onset and ending of imagined words in continuous EEG signals. The results showed an average True Positive Rate of 0.55, 0.63 and 0.69 for detecting the onset of imagined words using a Time Error Tolerance Region of 2,2.5 and 3 seconds respectively. For the ending of imagined words, the results showed an average True Positive Rate of 0.58, 0.64 and 0.71 using a Time Error Tolerance Region of 2,2.5 and 3 seconds respectively. In the second algorithm, an incremental learning approach is proposed to detect idle states versus imagined words, the results showed that in more than 75% of subjects, the incremental learning model performed better than the batch learning model.

**Keywords:** Imagined speech, incremental learning.

## 1 Motivation

In recent decades, intensive scientific research has been done in the area of Brain Computer Interfaces (BCI) [1] to develop devices that can improve the living conditions of people with disabilities.

Among the ways a BCI can be implemented [2] are the ones based on imagined words in electroencephalogram (EEG) signals [10]. However, in order to build a fully asynchronous BCI based on imagined words, many problems need to be solved, among them is the problem of detecting the onset and ending of the words and the problem of implementing incremental learning models in order to face the changes in EEG signals due to the natural neuro-plasticity of human brain.

## 2 Previous Works in the Area

To the best of our knowledge, there is not any work that solves the problem of detecting the onset and ending of imagined words using incremental learning models. However, there exist some works that can lead the way.

Some works face the problem of identifying imagined sound production vs. idle states. In [5] they investigated a method for classification of siren sound covert production and the idle state in an off-line system. Wavelet packet decomposition was employed for feature extraction and a Support Vector Machine (SVM) was used for classification. they reported an average True Positive Rate (TPR) of 79.2% for five subjects. In [6], they use sound imagery in a self-paced BCI. They implemented an online interface where the subject tried, by imagining a high pitched tone, to open a message that randomly appeared on a screen. Before the message randomly appeared, the screen was either playing a movie or the subject was reading a text, this was to simulate common tasks in daily life. Autoregressive coefficients, band power, common spatial patterns and discrete wavelet transform were used for feature extraction to cover all time, frequency, and spatial domains. Linear Discriminant Analysis was used for classification. The averaged TPR with six subjects was 88.9% in the watching video scenario and 78.9% in the reading text case. The average False Positive Rates (FPR) were 4.2% and 3.9%, respectively.

There are also works that deal with the problem of identifying imagined words vs. idle states. In [7], they attempted to classify imagined words states against two non linguistic states: relaxed and visual attention. They used features from spatial domain and time domain. In [8], they studied classification between idles states and imagined words. They use two corpus and two feature extraction methods: Wavelet energies and statistical values. They used three classifiers: Random Forest (RF), SVM and Naive Bayes. The higher accuracies they reported for the first corpus was 83% with the statistical features and the RF classifiers. For the second dataset, the higher accuracy they reported was 91% with the RF classifier and the statistical features also.

On the problem of detecting the onset of imagined words, there are some approaches that deal with the Onset detection of movement imagery [9]. And with the problem of Onset detection of high pitch sound imagery [10] using a Timing Error Tolerance Region (TETR).

## 3 Hypothesis and Research Objectives

The hypothesis driving this research is:

*An incremental learning approach can lead to a better adaptive performance of personalized classification models in the task of detecting from continuous EEG signals, the onset and ending of imagined words.*

The main objective of this research is:

*To design and develop an incremental learning model capable of identifying from continuous EEG signals, when a subject starts and finishes to imagine a word.*

The specific goals of the research are:

1. Find the features in which the information needed to discriminate when a subject's mind is in a state non relevant to the BCI, and when the subjects imagines a word is encoded.
2. Design a model capable to identify when a subject starts and finishes to imagine a word in continuous EEG signals.
3. Design a model capable of detect imagined words in the incremental learning approach in a way that is able to learn from new samples given by the user.

## 4 Methodology

1. **Classification of imagined words vs. non relevant states:** Experiment several feature extraction methods (time and frequency domain). Test several models (RF, SVM, ANFIS).
2. **To identify the onsets and endings of imagined words:** Design algorithm to evaluate the signal sequentially. Use of metrics for onset and ending detection (TFPR, TF, ROC, etc).
3. **To design models in an incremental learning approach:** Make use of already available algorithms (SVM, MLP). Design new algorithms in incremental learning way (RF, ANFIS).

## 5 State of the Research

The research is in its second year. Two algorithms are studied and performed using a dataset of 27 subjects that imagine 5 different words.

The first algorithm is focused on identifying the onset and ending of imagined words in a continuous EEG signal. The feature extraction consists of calculating The Generalized Hurst Exponent (GHE) [11] and a RF classifier is trained in order to classify the signal sequentially in windows of 1 second and determine when the imagined words start and finish.

The second algorithm is focused on studying the performance of incremental learning models in the task of identifying segments as imagined words vs. idle states. Instant Wavelet Energy and GHE are calcualted as features and a SVM is trained in both approaches (Batch learning and Incremental learning).

In order to finish the research, it is still needed to implement an algorithm that solves the problems of the two algorithms currently being studied. It is needed to have a better resolution in order to reduce the TETR and being able to have a better detection of the onset and ending of imagined words. Though the incremental learning algorithms currently studied outperform the batch learning approach, they still achieve poor accuracy in some subjects which suggest to try another incremental learning models or other feature extraction methods. Finally the resulting model must be tested with other data sets.

## 6 Results

In the task of detecting the onset of imagined words. The results reported an average TPR of 0.55, 0.63 and 0.69 with a TETR of 2, 2.5 and 3 seconds respectively.

In the task of detecting the ending of imagined words the same feature set was calculated. The results showed an average TPR of 0.58, 0.64 and 0.71 with a TETR of 2, 2.5 and 3 seconds respectively.

In the test of performance of incremental learning models in the task of classifying segments as imagined words or idle states, the incremental learning approach showed after 80 samples, a higher accuracy vs the batch learning approach.

## References

1. Wolpaw, J.R., McFarland, D.J.: Brain-computer interfaces for communication and control. Clinical Neurophysiology, 113(6), pp. 767-–791 (2002)
2. McFarland, D.J., Wolpaw, J.R.: EEG-based brain–computer interfaces. Current Opinion in Biomedical Engineering, 4, pp. 194—200 (2017)
3. Brigham, K., Vijaya-Kumar, B.V.K.: Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepathy. In: 4th International Conference on Bioinformatics and Biomedical Engineering, (iCBBE), pp. 1—4 (2010)
4. Alsaleh, M.M., Arvaneh, M., Christensen, H., Moore, R.K.: Brain-computer interface technology for speech recognition: A review. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1—5 (2017)
5. Song, Y.J., Sepulveda, F.: Classifying siren-sound mental rehearsal and covert production vs. idle state towards onset detection in brain-computer interfaces. In: International Winter Conference on Brain-Computer Interface (BCI) (2015)
6. Song, Y.J., Sepulveda, F.: A novel onset detection technique for brain-computer interfaces using sound-production related cognitive tasks in simulated-online system. Journal of Neural Engineering, 14(1) (2017)
7. Alsalehl, M.M., Christensen, H., Moore, R.K., Arvaneh, M.: Discriminating between imagined speech and non-speech tasks using EEG. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS) (2018)
8. Moctezuma-Pascual, L.A.: Distinción de estados de actividad e inactividad ling´ística para interfaces cerebro computadora. Tesis de Maestria Benemérita Universidad Autónoma de Puebla (2017)
9. Townsend, G., Graimann, B., Pfurtscheller, G.: Continuous EEG classification during motor imagery–simulation of an asynchronous BCI. In: IEEE Transactions on Neural Systems and Rehabilitation Engineering, 12(2), pp. 258–65 (2004)
10. Song, Y.J., Sepulveda, F.: An online self-paced brain-computer interface onset detection based on sound-production imagery applied to real-life scenarios. In: International Winter Conference on Brain-Computer Interface (BCI), pp. 46-–49 (2017)
11. Di Matteo, T.: Multi-scaling in finance. Quantitative Finance, 7(1), pp. 21-–36 (2007)

ISSN 1870-4069

# Cellular Evolutionary Algorithms with Estimation of Distribution

Yoan Martínez-López[1], Julio Madera[1], Ansel Rodríguez[2]

[1] University of Camagüey,
Cuba

[2] Centro de Investigación Científica y de Educación Superior de Ensenada,
Unidad de Transferencia Tecnológica UT3,
Mexico

{yoan.martinez, julio.madera}@reduc.edu.cu, ansel@cicese.mx

**Abstract.** In this research, a class of evolutionary algorithms (EA), called Cellular Estimation of Distribution Algorithms (CEDA) are proposed for discrete and continuous optimization problems. The main intention is to study the efficiency of EDA from two perspectives: to decrease the number of objective function evaluations (evaluative efficiency) and the execution time of the algorithm (time efficiency). First, a study is reported on the use of Bayesian models that use constraint-based independence detection for learning the structure of the probabilistic model. As a result, three algorithms are created: CEBA, CEGA and CUMDANCAUCHY, which were validated with different datasets.

**Keywords:** Cellular EDA, evolutionary algorithms, learning, probabilistic graphical models.

## 1 Introduction

Evolutionary Computation (EC) is one of the branches of Artificial Intelligence applied to solve combinatorial optimization problems, which is inspired by the mechanisms of biological evolution proposed by Darwin, Medel and Lamark where Darwin proposed the "Natural Selection of the Most Adapted", Medel proposed the "Corpuscular Theory of Inheritance" and Lamark proposed the "Inheritance of Acquired Characteristics" [1, 3]. EDAs (Estimation Distribution Algorithms) are a group of evolutionary algorithms that allow fitting the model to the structure of a given problem, made by an estimation of probability distributions from selected solutions. The model reflected by the bias is a probability distribution [2]. Cellular evolutionary algorithms belong to the class of evolutionary algorithms of discrete groups based on spatial structures, where each individual interacts with his or her adjacent neighbor. An overlapping neighborhood assists in the exploration of the search space, while exploitation takes place within a neighborhood by stochastic operators [2]. A cellular EDA is a collection of decentralized collaborative collections of EDAs, also called member algorithms, that develop overlapping populations, where the organization of cellular EDAs is based on

the traditional 2D (two-grid) structure of overlapping neighbors, where one grid contains chains, and the other contains disjointed sets of chains (cells) [2].

## 2 Research Question

How to use structured populations in EDA to building efficient algorithms from the fitness function evaluation point of view?

## 3 General Objective

To develop EDA algorithms with evidence-based learning of independence and decentralized schemes to reduce the number of evaluations in solving discrete and continuous optimization problems.

## 4 Specific Objectives

To create decentralized (cellular) EDA models that are more efficient evaluative than the centralized one.

To create decentralized EDA algorithms that use independence tests in learning to reduce the number of evaluations of state-of-the-art algorithms that use metrics optimization.

## 5 Hypothesis

Developing decentralized EDA algorithms that use independence testing in learning increases assessment efficiency with respect to centralized EDA.

## 6 Proposed Learning Strategies

Three learning strategies are proposed in this research:

1) Learning the structure and parameters from local populations.
2) Learning structure from the global population and parameters from local populations.
3) Learning structure from the global population, but calculating the necessary statistics locally and then integrating the results with an appropriate statistical method.

## 7 Contributions

Our contribution was the development of three algorithms: CEBA (Cellular Estimation Bayesian Algorithm), CEGA (Cellular Estimation Gaussian Algorithm), and

**Table 1.** Result of the first five places of GECCO/WCCI 2020[1].

| Algorithm | Ranking Index |
|---|---|
| CUMDANCauchy++ | 5.48 |
| DEEDA | 5.41 |
| CBBO, Cauchy and DEEPSO | 5.34 |
| HFEABC | 5.00 |
| DE-TLBO | 4.54 |

CUMDANCAUCHY (Cellular Univariate Marginal Distribution Algorithm with Normal-Cauchy distribution), which were evaluated with different datasets of discrete and continuous domain. Moreover, CUMDANCAUCHY was applied in some real-world problems, such as: energy resource management in smart grid and CVRP (Capacited Vehicle Routing Problem) problem applied health resources.

## 8 Experiments and Results

We used three domains for performing experiments: benchmarks discrete function, continuous function, and real-world problem.

**Discrete functions benchmark.** The preliminary experiments on discrete functions verify the effectiveness of the CEBA. The aim of these benchmark functions is to test the performance of the discrete optimization algorithms. The IsoPeak, FirstPolytree3, OneMax, Deceptive3 and Plateu functions were used. All results are shown in [9].

**Continuous functions benchmark.** The behavior of the proposed algorithm CEGA is evaluated using a comparison in terms of the approximation to the optimum and the numbers of iterations and evaluations needed for different neighborhoods are presented over the continuous functions: Griewangk, Rastrigin's, Rosenbrock's, Sphere and Ackley's. CEGA is compared with the other continuous EDAs reported in the literature for the same continuous functions. All results are shown in [7].

**Real-world problem.** To test the proposed algorithm, the framework developed for the competition: "Evolutionary Computation in Uncertain Environments: A Smart Grid Application" was used [4, 5]. CUMDANCauchy outperforms the winner's algorithms of the 2018 edition, emerging as a good tool to solve the ERM problem under uncertainty. Moreover, EDAs and hybrid EDAs methodologies achieved good results compared with the baseline algorithm (DE). In general, the experimental results showed that EDAs and hybrid EDAs methodologies are good tools to solve the ERM problem under uncertainty. Next, experimentation with 15 models to study the behavior of the algorithms applied to the problem, using the CVRP library, implemented in MATLAB was done. Three metaheuristics models were implemented: Estimation of Distributions Algorithms, Simulated Annealing and Variable Neighborhood Search. From the modelled problem, it was proceeded to solve the FSMVRPTW problem using EDA, SA and VNS algorithms [3]. The studies about the CVRP have demonstrated its usefulness in different complex situations as pandemic, to optimize the distribution of

[1] http://www.gecad.isep.ipp.pt/ERM-competitions/2020-2/

resources. Table 1 shows the result of the GECCO/WCCI 2020 competition, taking into account the Ranking Index, where CUMDANCauchy++ and DEEDA were the best algorithms [4, 5].

## 9 Conclusions

Cellular EDAs with evidence-based learning of independence may increase the evaluative efficiency of the EDA by the way the population is structured and the interaction between the different algorithms. As future work, adapting cEDA for practical problems where there are dependencies amongst the variables and the number of evaluations of the fitness functions is restricted will be an interesting research direction. Another interesting work will be comparing cEDA with other algorithms like Differential Evolution Algorithm, Particle Swarm Optimization and Firefly Algorithm in this kind of problems.

## References

1. Darwin, C.: El origen de las especies por medio de la selección natural. 2, Calpe (1921)
2. Alba, E., Madera, J., Dorronsoro, B., Ochoa, A., Soto, M.: Theory and practice of cellular UMDA for discrete optimization. Parallel Problem Solving from Nature, Springer, pp. 242–251 (2006)
3. Harik, G.R., Lobo, F.G., Goldberg, D.E.: The compact genetic algorithm. In: IEEE transactions on evolutionary computation, 3(4), pp. 287–29 (1999)
4. Martínez-López, Y., Rodríguez-González, A.Y., Quintana, J.M., Moya, A., Morgado, B., Mayedo, M.B.: CUMDANCauchy-C1: A cellular EDA designed to solve the energy resource management problem under uncertainty. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp. 13–14 (2019)
5. Martínez-López, Y., Rodríguez-González, A.Y., Quintana, J.M., Mayedo, M.B., Moya, A., Santiago, O.M.: Applying some EDAs and hybrid variants to the ERM problem under uncertainty. In: Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion, pp. 1–2 (2020)
6. Martínez-López, Y., Madera-Quintana, J., Leguen-de Varona, I.: Algoritmos evolutivos con estimación de distribución celulares. Revista Cubana de Ciencias Informáticas, 10, pp. 159–170 (2016)
7. Martínez-López, Y., Madera, J., Rodríguez-González, A.Y., Barigye, S.: Cellular estimation gaussian algorithm for continuous domain. Journal of Intelligent & Fuzzy Systems, 36(5), pp. 4957–4967 (2019)
8. Martínez-López, Y., Oquendo, H., Caballero, Y., Guerra-Rodríguez, L.E., Junco-Villegas, R., Benítez, I., Rodríguez, A., Madera, J.: Aplicación de la investigación de operaciones a la distribución de recursos relacionados con la COVID-19 en Cuba. Retos de la Dirección, 14(2), pp. 87–106 (2020)

9. Martínez-López, Y., Madera, J., Mahdi, G.S.S., Rodríguez-González, A.Y.: Cellular estimation Bayesian algorithm for discrete optimization problems. Revista Investigación Operacional (2020)

ISSN 1870-4069

# Cybersecurity on Transactions in Smart Metering Systems Using Blockchain

Juan C. Olivares-Rojas, Enrique Reyes-Archundia,
José A. Gutierrez-Gnecchi

Tecnológico Nacional de México / Instituto Tecnológico de Morelia,
México

{juan.or, enrique.ra, jose.gg3}@morelia.tecnm.mx

**Abstract.** Smart Metering Systems (SMS) provide real-time monitoring of energy consumption and production, enabling advanced services such as automatic connection and disconnection, demand-response events, and dynamic pricing. The most widely implemented SMS is the Advanced Metering Infrastructure (AMI). Despite their advantages, SMS face significant challenges, particularly in cybersecurity, including data manipulation, false information injection, and communication interruptions. Furthermore, SMS data is not fully exploited for advanced analytics, which could support applications such as energy theft detection, power quality monitoring, and fault prediction.This work proposes an extended AMI architecture that integrates a multi-tier blockchain framework with embedded data analytics capabilities. The proposed approach ensures secure and reliable energy transactions while enabling intelligent services directly at the smart meter and edge level. The research introduces a novel blockchain consensus algorithm, Proof-of-Efficiency (PoEf), combined with time-series forecasting, statistical methods, and reinforcement learning for anomaly detection and energy efficiency optimization. Experimental progress includes prototype development, security enhancements, and initial data analytics implementations to support fraud prevention, energy quality assessment, and failure detection in electrical networks.

**Keywords:** Smart metering systems, advanced metering infrastructure, cybersecurity, blockchain, proof-of-efficiency, data analytics, machine learning, energy theft detection, fog-edge-cloud computing.

## 1 Introduction

The Smart Metering Systems (SMS) allow the final-user, real-time monitoring of their energy consumption through Smart Meter (SM). The SM measures the energy consumption and energy production, and reports this data to the utilities. Also, SMS allows the automatic connection and disconnection, demand-response events and dynamic electrical price according to offer and demand [1]. The most implemented SMS is the Advanced Metering Infrastructure (AMI) [2].

Despite the enormous advantages of SMS, it has a lot of challenges and opportunities; one of them, it is related with the data cybersecurity. There are a lot of threats on cyber security in SMS: the interruption of the measurement (disconnection of the meter, deletion of the event log), investment of the meter (for less consumption data record), the deletion of records, the alteration of stored data, the interruption of communications to prevent data from being reported, the tamper of consumption data "on the fly" when they are reported, as well as the injection of false information (for example, alteration of dynamic energy prices), and the retransmission of packets (duplicate packets), among others [3].

On the other hand, SMS is not used actually for data analytic applications inside of SMs. The data reported by the SMs could be used for diverse applications such as fraud energy prevent, power quality prediction, fault detection, among others.

## 2 Previous Work in the Area

The main security mechanisms implemented to solve this problem lie in the use of cryptographic techniques [4]. Other techniques used are digital signature schemes and public-key schemes (PKI), Preventing and Intrusive Detection Systems (IPS and IDS), among others [5]. Recently, blockchain techniques (Blockchains) have been used because it is the combination of multiple cybersecurity techniques [6].

On the other hand, there are works related to data analytics, machine learning, and artificial intelligence with data in SMS [7, 8], in specific fields such as energy thief [9]. The most related works are presented in [9-15].

## 3 Research Objectives

The general objective of this work is develop a new architecture for AMI that in addition to the basic services related to measurement and billing can provide additional services to improve energy efficiency and contribute to better respond to the demand of electricity, through the use of data analytics directly in the measuring devices that allow to measure and control the quality of the energy as well as the detection of faults in the electrical network considering for all these services the security and privacy of the data.

## 4 Methodology

For the development of the proposed work, it is necessary to carry out the following steps:

1. Study and evaluation of AMI infrastructure that allows knowing the best way to implement AMI.
2. Implementation of a functional AMI prototype for testing.
3. Implementation of a multi-tier blockchain architecture for AMI that guarantees security and confidence in electricity transactions.

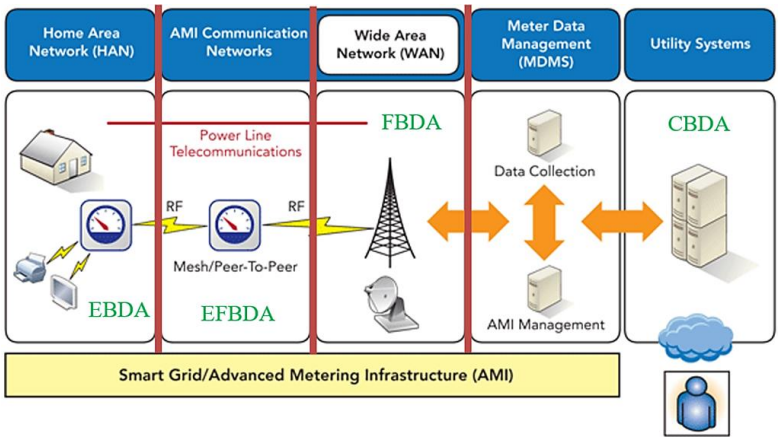



**Fig. 1.** AMI General Architecture with the inclusion of a Multi-Tier Blockchain and Data Analytic Platform.

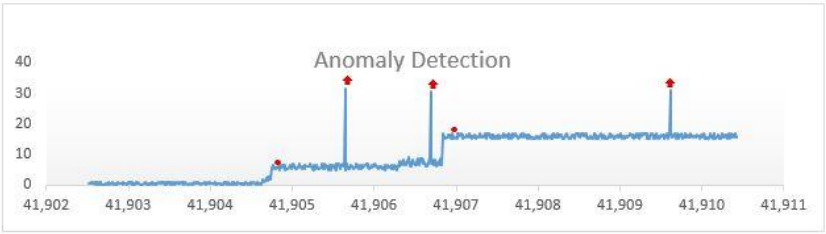



**Fig. 2.** Anomaly Detection consider as Energy Theft.

4. Strengthening of security and infrastructure in AMI.
5. Development of data analysis and machine learning techniques to improve energy efficiency and failure detection in electrical networks.

## 5 State of the Research

The authors are developing a new consensus algorithm for blockchain named Proof-of-Efficiency (PoEf) in their first version works with Moving Average (MA) and a basic Heuristic for data analytic based on Time Series Approaches like ARIMA. Right now, the authors are improving the algorithm to predict possible energy theft, check the quality of energy and determines the most efficient household. Each data analytic application has a different algorithm to process data. The analytic platform works with a reinforcement learning approach [16] with the idea of improving the results of estimations in each iteration.

Figure 1 shows the general AMI architecture extended with a multi-tier architecture of blockchain and data analytics platform. The readers can note that exist 4 main areas: CBDA (Cloud Blockchain and Data Analytics) represented with the utility's data center

servers, FBDA (Fog Blockchain and Data Analytics) represented with Data Concentrators (DC) in Substations, EFBDA (Edge-Fog Blockchain and Data Analytics) represented with DC in Neighbor Area Network (NAN) and SM, and the last tier EBDA (Edge Blockchain Data Analytics) represented with SM and Smart Appliances (SA). Figure 2 shows the main idea of detecting anomaly consumption/production inside SM. The techniques used are a hybrid approach between inferential statistics with machine learning classifiers.

## 6 Conclusions

There are some advances in the cybersecurity field programming a multi-tier blockchain in SMS. There is work in progress developing a new consensus algorithm using data analytics to rewarding energy transactions. Right now, we are working on improving a fog-edge-cloud computing architecture for secure data analytic application in SMS.

## References

1. Coelho, P., Gomes, M., Moreira, C.: Smart metering technology. Microgrids Design and Implementation, pp. 97–137 (2018)
2. Weranga, K.S.K., Kumarawasu, S., Chandima, D.P.: Smart metering design and applications. Springer (2014)
3. Knapp, E.D., Samani, R.: Applied cyber security and the smart grid: Implementing security controls into the modern power infrastructure. Syngress (2013)
4. Borges de Oliveira, F.: On privacy-preserving protocols for smart metering systems: security and privacy in smart grids. Science Direct (2017)
5. Song, H., Flick, G.A., Jeschke, S.: Security and privacy in cyber-physical systems: foundations, principles, and applications. John Wiley & Sons (2018)
6. Dorri, A., Kanhere, S.S., Jurdak, R., Gauravaram, P.: Blockchain for IoT security and privacy: The case study of a smart home. In: IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 618–623 (2017)
7. Ye, F., Qian, Y., Qingyang-Hu, R.: Smart grid communication infrastructures: Big data, cloud computing, and security. In: Communication, Networking and Broadcast Technologies; Components, Circuits, Devices and Systems; Computing and Processing; Photonics and Electrooptics; Power, Energy and Industry Applications; Geoscience, pp. 304, Wiley–IEEE Press (2017)
8. Al-Shaer, E., Rahman, M.A.: Security and resiliency analytics for smart grids: Static and dynamic approaches. Springer (2017)
9. Badrinath-Krishna, V., Lee, K., Weaver, G.A., Iyer, R.K., Sanders, W.H.: F-DETA: A framework for detecting electricity thef attacks in smart grids. In: 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (2016)
10. Aitzhan, N.Z., Svetinovic, D.: Security and privacy in decentralized energy trading through multi-signatures, blockchain and anonymous messaging streams. In: IEEE Transaction on Dependable and Secure Computing (2016)

11. Gao, J., Omono-Asamoah, K., Boateng-Sifah, E., Smahi, A., Xia, Q., Xia, H., Zhang, X., Dong, G.: Gridmonitoring: Secured sovereign blockchain based monitoring on smart grid. In: IEEE Access, 6, pp. 9917–9925 (2018)
12. Sompolinsky, Y., Wyborski, S., Zohar, A.: PHANTOM: A scalable BlockDAG protocol. School of Engineering and Computer Science (2018)
13. Cebe, M., Akkaya, K., Aksu, H., Uluagac, S.: Block4Forensic: An integrated lightweight blockchain framework for forensics applications of connected vehicles. In: IEEE Communications Magazine, 56(10), pp. 50–57 (2018)
14. Sharmar, P.K., Chen, M.Y., Hyuk-Park, J.: A software defined fog node based distributed blockchain cloud architecture for IoT. In: IEEE Access, 6, pp. 115–124 (2018)
15. Gu, J., Sun, B., Du, X., Wang, J., Zhuang, Y., Wang, Z.: Consortium blockchain-based malware detection in mobile devices. In: IEEE Access, 6, pp. 12118–12128 (2018)
16. Rayati, M., Sheikhi, A., Ranjbar, A.M.: Applying reinforcement learning method to optimize an energy Hub operation in the smart grid. In: IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5 (2015)

ISSN 1870-4069

# A Multilayered Model based on Blockchain that Fortifies the Integrity and Security of Public Information

Fernando Rebollar, Rocío Aldeco-Pérez, Rosa M. Valdovinos, Marco A. Ramos

Universidad Autónoma del Estado de Mexico,
Facultad de Ingeniería,
Mexico

{frebollarc,rvaldovinosr}@uaemex.mx,
marco.corchado@gmail.com,raldeco@unam.mx

**Abstract.** Digital services have increased proportionately with Internet access. Then, governments in different countries also have increased the digital services they offer to facilitate multiple procedures. As consequence, the use of reliable and transparent systems has become a need. A solution is the use of cryptographic techniques that has culminated in the emergence of blockchain, which allows to decentralize information and build trust and adding smart contracts functionality. However, scalability and speed of transactions are still challenging. Governments from different countries have already evaluated the use of blockchain in their systems, but adaptation strategies and optimization improvements are still required to be successful. This investigation presents a possible solution, adapted to use in government systems, dividing blockchain into multiple organized layers, allowing the use of smart contracts functionality and document storage.

**Keywords:** Multi-layered blockchain, blockchain, smart contracts, government service.

## 1 Motivation

As a result of an increased access to Internet, the offer of digital services have grown exponentially [8]. Examples are the governments of developed countries that have increased the number of digital services offered to their citizens and taxpayers [5]. Some of these digital services are created to allow citizens to make payments or to get information about how taxes are invested, spent or allocated. These digital services, even reduce cost and improve efficiency, they lack of transparency on the use of information and consequently decreasing the level of trust in such information that is presented from this government to their citizens.

On these type of digital systems, trust lies on a centralized entity that can be a government employee or an organization depending on the government.

Under this political centralization, digital services become vulnerable, since there is nothing left but to rely on the honesty of these centralized entities. This honesty cannot be enforced neither verified. For this reason, mechanisms that guarantee transparency are necessary so they can facilitate auditing in a decentralized way.

Blockchain has proven to be effective on providing information decentralization [4], while allowing ease of verifying this information transparency, integrity and immutability [10]. Blockchain also makes possible to create the so-called smart contracts [2], which are digital assets containing secure digital agreements that self-execute once the event(s) for which they were designed happen.

The advantages of blockchain have already been recognized by governments of different countries, some of these have already launched projects to incorporate blockchain into digital services, however, they have also found difficult to make it function in the best way [6].

Unfortunately, when blockchain is implemented, a number of challenges that limit its usage are present. For instance, when a considerable number of transactions are created the speed under those transactions are performed and later verified decreases. If smart contracts are added to the picture, then the storage of them became a problem. These challenges are extensively discussed on [9].

A possible solution to the above mentioned challenges is the use of a "multi-layer blockchain". A multi-layer blockchain, as its name indicates, create layers for dividing the different types of information that are stored on such a blockchain. The goal is creating separate blocks that will be accessed only when is required. In that way, performance is improved and information storage can be better manage. [1]

## 2 Related Work

Cheng and Zhang (2017) in [7] present a proposal for a blockchain-based network model to improve the implementation of IoT devices, maintaining network security by combining cloud storage protocols and proposing two types of layers: external layers and high-level layers. The external layers operate centrally as cloud servers commonly do, while the high-level layers connect to external layers and behave like connected IoT devices. Unlike the outer layer, at the high-level layer, the network is decentralized.

Badr et al. (2018) in [1] describe a multilayer model for clinical patient data, resuming the model proposed by Cheng and Zhang but presenting a configuration in which layers contain 3 main levels. The first level for the patient's devices and sensors, the second level corresponding to hospitals, laboratories, medical bodies, etc. The third tier for centralized cloud storage.

Chang et al. (2018) in [3] propose a two-layer blockchain-based model to preserve clinical patient data while ensuring their privacy. Added

deep learning algorithms are used to guarantee data distribution without sacrificing data privacy.

Zhou et al. (2018) in [11] proposes a multilayer architecture from which a cryptocurrency called MOAC is implemented. The proposal maintains security, decentralization and speed of transactions using smart contract functionality through a two-layer splitting. The base layer is used for storing files and the top layer to execute smart contracts.

The majority of the discussed proposals present a two-layer model in which information is separated based on its functionality. However, two layers are not enough because the different types of information contained in the blocks are not separated, which still generates a loss of speed in transactions, this problem gets worse when features like smart contracts are added.

## 3 Hypothesis

Whit the development of a new multiple layer blockchain were each layer will use a different consensus algorithm, will be possible to guarantee the integrity and security of information while increasing its reliability.

## 4 Methodology

The proposed methodology considers the following steps:

1. *To study and to analyze the operation of blockchain.* To study and analyze the main cryptographic algorithms used by blockchain, consensus algorithms and mechanisms used to carry out transactions in distributed systems.
2. *To generate the multi-layer model based on blockchain* Once the advantages, disadvantages, compatibilities and incompatibilities of the blockchain types and their consensus algorithms have been identified, a proposal will be built.
3. *To test the previously created proposal and to document the results* Perform a proposal validation through modeling and formalization to validate or refute the hypothesis. The modeling will be carried out using a probabilistic qualitative approach, to infer its behavior.

## 5 State of the research and Preliminary results

After two years of work, the proposal design is finalized. This proposal presents 4 layers, layer 1: Index-Keys, layer 2: Transactions, layer 3: SmartContracts and layer 4: Files.

By using 4 layers, information can be classified into different types based on the use of each type of information and later be included into the corresponding layer. In this way it is expected to improve the efficiency as transactions will occur in different layers based on functionality.

The next step is to validate the proposal that corresponds to step 3 of the methodology. We believe that given the current state of blockchain a formalization of each layer will help us to verify their behavior. Previous to a formalization is necessary to have a complete model of our proposal. This model will be validated by simulations.

## 6 Conclusions

Using blockchain-based digital systems can allow transparency and verification of information from government digital systems. It is still necessary to improve the operation of blockchain.

The proposal still needs to be validated and tested so there are no measurable results. However, it is expected that by dividing and organizing the different types of data, a more efficient process that maintain a high rate of transactions will make the use of the blockchain viable in various digital services.

By improving transaction speed and leveraging the functionality of smart contracts, it is possible to streamline, make transparent and automatic various services offered by governments, such as make payments, collections, streamline procedures, etc.

As future work, automatic checks could be added where government digital services are open source. In this way, anyone can download and compare code versions with their cryptographic hash validating that the code that has being released is the same as the one running on the server.

## References

1. Badr, S., Gomaa, I., Abd-Elrahman, E.: Multi-tier blockchain framework for iot-ehrs systems. Procedia Computer Science, vol. 141, pp. 159–166 (2018)
2. Buterin, V.: A next-generation smart contract and decentralized application platform. Web: ethereum.org/en/whitepaper/ Accessed 18/09/2020, (2014)
3. Chang, E. Y., Liao, S.-W., Liu, C.-T., Lin, W.-C., Liao, P.-W., Fu, W.-K., Mei, C.-H., Chang, E. J.: Deeplinq: Distributed multi-layer ledgers for privacy-preserving data sharing. In: 2018 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR). pp. 173–178. IEEE (2018)
4. Crosby, M., Pattanayak, P., Verma, S., Kalyanaraman, V., et al.: Blockchain technology: Beyond bitcoin. Applied Innovation, vol. 2, no. 6-10, pp. 71 (2016)
5. Gabison, G.: Policy considerations for the blockchain technology public and private applications. SMU Sci. & Tech. L. Rev., vol. 19, pp. 327–336 (2016)
6. Jun, M.: Blockchain government-a next form of infrastructure for the twenty-first century. Journal of Open Innovation: Technology, Market, and Complexity, vol. 4, no. 1, pp. 7 (2018)
7. Li, C., Zhang, L.-J.: A blockchain based new secure multi-layer network model for internet of things. In: 2017 IEEE International Congress on Internet of Things (ICIOT). pp. 33–41. IEEE (2017)
8. Lipschultz, J.: Free expression in the age of the Internet: Social and legal boundaries. Routledge (2018)

9. Moin, S., Karim, A., Safdar, Z., Safdar, K., Ahmed, E., Imran, M.: Securing iots in distributed blockchain: Analysis, requirements and open issues. Future Generation Computer Systems, vol. 100, pp. 325–343 (2019)
10. Wüst, K., Gervais, A.: Do you need a blockchain? In: 2018 Crypto Valley Conference on Blockchain Technology (CVCBT). pp. 45–54. IEEE (2018)
11. Yang, X., DavidChen, X., Zhou, S., Wang, R.: The moac platform: Advancing performance with layered multi-blockchain architecture for enhanced smart contracting. Web: moac.io/uploads/MOAC_White_Paper.pdf Accessed 18/09/2020, (2018)

ISSN 1870-4069

# COVID-19 on the Time, Mexico and the World

Juan J. Martínez, Ildar Batyrshin, Alexander Gelbukh

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

{juanjmtzs,batyr1}@gmail.com,
gelbukh@gelbukh.com

**Abstract:** This paper describes a brief analysis on the COVID-19 over the time Globally and specifically centered in Mexico. In this work, it was created a website and using Google Data Studio for analytical dashboards it is tracked the daily data dynamics of the pandemic, which is collected and represented graphically. For all data collecting it was developed various web scraping scripts mainly based on bash scripting and python which extract data from specific web sites and once the initial inputs are obtained, the transforming process is started making aggregations, key performance indicators, correlations and mappings giving the facility of using that transformed data for future works. Furthermore, it is also used an specific model for studying the changing aspects of the epidemic and this is presented to analyze the rates, model is discussed and it is shown how it can be used to track impactful decisions to reduce the positive cases identified. The data has been collected and treated for study from different sources [1, 2, 3, 4]. Additionally, all the results and final data after transformations is being published on a daily basis in the following sites [5, 6, 7, 8].

**Keywords:** COVID-19, coronavirus, SARS-CoV-2, COVID19 mx, contagion rate.

## 1 Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe respiratory syndrome coronavirus (SARS-CoV-2). It has being rapidly spread all over the world in only six months. On 31 December 2019 Wuhan Municipal Health Commission, China, reported a cluster of cases of pneumonia in Wuhan, Hubei Province. On January the 4th 2020, WHO (World Heath Organization) reported on social media that there was a cluster of pneumonia cases – with no deaths – in Wuhan. On the 13th of January 2020 officials confirmed a case of COVID-19 in Thailand, the first recorded case outside of China. On 11 March 2020 deeply concerned both by the alarming levels of spread and severity, and by the alarming levels of inaction, WHO made the assessment that COVID-19 could be characterized as a pandemic. Nowadays more than 11.62 million people has being infected all over the world.

The virus is primarily spread between people during close contact, most often via small droplets produced by coughing, sneezing, and talking. The droplets usually fall

to the ground or onto surfaces rather than travelling through air over long distances, although in some cases they may remain airborne for tens of minutes. Less commonly, people may become infected by touching a contaminated surface and then touching their face. It is also known the COVID-19 is most contagious during the first three days after the onset of symptoms, although spread is possible before symptoms appear, and from people who do not show any kind of them.

On this work, many data science techniques are applied and represented visually in order to disclose data analytics for the people to stay informed and compare virus dynamics between countries.

## 2 How Infection Rate is Analysed

To analyse daily behaviour of the pandemic the following model is used where exist an initial formula, which is the Delta of the Infected People on a day $\Delta I$ that is directly related to three factors:

- **Exposure:** $(E)$ describes the contact with the virus not saying at home and being in continuous interactions with people, in other words, not respecting social distancing staying at least 2m apart from others.
- **Contagion probability:** $(P)$ describes the compliance of instructions to avoid contagion like washing your hands, using anti-bacterial hand gel, using mask of at least three layers and N95 in case of contact with infected people, avoid touching your face and eyes, etc.
- **Infected people:** $(I_n)$ the total confirmed people with the infection on a $n$ day.

The initial formula is as follows:

$$\Delta I = E * P * I_n \ .$$

Now, $\Delta I$ is simply calculated by the difference of those infected on the day $I_{n+1}$ minus the day $I_n$, in other words today's infections subtracting yesterday's infections:

$$\Delta I = I_{n+1} - I_n.$$

Then substitute the new equation in the original one:

$$I_{n+1} - I_n = E * P * I_n.$$

Once with the substitution, process continues isolating the variable $I_{n+1}$ and factoring the equation:

$$I_{n+1} = E * P * I_n + I_n,$$

$$I_{n+1} = I_n \ (E * P + 1).$$

Now $(E * P + 1)$ can be named as **Contagion rate** $(Cr)$ which is the actual problem factor on the propagation of the virus. Substituting and isolating the new variable the resulted equation is:

$$I_{n+1} = I_n \ Cr,$$

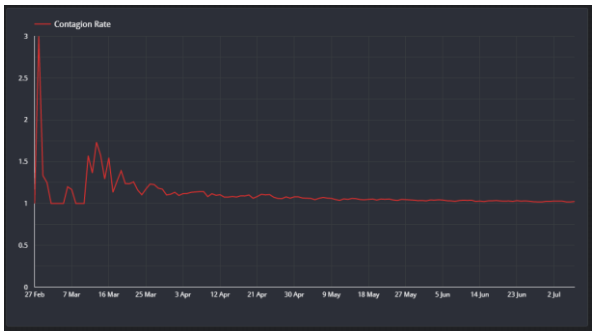



**Fig. 1.** Contagion rate in Mexico from 27 of February to 7th July 2020.

$$Cr = \frac{I_{n+1}}{I_n}.$$

For knowing that the virus propagation is stopped, $Cr$ must be one which means that there are no more new infections every day. Applying the proposed model, for the official data of Mexico, the representation looks as shown in Fig. 1.

Mexico and Latin America have been widely affected, on this regard an important methodology taken in Mexico was the National Period of Healthy Distance (In Spanish, Jornada Nacional de Sana Distancia) which started on March 23rd 2020. Its main rules consisted in [9]:

- **Basic Prevention:** Frequent hand washing; when sneezing, cover nose and mouth with the internal angle of the elbow; avoid physical greeting; and do not leave home if there are symptoms comparable to the coronavirus.
- **Temporary Suspension of Non-Essential Activities:** The Ministry of Health recommended to Mexican society to suspend the activities considered "not essential", that is, those that do not affect the "substantive activity" of companies, organizations and the government itself as of March 23. Also, those that gather people or that imply that there are constant displacements in public transport. Among the instructions given were:

  a. Avoid conglomerations of more than 100 people.
  b. Maximum limitation of public sector activity.
  c. Interruption of school activity at all educational levels.
  d. Closure of cultural sites such as museums, cultural centres and archaeological zones.
  e. Limitation of recreational sites such as activity in cinemas, theatres, restaurants and bars.

- **Reprogramming of Events of Massive Concentration:** Such as concerts or tourist or employer's fairs.
- **Provide Protection and Care for Elderly People:** As they were considered the most vulnerable sector against possible COVID-19 infection.

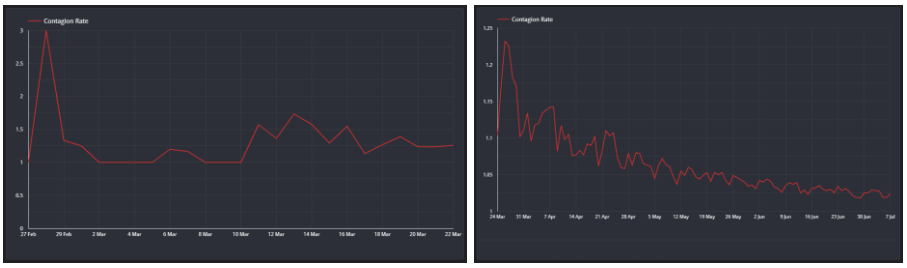



**Fig. 2.** Contagion rate in Mexico from 27 of February to 23 of March 2020. Contagion rate in Mexico from 23 of March to 7th July 2020.

In order to analyse the data in a better way and exemplify our model, a segregation was done over the time series, separating it in two sections for now, before and after 23 of March 2020.

Comparing both periods it is demonstrated a positive result with the National Period Of Healthy Distance strategy implemented by Mexican Government using the proposed model, as in Fig. 2 first chart $Cr$ is observed a SIDEWAYS TREND while in the second chart is seen a DOWNWARD TREND reducing the infection propagation in the country.

With this support, is suggested to use this model actively as a key indicator of the external factors and its impact on the pandemic dynamics to know if a decision is having good results or wrong ones.

## 3 Current Status in Mexico and the World

It has passed over 168 days on the world since first detected case and 132 days in Mexico. In each country different decisions have been made by their governments in order to reduce the infection propagation. Here some data statistics are presented.

### 3.1 World

As for July the 7th the related numbers for the pandemic globally are:

- 11,829,602-Confirmed Cases,
- 544,163-Deaths,
- 6,447,656-Recovered People,
- 1.018-Contagion Rate,
- 1.011-Deaths rate,
- 1.023-Recovered Rate.

Reviewing the progress on the time for confirmed cases and daily new confirmed cases in the world is seen in Fig. 3 that daily new cases is in a SIDEWAYS TREND and the plateau was reached. Moreover, in the same picture is shown top countries by confirmed cases being USA the most critical country with a difference of around 1.4 million from the second place that is being occupied by Brazil.

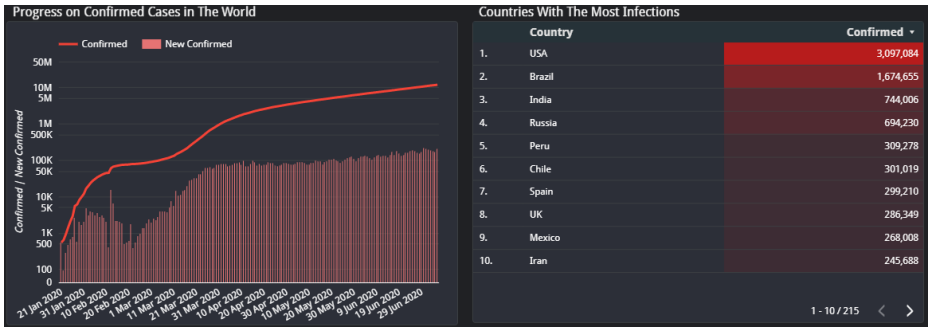



**Fig. 3.** Time series chart for confirmed cases and daily confirmed cases in the world. Top Countries by total confirmed cases in the world [6].

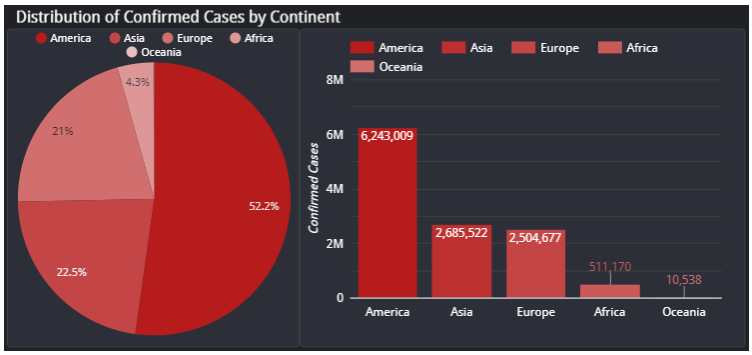



**Fig. 4.** Confirmed cases by continent in the world [6].

In Fig. 4 is represented graphically the distribution of confirmed cases and in spite of the epidemy started in Asia (Wuhan-China), nowadays America's region has overpassed Asia by 232% having the 52.2% of the total confirmed cases all over the world.

Talking about deaths is identified a DOWNWARD TREND for daily new deaths globally in Fig. 5 reaching the resistance zone by first half of April. Top countries by deaths are represented also in the picture with USA in the first place and a difference of around 67 thousand (200%) from the second place that is being occupied by Brazil. Reviewing Fig. 6, is seen an UPWARD TREND for recovered people from COVID-19 and also the table for top countries by recuperated people is represented.

## 3.2 Mexico

On the day 132 (7th July 2020), numbers representing the pandemic in Mexico are as follows:

- 268,008-Confirmed Cases,
- 273,289-Projection (Moving Average Model, 3-day window),

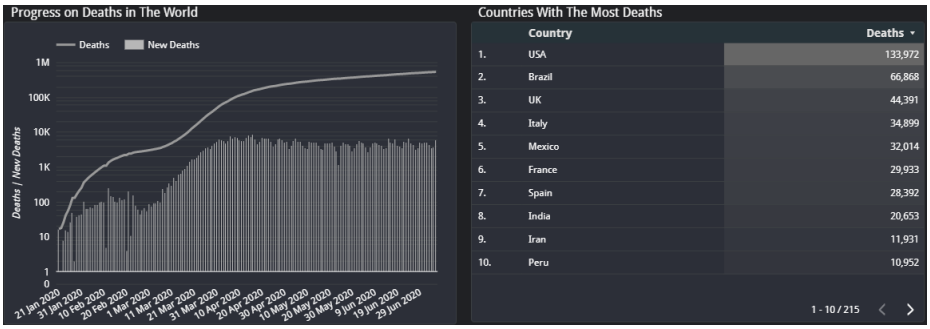



**Fig. 5.** Time series chart for deaths and daily new deaths in the world. Top Countries by total deaths in the world [6].

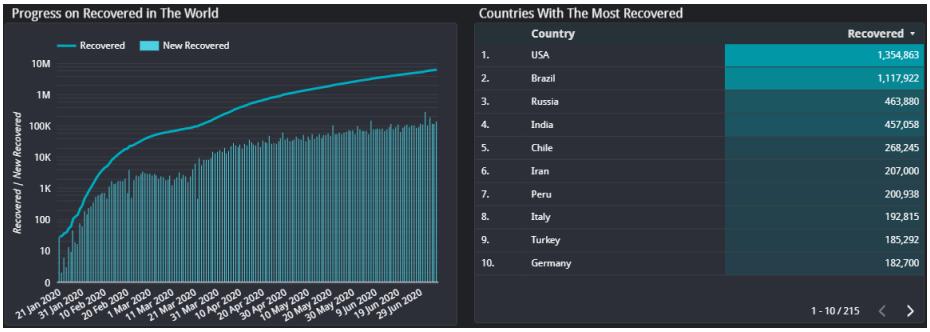



**Fig. 6.** Time series chart for recovered people and daily new recovered cases in the world. Top Countries by total recovered people in the world [6].

2,331,670-Sentinel Model,

77,703-Suspected,

322,826-Negative Cases,

32,014-Deaths,

209,437-Recovered,

1.024-Contagion Rate,

1.029-Deaths Rate.

Sentinel Model was the Mexican government's strategy to monitor the behaviour of the pandemic for not identified cases in the country. It is an epidemiological surveillance program and is used in several countries to monitor diseases such as seasonal influenza. Nowadays is being taken last reported factor by the Mexican Government of 8.7x and continued the tracking and calculation on a daily basis.

About daily forecasting, it is used a Moving Average method over 3-day window, in Fig. 7 is shown the tracking over predictions vs real confirmed cases with a mean error in the last month of +- 0.003%.

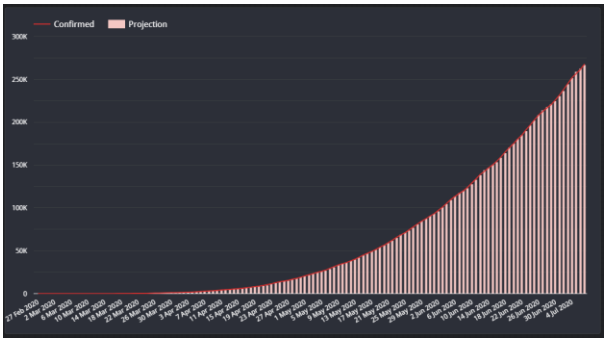



**Fig. 7.** Confirmed cases compared with daily predictions in Mexico.

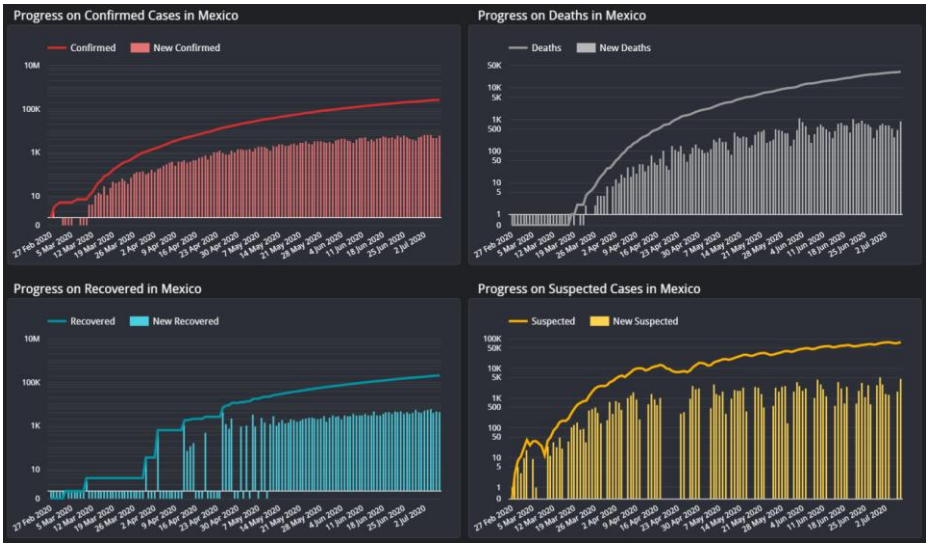



**Fig. 8.** Time series charts for confirmed cases and daily confirmed cases in Mexico. Time series chart for deaths and daily new deaths in Mexico. Time series chart for suspected cases and daily new suspected cases in Mexico. Time series chart for recovered cases and daily new recovered cases in Mexico [6].

In Fig. 8 are represented firstly the progress of the pandemic over the time for confirmed cases and daily new confirmed cases since 27$^{th}$ of February when the first case was reported can be seen in where it is identified that plateau and resistance zone have been reached for daily new cases. Secondly, the progress of the deaths over the time in Mexico and daily new deaths. Thirdly, another time series that is being tracked is the progress of suspected cases and daily new suspected people with COVID-19. Finally, the progress of the recovered cases over the time and daily new recovered people that resulted positive with COVID-19 is shown in the same figure.

As can be seen the recovered time series chart presents an anormal behaviour on the first months as the Mexican Government changed the definition of recovered cases, naming active cases just the ones that resulted positive in the last 14 days. However,

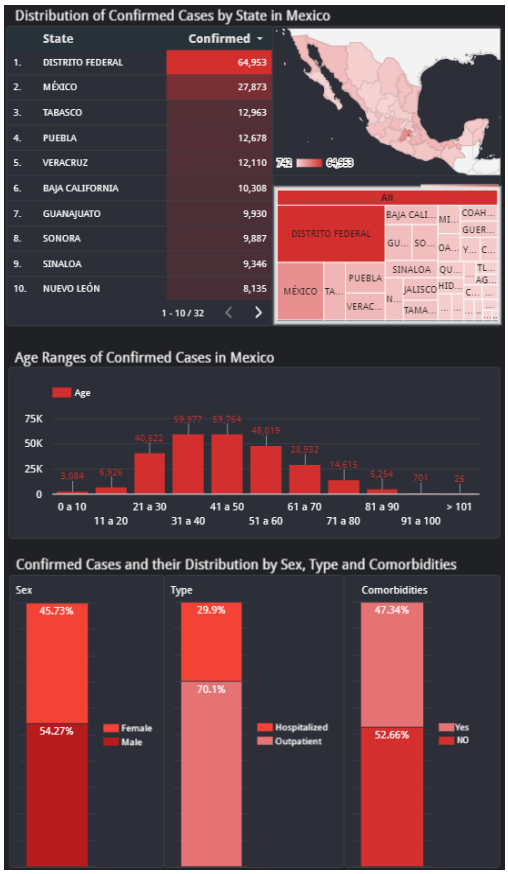



**Fig. 9.** Confirmed cases distribution by states and top states with confirmed infected people in Mexico. Confirmed cases distribution by ages in Mexico. Confirmed cases distribution by gender, type of treatment and with or without comorbidities [6].

apart from that there is not certain information on the number of recovered people in Mexico. Taking those official numbers means that 78.14% of the infected people in Mexico is already recovered which turns Mexico into one of the countries with most recovered ratio in the world, in contrast USA that is the most critical country reports only 43.57% of recovered people. Based on those numbers, it is difficult to trust on the Mexican Government numbers.

### 3.2.1 Confirmed Cases Segmentation

The actual concentration of the pandemic per State in Mexico is represented in Fig. 9. As it is seen, Mexico City is the state that on the time has had the majority of confirmed cases with 24.23% followed by Mexico State with 10.40% of the total confirmed cases in the country. Also represented in the picture, age ranges for confirmed cases has the major concentration of infected people between 31-50 years old. Also, the distribution

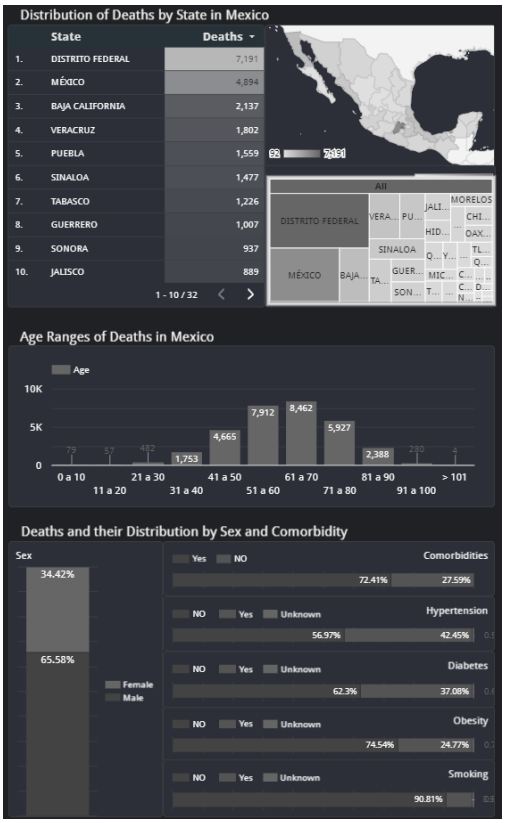



**Fig. 10.** Deaths distribution by states and top states with deaths in Mexico. Deaths distribution by ages in Mexico. Deaths distribution by gender and top comorbidities in Mexico [6].

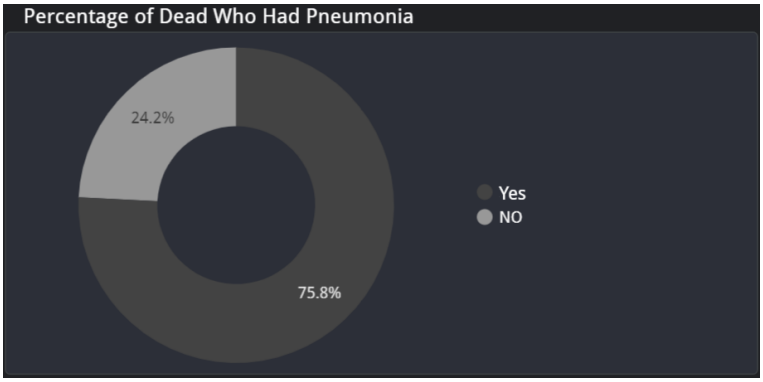



**Fig. 11.** Deaths distribution by pneumonia presence in Mexico [6].

of the infected people divided by gender, type of treatment (Hospitalized/Outpatient) and percentage of people having any comorbidity is illustrated in the same figure.

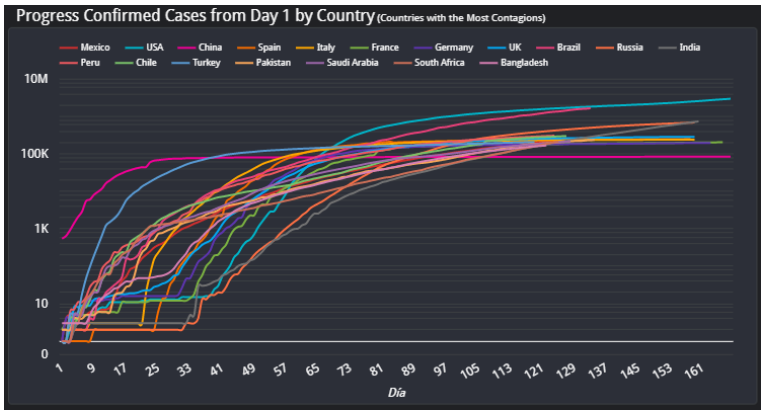



**Fig. 12.** Confirmed cases progression in Mexico and top countries with most confirmed cases taking day one in each country when the first reported case [6].

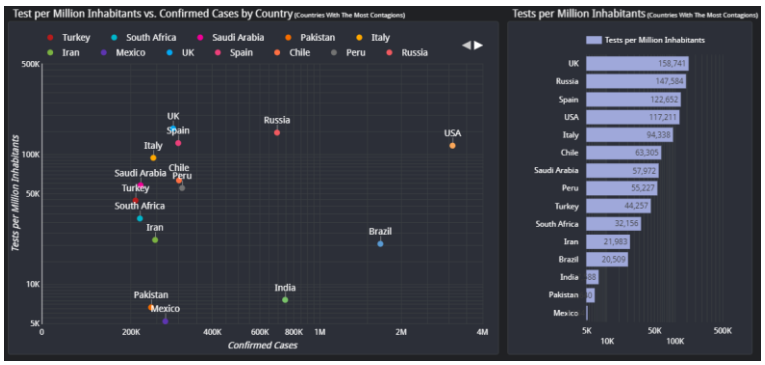



**Fig. 13.** Comparative for tests per million inhabitants vs. confirmed cases in Mexico and countries with more than 100,000 cases and tests/million. Comparative for tests per million inhabitants in Mexico and countries with more than 100,000 Cases and tests/million [6].

### 3.2.2 Death Cases Segmentation

Concentration of the deaths due to the pandemic per state in Mexico is represented in Fig. 10, being Mexico City the State that on the time has had most deaths followed by Mexico State. The age distribution for the death people due to the infection has the concentration between 51-70 years old which represents 51.14% of the total deaths in the country.

In Mexico, the majority of the death people has been males covering the 65.58% of the total and the segmentation by comorbidity can be seen also in the Fig. 10 being hypertension the one that most of deadly cases presented apart from COVID-19. However, 27.59% of the cases did not have any other chronic disease during their infection.

Another interesting statistic is that 75.8% of the deaths presented pneumonia as illustrated in Fig. 11.

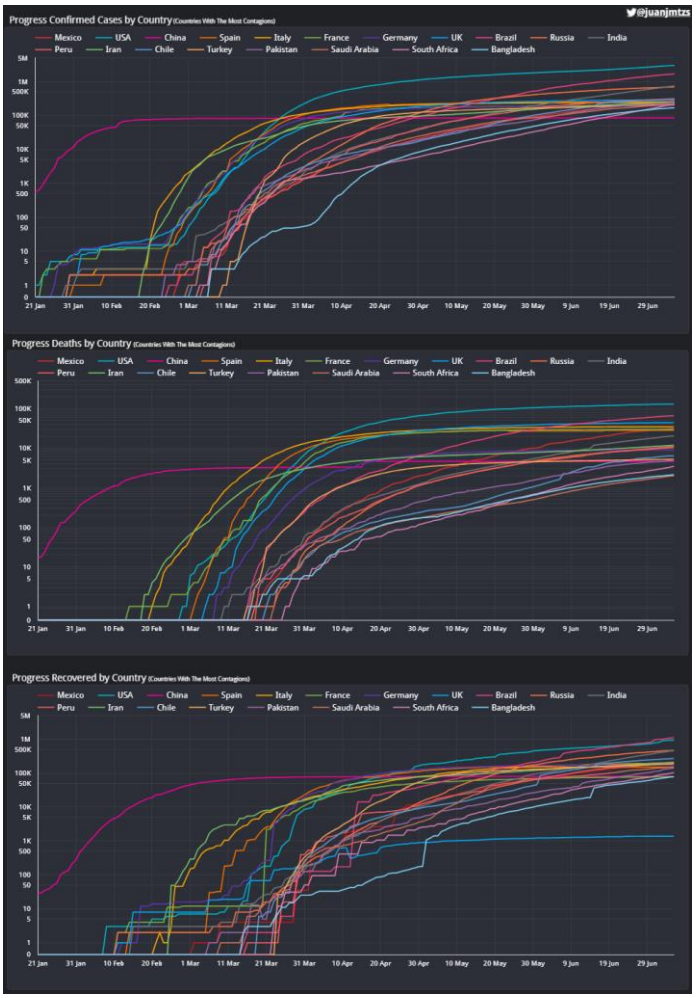



**Fig. 14.** Confirmed cases progression in Mexico and top countries with most confirmed cases. Deaths progression in Mexico and top countries with most confirmed cases. Recovered people progression in Mexico and top countries with most confirmed cases [6].

### 3.3 Mexico in Contrast with the World

As part of the developed research it is analysed confirmed cases progression in Mexico and top countries with most confirmed cases taking day one in each country as the first reported case Fig. 12. What is seen is that despite Mexico did not have same velocity for increasing the total cases, other countries reached their plateau in less days.

Globally, one of the most discussed topics has been number of test taken per millions of inhabitants and the conclusion obtained based on the illustration Fig. 13 is that there is not a strong correlation between number of test/million and confirmed cases, total confirmed cases may increase if more tests are made, however, it must be considered

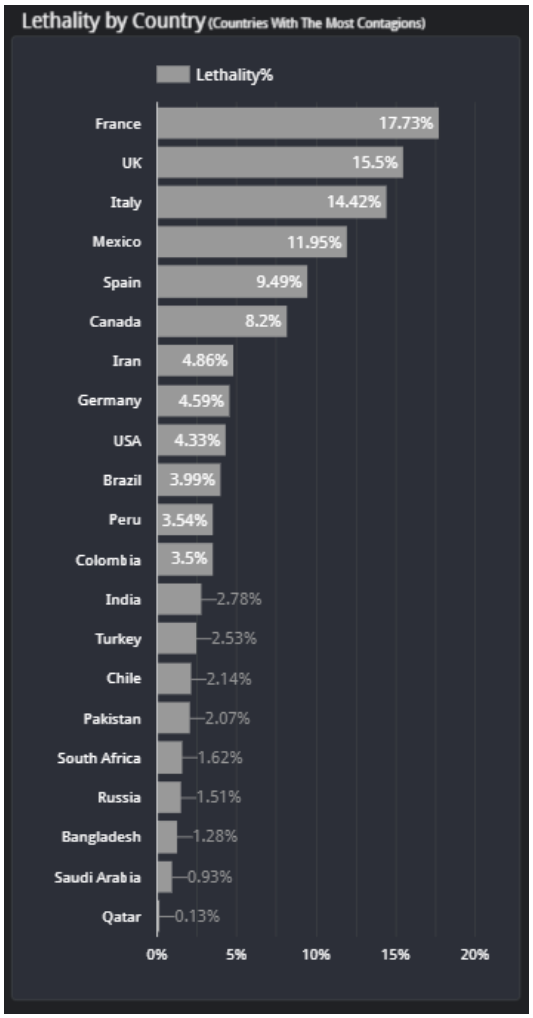



**Fig. 15.** Lethality in Mexico and top countries with most confirmed cases [6].

the possibility of doing it based on the total population of each country, for example Monaco is the territory with more tests per million but with a population of only 39,227. On this regard, Mexico has the last place on tests per million of inhabitants compared with countries with more than 100,000 cases and more tests/million as represented in Fig. 13.

Another visualisation that is presented was making comparations between Mexico and top countries with most COVID-19 contagions on confirmed, deaths and recovered cases as shown in Fig. 14.

As the final analysis for this work, in Fig. 15 it is identified Mexico as one of the countries with more lethality compared with the countries that have most confirmed cases in the world occupying the 6$^{th}$ place with 11.95%.

## 4 Conclusions and Future Work

Here is presented a basic mathematic model for tracking contagion rate, a simple forecasting model for daily rebalancing and most work was based on presenting visualisations on the pandemic over the time and current state in Mexico and the World. In most of the countries plateau has been reached and downward trends are being identified for confirmed cases and deaths. However, a regrowth may happen in coming days once activities are restarted.

For future research is planned to work over a longer time forecasting and countries classification depending on a multi-factor clustering, with the advanced work made here, the research path will continue taking the transformed data and improving the dashboards.

Moreover, specific work is being done to compare contagion, deaths and recovered curves on a Moving Average of 14 days to avoid peaks on data due to tracking methods in each country, on this path it will be easier to identify real trends and know over real data when a country has overpassed the pandemic or still is going upwards. Additionally, is intended to have countries classifications on different aspects and look for answers on why some territories made it better dealing with the pandemic or maybe there is hidden data in other ones explaining the reason of high lethality rates.

## 5 Conclusions

Some interesting related works that have been developed by different organizations like CONACYT [13], Scriby [12], Youyang Gu [11] and Institute for Health Metrics [10] and Evaluation, have similar work over data statistics on the pandemic and also different forecasting techniques using machine learning. These are carefully tracking data related to the spread of COVID-19 in the world encouraged by what is being seen in some areas and concerned about what is seen in others. As world move forward, is needed to continue making decisions based on the science, data and facts related to the specific conditions in our communities. Those researches are committed to providing accurate, reliable reports to the public and the information presented is updated daily and is dependent on reporting by numerous agencies across the world.

## References

1. Gobierno de México: Datos Abiertos-Dirección General de Epidemiología (2020)
2. Gobierno de México: Coronavirus (COVID-19)-Comunicado Técnico Diario (2020)
3. Johns Hopkins University: COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) (2020)
4. Worldometer: Covid-19 coronavirus pandemic (2020)
5. Martínez, J.J.: @JuanJMtzS. Retrieved from Twitter (2020)
6. Martínez, J.J.: BI-COVID19. Bi-covid19.hiketech (2020)
7. Martínez, J.J.: BI-COVID19. GitHub (2020)
8. Martínez, J.J.: Juan J. Martínez LinkedIn (2020)

9. Gobierno de México: Jornada Nacional de Sana Distancia (2020)
10. Institute for Health Metrics and Evaluation: COVID-19 Projections. Healthdata (2020)
11. Gu, Y.: COVID-19 Projections Using Machine Learning. Covid19-projections (2020)
12. Scriby Inc.: COVID-19 Map. Coronavirus app (2020)
13. CONACYT: Covid-19 México. Gobierno de México (2020)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación