Cognitive Computation manuscript No.
(will be inserted by the editor)

An Exploration and Forecast of COVID-19 in Mexico with Machine Learning

Daniela A. Gomez-Cravioto · Ramon E. Diaz-Ramos* · Francisco J. Cantu-Ortiz · Hector G. Ceballos

Received: date / Accepted: date

Abstract Background: To understand and approach the COVID-19 spread, Machine Learning offers fundamental tools. This study presents the use of machine learning techniques for the projection of COVID-19 infections and deaths in Mexico. The research has three main objectives: first, to identify which function adjusts the best to the infected population growth in Mexico; second, to determine the feature importance of climate and mobility; third, to compare the results of a traditional time series statistical model with a modern approach in machine learning. The motivation for this work is to support health care providers in their preparation and planning. **Methods:** The methods used are linear, polynomial, and generalized logistic regression models to evaluate the growth of the COVID-19 incidents in the country. Additionally, machine learning and time-series techniques are used to identify feature importance and perform forecasting for daily cases and fatalities. The study uses the publicly available data sets from the John Hopkins University of

* Corresponding author.

D. Gomez-Cravioto
School of Engineering and Sciences, Tecnologico de Monterrey, 64849 Monterrey, N.L.
Tel.: +52 81 8358 2000
E-mail: a01181520@itesm.mx

R. Diaz-Ramos
School of Engineering and Sciences, Tecnologico de Monterrey, 64849 Monterrey, N.L.
Tel.: +52 81 8358 2000
E-mail: a01133921@itesm.mx

F. Cantu-Ortiz
School of Engineering and Sciences, Tecnologico de Monterrey, 64849 Monterrey, N.L.
Tel.: +52 81 8358 2000
E-mail: fcantu@tec.mx

H. Ceballos
School of Engineering and Sciences, Tecnologico de Monterrey, 64849 Monterrey, N.L.
Tel.: +52 81 8358 2000
E-mail: ceballos@tec.mx

Medicine in conjunction with mobility rates obtained from Google's Mobility Reports and climate variables acquired from Weather Online. **Results:** The results suggest that the logistic growth model fits best the behavior of the pandemic in Mexico, that there is a significant correlation of climate and mobility variables with the disease numbers, and that LSTM is a more suitable approach for the prediction of daily cases. **Conclusion:** We hope that this study can make some contributions to the world's response to this epidemic as well as give some references for future research.

Keywords Covid19 · Data Science · Time-Series Forecasting · Recurrent Neural Networks.

1 Introduction

As referenced by the World Health Organization, the first case of COVID-19 (also known as 2019 Novel Coronavirus) was confirmed in Wuhan, China on December 31st, 2019 [1]. Even though the disease is now successfully contained in China, it has spread all over the world. On May 21th there had been over 5,102,424 confirmed cases which resulted in more than 332,924 fatalities around the world [2]. The pandemic is severe and it continues to affect billions of people.

In this study, we compare three curve fitting models: a Linear, Polynomial and Generalized Logistic Model (GLM) and two multivariate time-series models: a Long-Short Term Memory (LSTM) neural network and a traditional time-series, Vector Autoregression (VAR) model to predict the number of COVID-19 daily cases and fatalities in Mexico.

The motivation of this study is to contribute to the knowledge necessary to fight the disease and characterize its course in Mexico, with the attempt to display more preparedness and promote more logical actions by the policymakers and the population in general.

The generalized logistic model has been successfully applied in other studies to describe previous epidemics [3] and the LSTM which uses a type of recurrent neural network (RNN) has previously been used in other studies to predict infections over time [4]. There are risk factors such as climate features and the adherence to social distancing that has previously been hypothesized to affect the number of daily cases. However, we did not find a previous study analyzing the significance of these factors with the use of machine learning techniques in time-series forecasts.

For the data exploration and model training, we used the dataset obtained from the Resource Center at John Hopkins University of Medicine GitHub repository [5] and supplemented this with information with climate information obtained from weather online API [6], and social mobility rate obtained from Google's COVID-19 Community Mobility Reports [7].

The limitations of this study are the data collection bias; the number of reported cases is in function of the number of tests that are applied, and the willingness of the government to report the numbers. Thus the potential

censoring in the data can affect the predictions. Another limitation is that the number of daily cases is reported based on the number of available tests when there are zero tests, there are zero known cases in the data but this does not necessarily reflect the reality; in other cases, there are observations with a huge spike in daily cases which can reflect that there were many tests available. One last identified limitation is the lack of sufficient data which may deter the predictions; the pandemic is currently in progress and therefore the model should be constantly updated with larger amounts of observations.

The remainder of this paper is structured as follows: section I describes the related work, section II describes the methods and dataset used in this research. In section III, we present a data exploration and preparation for modeling. Section IV presents the results of the models. Finally, section V presents the conclusions and future works proposed.

1.1 Related Work

With the same purpose of forecasting COVID-19 confirmed cases, we were able to identify the following related work, which mainly consists of studies using multivariate time-series regressions and curve-fitting models.

Related work includes the work of Chae, Kwon and Lee [4] who compared deep neural network (DNN) and long-short term memory (LSTM) with the ordinary least squares methods (OLS) and the autoregressive integrated moving average (ARIMA) to predict three infectious diseases (chickenpox, scarlet fever, and malaria). The result of this study showed that both deep learning models had a better performance than the traditional methods of OLS and ARIMA, with an average of 20% improvement on the RMSE.

Another related work is the work of Liu et al. [8], who analyzed the impact of meteorological factors on COVID-19 in China's provinces. The results obtained from this study indicated that the transmission may be affected by factors such as low temperature, low humidity, and mild diurnal temperature range.

Finally, this work is similar to the work done by Tomar and Gupta [9], where curve-fitting methods and LSTM were used to predict the number of COVID-19 cases in India and measured how the preventive measures like social isolation and lockdown affected the spread of COVID-19. The results indicated that preventive measures (social isolation and lockdown) have worked well in containing this contagious virus in India. It also showed a graph showing how the forecasted numbers with the logistic curve fitting closely resembled the official data.

2 Methods

In this research, we compare different techniques to forecast COVID-19 incidences and obtain insights into the COVID-19 outbreak. The exploration

and visualization of the data, as well as the machine learning modeling, were performed using Python programming and ran in the open-source Jupyter Notebook platform.

2.0.1 Datasets

The main dataset used for this analysis comes from the Resource Center at the John Hopkins University of Medicine. The data collected is open source, and available through a GitHub Repository [5], which is updated daily at 9 am EST.

The dataset contains information for the accumulated confirmed cases and fatalities in 173 countries. The features that are available in this data set are the following:

- Country: provided for 173 countries,
- Province: only for Australia, Canada, China, Denmark, France, Netherlands, United Kingdom, United States,
- Date: days since January 22$^{\text{nd}}$ to March 31$^{\text{st}}$ (70 days),
- Confirmed Cases: total number of confirmed Covid-19 cases, and
- Fatalities: total number of deaths.

The John Hopkins data was complemented with additional covariates: weather variables and social mobility rates. The climate information was obtained from weather online API [6] and the variables include max temperature, min temperature, UV index, humidity, precipitation, pressure, and wind speed. On the other hand, the social mobility rate was obtained from Google's COVID-19 Community Mobility Reports [7] and it includes the following variables which represent a percent change from the baseline: retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, and residential. These rates show how visits and stays differ from the baseline, which is the median value for the corresponding day of the week.

2.0.2 Uni-variate Growth Curve Models

The growth curve models, also called curve fitting models, are multilevel models mainly used to describe how a continuous outcome changes over time, focused mainly on the between-individual variations [10]. Different types of models to fit the curve include linear, polynomial of various degrees, logarithmic curve fit, and non-linear curve fit [11].

In this paper, we used three different growth models, linear regression, polynomial regression, and generalized logistic regression, and fit them to the data of the confirmed accumulated cases and fatalities in Mexico. This was done to identify the mathematical function that provided the best fit to the line or curves in the dataset. The linear regression approximates a straight line, while polynomial and generalized logistic regression are non-linear regressions that approximate the data by a curved equation.

The hypothesis of the curve fitting models is that the GLM is the best in adjusting to the population growth (COVID-19 cases) and by obtaining the lower part of the curve we can obtain the parameters of the function and hence obtain the complete curve, which can help us in estimating when the inflection point and limiting value will be reached.

The equations for each of these techniques are shown below. The first Equation (1) is for the linear regression and considers the slope of the line as b and c as the intercept of the value of y when $x = 0$. The Equation 2 shows the fitting of a polynomial regression with c being the set of coefficients and n the degree of the polynomial. Finally Equation 3 considers e as Euler's number, x_0 is the x value of the sigmoid's midpoint, L is the curve's maximum value, and k is the logistic growth rate.

$$y = cx + b \tag{1}$$

$$f(x) = c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4 + c_n x^n \tag{2}$$

$$f(x) = \frac{L}{1 + e^{-k(x - x_0)}} \tag{3}$$

The reason for using these models is that they can capture a variety of trends and patterns. In the first model, we adopt a pessimistic approach with the assumption that the exponential trend will continue indefinitely in the future, the second model is done to capture a variety of additive and multiplicative patterns in the data, and finally, the third model, assumes convergence, meaning that a stable state will be achieved.

2.0.3 Point of Inflection and Limiting Value

In this study, we predicted the point of inflection and limiting value by using the generalized logistic function. The point of inflection is the steepest part of the graph, which represents the time of the most rapid growth of the curve. The limiting value is the carrying capacity of the population and shows us the total number of predicted cases in the final stage of the epidemic [12].

2.0.4 Multivariate Time-Series Models

The most traditional statistical methods when approaching time-series forecasting are autoregressive integrated moving average (ARIMA), exponential smoothing techniques [13], and variate autoregression (VAR) methods [14]. In machine learning the most common techniques to approach this problem is the Long Short Term Memory (LSTM) network, however, other non-parametric algorithms can also be useful in this approach. In this study, we compare the results of a traditional time-series model (VAR) with a neural network model (LSTM) to determine which can better predict the number of cases and fatalities in Mexico.

2.0.5 VAR

The Vector Autoregressive Model is an extension of the univariate autoregression model for multivariate time series data. We decided to use this method since VAR has proven to be one of the most suitable and flexible models for the analysis of multivariate time series.

The model consists of a multi-equation system that treats all variables as endogenous (dependent) and is a linear function of past observations [15]. In its reduced form the equation includes lagged values for each of the dependent variables in the system. This form is shown in equation 4 where Y_t represents the vector of the time series variable, a is the vector of intercepts, A_i is the coefficients matrices and ϵ_t is a vector of white noises.

$$Y_t = a + A_1 Y_{t-1} + A_2 Y_{t-2} + \ldots + A_p Y_{t-p} + \epsilon_t \tag{4}$$

2.0.6 LSTM

The Long-Short Term Memory is an artificial recurrent neural network (RNN) architecture used in the field of Deep Learning [16]. This multi-layered neural network has the capability to avoid the long-term dependency problem by adapting non-linearities in the datasets [17], which makes it a significant performer in time-series data [18].

The concept of this network consists of the use of three nonlinear gates: the forget gate, the input gate and the output gate, and one "memory" cell. The "memory" cell transports relevant information through a sequence chain, and it can maintain its state value over a long time. In the process, the cell loses and wins information, and the gates are responsible for deciding what information should be added to the next time step and what should be removed. In this gates, the logistic or sigmoid function is used to transform the information into values between 0 and 1, to make a "Yes"/"No" decision and a hyperbolic tangent(τ) is used to transform the information to values between -1 and 1 to make a "negative"/"neutral"/"positive" decision [17].

The role played by the first gate, the forget gate(f), is in deciding what is to be forgotten from the previous state data and which weighted previously hidden state information is to be remembered. The second gate, the Input Gate (i) determines what information is relevant to be written onto the Internal Cell State. Inside the LSTM cell unit there are three outputs: C(t) , y(t) and h(t), the calculation performed is shown in the equations below Equations 5, 6, and 7. Where the w_0 represents the weights, the g_0 represents a nonlinear function, which can be the sigmoid function, and the $fff(t)$ represents an internal forget gate inside the input gate.

$$C(t) := f(t)C(t-1) + i(t) \tag{5}$$

$$y(t) := g_0((w_0 h(t))) \tag{6}$$

$$h(t) := fff(t)\tau(C(t)) \tag{7}$$

Finally, the Output Gate(o9) determines what the output(hidden state) is to be generated from the Internal Cell State; this is done by multiplying the $fff(t)$ result by the current cell state with values between -1 and 1.

2.1 Evaluation Metrics

To measure the performance of each of the models, the following metrics are computed: The Root Mean Squared Error (RMSE) is obtained to measure how close the fitted values are to the real values and the Akaike information criterion (AIC) is used to obtain the estimated likelihood to predict a model and to test how well the model fits the data without overfitting it [19]. The formulas for RMSE and ACI are shown in Equation 8 and Equation 9 respectively. The n in the RMSE formula represents the number of samples, the p_i is the forecasted values and the o_i is the real observed values. In the AIC formula, the k is the number of model parameters and the $\mathcal{L}$ is the log-likelihood measure as a measure of model fit.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2} \tag{8}$$

$$AIC = 2k - 2ln(\mathcal{L}) \tag{9}$$

3 Exploratory Data Analysis

During the data exploration, we compute the main statistics of the data and perform analysis through graphs and plots visualization. First, an initial table is obtained from the variables in the data set (Table 1). A bar graph with the cumulative number of confirmed cases(blue line) and the number of fatalities (orange line) reported worldwide can be observed in Fig. 1.

Table 1 Statistical Summary

	ConfirmedCases	**Fatalities**
count	33,909	33,909
mean	5,757	380
std	49,991	3414
min	0	0
25%	0	0
50%	28	0
75%	547	0
max	1,699,176	100,417

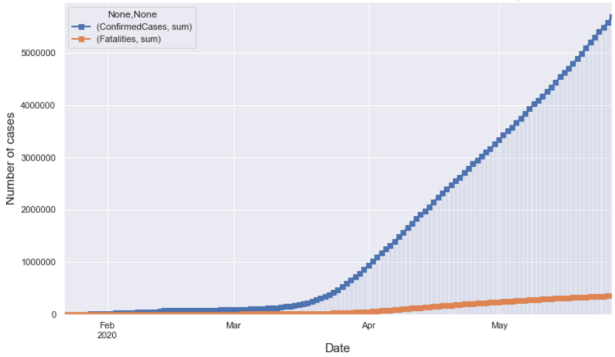

Fig. 1 Worldwide COVID-19 confirmed cases since January 22^{th} 2020.

Next, to visualize the course of the pandemic in Mexico in comparison to other countries in Latin America, we evaluated the confirmed cases and fatalities from five different countries: Mexico, Chile, Brazil, Peru, and Ecuador. The statistics of the different features for each of the selected countries is shown in Table 2. The growth factor of daily new cases is obtained by dividing daily new cases by the total number of cases accumulated in the previous day. The growth factor of daily new fatalities is the division of the daily new fatalities by the total number of accumulated fatalities of the previous day. Finally, the average mortality rate is obtained by dividing the daily fatalities by the daily cases.

Table 2 Data Summary

Country	**Mexico**	**Chile**	**Brazil**	**Peru**	**Ecuador**
Start	2/28/20	3/3/20	2/26/20	3/6/20	3/1/20
End	5/21/20	5/21/20	5/21/20	5/21/20	5/21/20
Mean Cases of Daily Accumulated	16,153	16,337	77,318	32,060	13,775
St. Dev. of Daily Accumulated	21668	20,796	110,300	42,013	14,005
Growth Factor of Daily New Cases	19.08%	17.24%	17.92%	19.93%	12.09%
Mean Fatalities of Daily Accumulated	2098	228	6,441	1,095	1,020
St. Dev Fatalities of Daily Accumulated	2471	221	7,503	1,191	1,102
Growth Factor of Daily New Fatalities	16.1%	12.84%	19.95%	12.49%	11.43%
Average Mortality Rate	5.67%	0.77%	4.18%	2.23%	4.3%

The countries of interest are plotted and show their confirmed cases in Fig. 2 and fatalities in Fig. 3. These graphs show the country's amount of cases increasing through time. We can observe that even though Brazil has a higher amount of daily cases, Mexico has a higher average mortality rate. Also, we

can see some gaps in the data from Ecuador, as there are some abrupt steps observed in the graph.

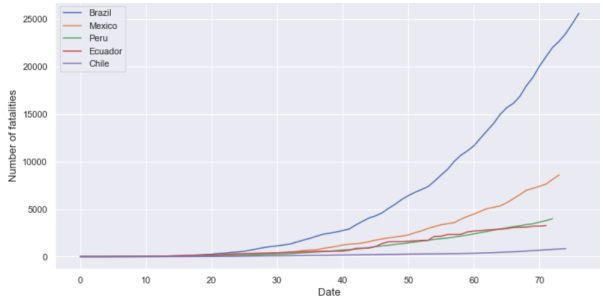

Fig. 2 Total number of confirmed cases.

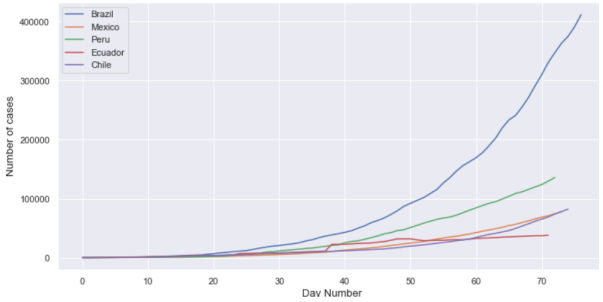

Fig. 3 Total number of deaths.

3.1 Data Preparation

The Data preparation phase consists mainly of data cleaning and feature reduction. The data set included confirmed cases and fatalities for Mexico, obtained from John Hopkins Repository is mainly clean in terms that it does not contain inconsistent or missing values. This is the same case for climate data. In regards to the social mobility rate, there were missing values in the most recent dates, the data is updated until March 21st. Due to this limitation, the models used in this study consider this date as the last one.

Additional to this, a data transformation was performed in the dependent variable of the linear growth model as it was identified in the visualization phase, that the data does not have a linear evolution through time but exponential. For this, we make a natural logarithmic transformation to the output variable (Confirmed Cases) to simulate a linear behavior and be able to use this for prediction. After performing this, we included a row with the resulting

logarithmic transformation in our table. We can observe in Fig. 4 how the right graph, with the log transformation, shows a straight line. A transformation was also made to add a new variable showing the number of days since the start date of the reported outbreak (January 22nd, 2019). The row shows day number 0 for January 22nd and day 69 for March 31st.

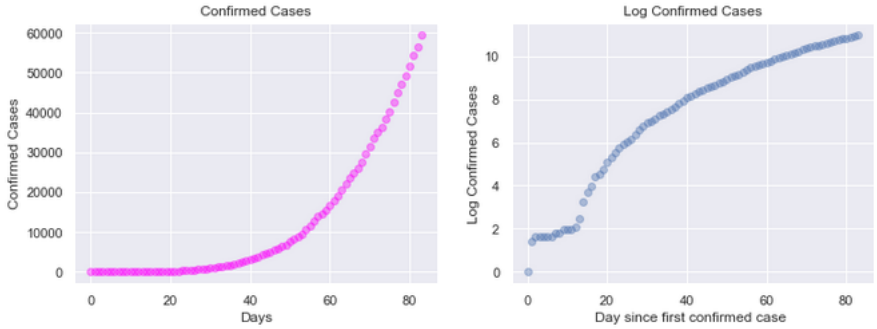

Fig. 4 Logarithm transformation of confirmed cases.

For the time-series models, additional preparation was required. For data to be fitted with time-series models, it is important to ensure stationarity of all the time series variables. For this purpose, we first included a new column to transform the confirmed accumulated cases into daily cases by removing the immediate prior date number of cases. Then, we performed a logarithmic transformation on the daily number of cases and the daily number of fatalities. Following, to transform the numeric values to a common scale, we use the z-score normalization method [20].

Specifically for the LSTM model, we are required to transform the time series data to a supervised learning problem. For this, we included the time lag variables for each of the covariates and dependent variables.

Then to avoid overfitting problems and to determine the most important input parameters, we performed a feature selection with the use of the Spearman Rank Correlation coefficient. This statistical method is used since it is robust when dealing with non-normally distributed data [21]. This method is a filer selection method so the selection is done before applying any machine learning algorithm.

Finally, to test the performance of each model, we separated the last 20% observations as a testing set and used 80% of the data to train the model. With this, we can use the hold-out data to test our predictions.

4 Results

In this study we evaluated the COVID-19 infected population growth in Mexico by comparing it to three curve fitting models: Linear, Polynomial, and

Sigmoid Curve models, and then considered the generalized logistic growth model to determine the inflection point in Mexico. For the second and third objective of this study, we used the Spearman Rank Correlation to select the most important features and use these features in two time-series models: VAR and LSTM. Finally, we compare the prediction results from these two models. The process followed in this second step is shown in Fig. 5.

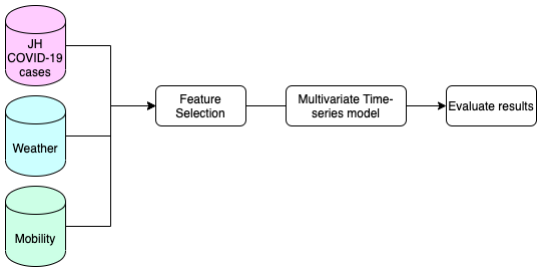

Fig. 5 Model diagram.

4.1 Population Growth Models

The linear model was constructed with the last 20 days of data, in an attempt to understand if the growth behaved exponentially. The linear regression model for Mexico's confirmed cases is plotted in Fig. 6 and the accumulated number of fatalities in Fig. 7. In the graph, we can observe that the linear regression, despite being a simple model can accurately fit the logarithm data. We can see that these last 20 days closely resembled an exponential growth except for the last observations.

Additionally, the results of the equation's coefficients, AIC, and RMSE for both target variables are shown in Table reftab:FittingResults. The RMSE is high for the confirmed cases in linear regression, it seems that the case numbers from Mexico are not behaving exponentially anymore. On the other hand, Mexico Covid19 fatalities have a low RMSE which indicates that the growth of fatalities in the last 20 days is still fitting well with this model.

The second model created to fit the data and predict the confirmed cases of coronavirus for the following weeks was a polynomial regression. Polynomial regression is a form of linear regression but where the dependent variable is modeled with an nth degree parameter. For this regression, we performed a tuning process and obtained the best results with a 4th-degree parameter performed on 80% of the available data. The results obtained are shown in Table 3. We can see from the graphs in Fig. 8 and Fig. 9 that the model fitted the data well only in the initial stage of the infection.

The third model used to fit the data and predict the confirmed cases of coronavirus for the following weeks is the generalized logistic model. The logistic function resembles the behavior of a pandemic, so the models created with

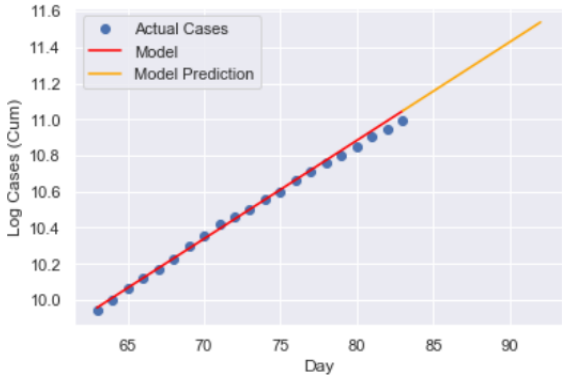

Fig. 6 Linear regression model for logarithm confirmed cases of the last 20 days.

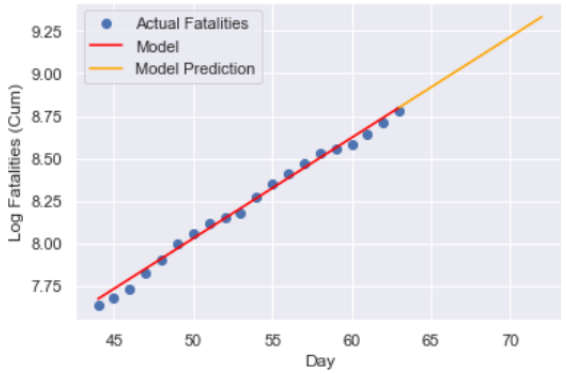

Fig. 7 Linear regression model for logarithm fatalities of the last 20 days.

Table 3 Growth Models Results

Models	**Confirmed Cases**			**Fatalities**		
	Coef.	RMSE	AIC	Coef.	RMSE	AIC
Linear Regression	c1=0.05 b=6.52	2299.24	81.40	c1=0.06 b=5.06	135.82	43.29
Polynomial Regression	c1=-16.67 c2=1.69 c3=-0.06 c4=0.01 b=34.14	3781.91	284.09	c1=-0.96 c2=-0.15 c3=0.03 c4=-0.01 b= 7.43	179.98	139.01
Sigmoid curve fitting	L=99,592 x_0=79.04 k=0.09	535.57	845.97	L=11,036 x_0=60.05 k=0.09	102.72	476.47

this function are expected to follow this behavior. The curve fit function was implemented to get the best possible coefficients that better adjust to the data behavior in the training set and the results obtained for Mexico are shown in Table 3.

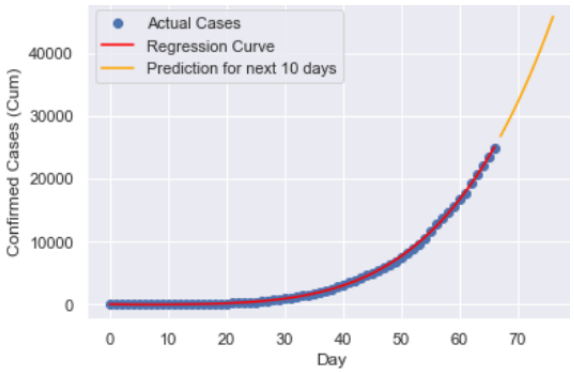

Fig. 8 Polynomial regression model for Mexico confirmed cases.

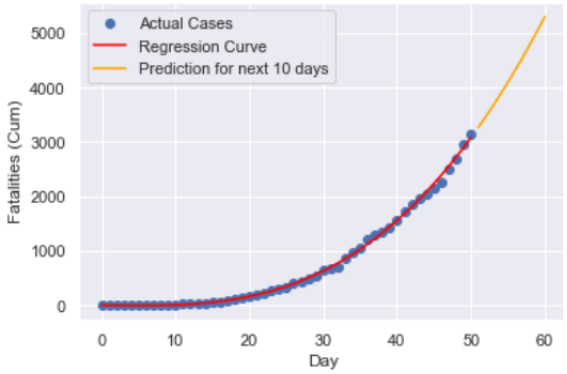

Fig. 9 Polynomial regression model for Mexico fatalities.

We can conclude from the results that the sigmoid curve, in comparison to the other two curve-fitting models, fits best the behavior of the infected population growth in Mexico for both the accumulated daily cases and daily fatalities.

Finally, this last model is used to make predictions for the next 100 days with the input of the complete dataset. We can see these predictions in Fig. 10 and Fig. 11. In the next section, the point of inflection and limiting parameters are obtained.

4.1.1 Point of Inflection and Limiting Value

In this section, we predict the point of inflection and limiting value by using the generalized logistic function. With this, we determine the shape and some general features of the infection growth, and we can see that even with this simple equation and using only one variable we can reach interesting results. Table 4 summarizes the results of both the accumulated cases and fatalities; this considering that the compliance of the lockdown remains the same.

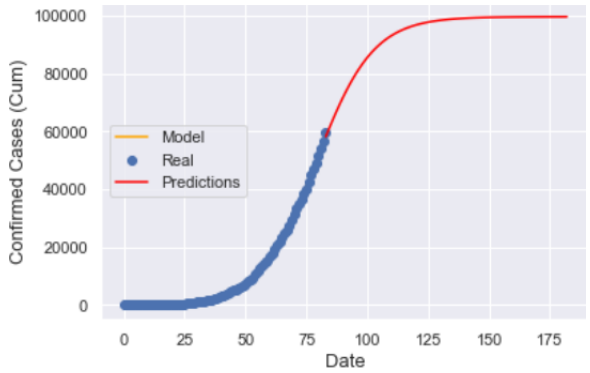

Fig. 10 Polynomial regression model for Mexico confirmed cases.

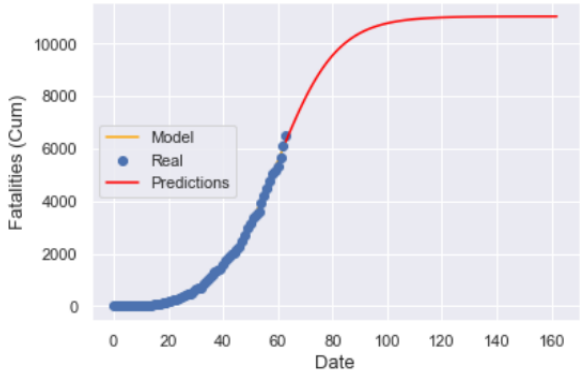

Fig. 11 Polynomial regression model for Mexico fatalities.

Table 4 Inflection point and limiting values of confirmed cases and deaths.

	Accumulated Cases	**Accumulated Fatalities**
Inflection Point	49,796 (May 18^{th})	5,518 (May 19^{th})
Limiting value	99,592 (September 29^{th})	11,036 (August 27^{th})

4.2 Feature Selection

A correlation analysis was performed to determine the most important input parameters which will be used to build the multivariate time-series models. To understand which correlation method to use for this, we first need to understand if the data is normally distributed. To test normality, we used the Shapiro-Wilk test. The results were a coefficient of 0.7579 and a p-value smaller than 0.05 and for the Fatalities a variable coefficient of 0.8152 with a p-value smaller than 0.05. There was enough statistical evidence to reject the null hypothesis and determine that the data was not normally distributed.

The plots in Fig. 12 show the density behavior of the data as a QQ-plot for confirmed cases.

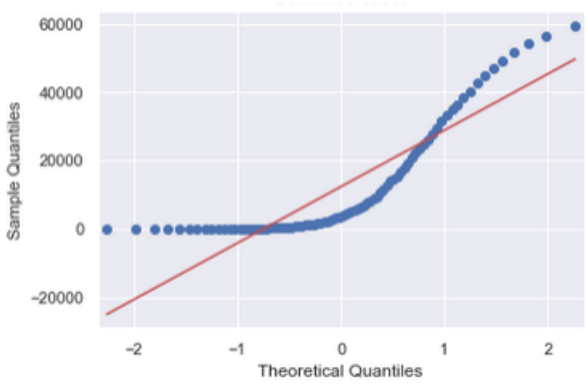

Fig. 12 Confirmed cases qqplot.

Following, the data were transformed into a supervised problem to obtain the t-n observations for time series in an attempt to understand which time lags and variables hold the highest linear coefficient correlation with the output variable. We used a Spearman coefficient matrix since this is more robust when dealing with non-normality. Next, an absolute mean of each variable was calculated for each time step. The results are shown in Table 5.

Table 5 Time step new daily cases and fatalities feature selection.

Variables	**ID**	**New cases Absolute Mean**	**Fatalities Absolute Mean**
New log cases	var1	0.84	0.53
MaxtempC	var2	0.30	0.26
MintempC	var3	0.31	0.24
UVIndex	var4	0.23	0.19
Humidity	var5	0.30	0.21
PrecipMM	var6	0.33	0.26
Pressure	var7	0.24	0.19
WindspeedKmph	var8	0.15	0.15
Retail and Recreation	var9	0.32	0.34
Grocery and Pharmacy	var10	0.38	0.32
Parks	var11	0.35	0.35
Transit stations	var12	0.40	0.35
Workplaces	var13	0.21	0.35
Residential	var14	0.21	0.33

We identified that the top five features with the highest correlation with respect to the daily cases are the amount of cases at $t-14$ (0.92) , the transit stations mobility rate $t-28$ (-0.79), parks mobility rate at $t-26$ (-0.76), grocery

and pharmacy mobility rate at $t-26$ (-0.75), and maximum temperature (°C) at $t-28$ (-0.70). On the other hand, for the daily fatalities the top five features with the highest correlation with respect to the daily fatalities are the daily cases at $t-7$ (0.84), the residential mobility at $t-28$ (0.78), parks mobility rate at $t-1$ (0.75), transit stations at $t-28$ (-0-.71), and grocery and pharmacy mobility rate at $t-25$ (-0.70).

The defined threshold used for the confirmed cases and fatalities was 0.30 and 0.25 respectively. All attributes that had an absolute mean value less than 0.30 were eliminated (UV index, pressure, wind speed, workplaces mobility rate, and residential mobility rate) from the confirmed cases dataframe. In regards to fatalities, all attributes with an absolute mean spearman correlation coefficient value less than 0.25 were eliminated (minimum temperature, UV Index, humidity, pressure, and wind speed).

4.3 Multivariate Time-Series Models

In this phase, we compare the scores of the two multivariate time-series models (LSTM and VAR) to identify which one is the best at predicting the new daily cases and fatalities caused by COVID-19 in Mexico.

Since Autoregressive models perform best when the time set is stationary, an Augmented Dickey-Fuller (ADF) test was done to prove stationarity. The null hypothesis of this test is that the data set has a unit root. Since the result was a p-value greater than 0.05, we rejected the null hypothesis with a 95% confidence and determined that the data set was not stationary.

The data was then transformed to a logarithm scale and a new output variable was created by using the differencing method. After the transformation, the time series was shrunk to 56 preview days. When performing the ADF test, we did not reject the null hypothesis and determined that the transformed data now hold a stationary property. Finally, the data were normalized with the z-score function.

Following this preparation, we fitted the VAR and LSTM model. After a series of experiments, the VAR model showed the best results with a time lag of 7 for both dataframes. For the neural network model, we used a two-layer LTSM with 200 neurons in the first layer, 100 in the second layer, and a time-lag of 28 days. The results are shown in Table 6. For this analysis we used t_0 as of May 21$^{\text{st}}$.

The results of the models are shown in Table 6. The computed RMSE and AIC help us do a model comparison between the models created to select the one with the smallest values. The values show that the best model for predicting both the daily cases and daily fatalities is the LSTM model with an RSME smaller in 47.16% for the cases and 33.27% for the fatalities.

Table 6 Time series models metrics summary.

Model	RMSE	AIC
LSTM daily cases	297.3051	74.28
LSTM daily fatalities	69.3535	55.78
VAR daily cases	630.3469	94.0
VAR daily fatalities	208.4456	79.0

5 Conclusion and Future Work

The contribution of this paper is threefold. First, we defined the population growth of the cumulative cases and cumulative fatalities with three growth models and used the use of logistic curve fitting to predict the inflection point and limit the value of the COVID-19 outbreak in Mexico. Based on the results of the logistic curve-fitting model, we determined the inflection point was on May 15^{th} 2020 and predict that the possible limit value of this outbreak in Mexico will be 99,592 cases and will be reach around the end of September; this considering that the compliance of the lockdown remains the same.

Secondly, we identified several relational features that can be used to predict COVID-19 daily cases with the use of exploratory data analysis. We identified the features with the highest correlation to the daily cases and fatalities were: the number of cases, the transit stations mobility, parks mobility, the grocery pharmacy mobility, the residential mobility, and the maximum temperature (°C).

Third, we demonstrated that it is better to use LSTM for this prediction in comparison to the traditional statistical model of VAR, as we obtained significantly better results with an RSME smaller in 47.16% for the new cases and 33.27% for fatalities.

Finally, there are several interesting approaches in which this work could be extended. For instance, this model can serve as a baseline and be adapted with social media information. Furthermore, the same analysis can be useful to make predictions for other countries and even globally. Due to time constraints, we were not able to test a stacking method, but we will like to update this work by combining the fitted VAR and using it to improve the performance of the LSTM model. We hope that this study can make some contributions to the world's response to this epidemic as well as give some references for future research.

Acknowledgements We thank the team from Johns Hopkins University Center for Systems Science and Engineering (CSSE) for their public service by collecting the data of COVID-19 from around the world and sharing it for this study. We also acknowledge the support of Tecnologico de Monterrey and CONACyT. Finally, we acknowledge Tecnologico de Monterrey's intelligent system research group for their endorsement.

Conflict of interest

Conflict of Interest: Daniela A. Gomez-Cravioto declares that she has no conflict of interest. Ramon E. Diaz-Ramos declares that he has no conflict of interest. Francisco J. Cantu-Ortiz declares that he has no conflict of interest. Hector G. Ceballos declares that he has no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. W.H. Organization, Emergencies preparedness, response, Disease outbreak news, World Health Organization (WHO) (2020)
2. Home - Johns Hopkins Coronavirus Resource Center (2020). URL https://coronavirus.jhu.edu/
3. G. Chowell, A. Tariq, J.M. Hyman, BMC medicine **17**(1), 164 (2019)
4. S. Chae, S. Kwon, D. Lee, International journal of environmental research and public health **15**(8), 1596 (2018)
5. GitHub - CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE. URL https://github.com/CSSEGISandData/COVID-19
6. Historical Weather API from World Weather Online. URL https://www.worldweatheronline.com/developer/api/historical-weather-api.aspx
7. COVID-19 Community Mobility Reports. URL https://www.google.com/covid19/mobility/index.html?hl=en
8. J. Liu, J. Zhou, J. Yao, X. Zhang, L. Li, X. Xu, X. He, B. Wang, S. Fu, T. Niu, Science of the Total Environment p. 138513 (2020)
9. A. Tomar, N. Gupta, Science of the Total Environment **728**, 138762 (2020). DOI 10.1016/j.scitotenv.2020.138762
10. B.B. Frey, The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation **2**, 772 (2018). DOI 10.4135/9781506326139.n296
11. P. Vidyullatha, D.R. Rao, International Journal of Electrical and Computer Engineering **6**(3), 974 (2016)
12. B. Crauder, B. Evans, A. Noell, *Functions and change: A modeling approach to college algebra* (Nelson Education, 2013)
13. R.J. Hyndman, G. Athanasopoulos, *Forecasting: principles and practice* (OTexts, 2018)
14. H. Lütkepohl, *New introduction to multiple time series analysis* (Springer Science & Business Media, 2005)
15. E. Zivot, J. Wang, Modeling Financial Time Series with S-Plus® pp. 385–429 (2006)
16. F.A. Gers, J. Schmidhuber, F. Cummins, (1999)
17. S. Skansi, *Introduction to Deep Learning: from logical calculus to artificial intelligence* (Springer, 2018)
18. Z. Karevan, J.A.K. Suykens, Neural Networks (2020)
19. E.E. Leamer. Chapter 5 Model choice and specification analysis (1983). DOI 10.1016/S1573-4412(83)01009-0
20. S. Patro, K.K. Sahu, arXiv preprint arXiv:1503.06462 (2015)
21. M. Savić, V. Kurbalija, M. Ivanović, Z. Bosnić, in *International Conference on Model and Data Engineering* (Springer, 2017), pp. 248–261