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.

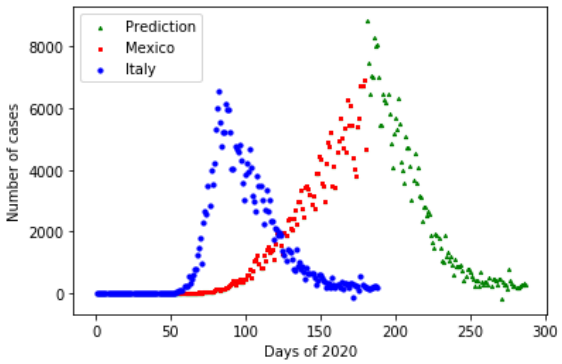

Fig. 1. Prediction of the pandemic in Mexico based on the data of Italy.

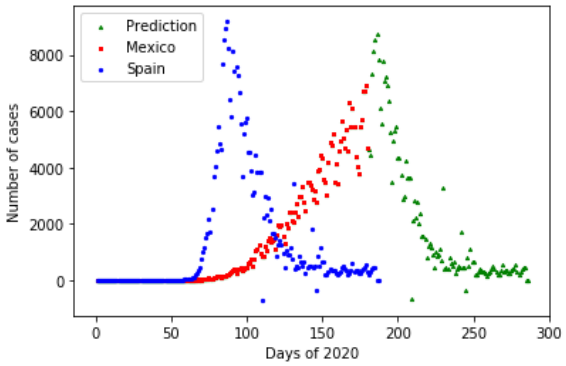

Fig. 2. Prediction of the pandemic in Mexico based on the data of Spain.

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	200th day of 2020	190th day of 2020
Pandemic End	300th day of 2020	290th 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 (200th 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 300th 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 (190th day of 2020), and the pandemic ends in Mexico in mid-October (290th day 2020).

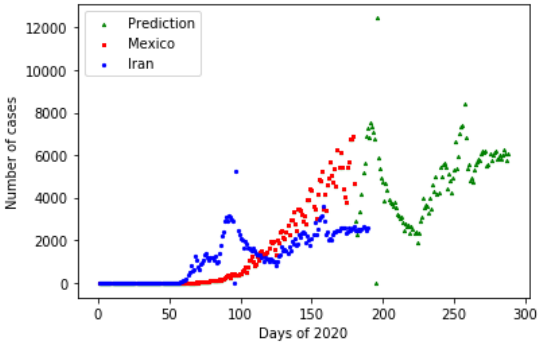

Fig. 3. Prediction of the pandemic in Mexico based on the data of Iran.

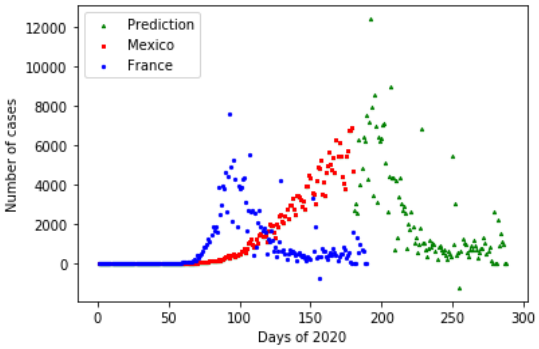

Fig. 4. Prediction of the pandemic in Mexico based on the data of France.

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	200th day of 2020	200th day of 2020
Second Peak	225th day of 2020	280th 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 (200th day of 2020), and the number of cases decreases after that. However, after mid-August (225th day of 2020), again the number of cases increases until Mexico experiences its second peak around mid-September (260th 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 (200th day of 2020) when Mexico experiences about 8500 affected cases per day, and the pandemic gets close to its end in late October (the 300th day of 2020); however, based on this model, we will see a second (or more) small peak between 250th and 300th 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)