Covid-19: An Epidemic Study

Manuel A. Hernández Alonso
Faculty of Mathematics and Computer Science
University of Barcelona
Gran Via de les Corts Catalanes, 585, Barcelona
manuelheralo@gmail.com

Abstract

The coronavirus disease 2019 originated by severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) generated a crisis due to a pandemic around the world. Forecasting and prediction on cases outbreaks was crucial for the proper management of logistics and resources of different countries. Several studies have been performed on the use of ARIMA (Autorregressive integrated moving average) models on forecasting and prediction of cases of COVID-19 with great success and accuracies. Additionally, some studies have reviewed the use of deep learning models such as LSTMs. The data used for this particular research has been sourced from Singapore and the United States to review the particularities of the different cases. This study limited the scope to the use of ARIMA models, ARIMAX models and LSTMs to generate insights in the data. We found that smoothing the data with last-7-days average really helps the model when the data has some erratic behaviour caused by different sampling rates aggregated into a single series. Lastly, LSTMs models have shown that they need complexity and lots of data to be used properly in these scenarios.

1 Introduction

The COVID-19 pandemic caused crisis among all countries, caused by the coronavirus disease 2019 originated by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), started in Wuhan, Hubei Province, China [12]. During this time several governments and organizations had to coordinate themselves to implement and apply new measures to contain the epidemic. Since the first outbreak, the virus has mutated into several variants such as alpha, beta and delta SARS-CoV-2 variants, which have produced several new waves of infections [17].

However, this pandemic generated new insights and data from disease epidemics that wasn't available before. Thus, this data was, and still is, being used for research into pandemic modeling, logistics, prognosis, and many other targeted research. For the scope of this paper, we are interested in those that have used methods from time series analysis.

Time series analysis, particularly through the auto regressive integrated moving average (ARIMA) model, has shown significant accuracy in forecasting infectious diseases [8, 18]. ARIMA has been applied to predict the number of new COVID-19 cases, deaths, and recoveries using daily reported data from various countries, aiding in the assessment of future outbreaks [1, 6, 15, 14]. Additionally, some deep learning models have been tried and tested on this data generated by the COVID-19 pandemic [3, 13]. Although there are many advanced data-driven time series methods available, developing new and more accurate prediction models remains essential during a pandemic.

Lastly, the objective of this study is to test these models and further inspect the modeling capability of these techniques over the data generated by health organizations and governments during the COVID-19 pandemic.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

The rest of the paper is organised as follows: Section 2 introduces the data and methods used in this study, Section 3 presents the results obtained with the models on the two different datasets, and Section **??** ends with the main remarks about the work carried out.

2 Methods

In this section we will explain the data used for this research, the methods employed to investigate and model the data collected, and the metrics used to compare the different models and statistics.

2.1 Data

During the COVID-19 pandemic, several hospitals and health organization collected data about the tests performed, positivity of the tests, deaths and other relevant information. In particular, this study is focused on two datasets collected and maintained by different organizations, (a) the dataset collected in Singapore by the Ministry of Health[4], and (b) the dataset collected in the United States by the government[7], and published by the State of California.

In the first dataset we can find the confirmed cases, the deaths, the recovered and other information per day. However, this dataset has missing ranges of data that make it difficult to work with. This dataset was used as a first approach to the data we are going to work later on, with different approaches and techniques applied to it to test the adaptability of the models to the COVID-19 data series.

On the other hand, the US dataset contains less information, just 3 columns of confirmed, deaths and tests performed. In this case, we don't have missing ranges and we can work with 1418 days of continuous data. This dataset is segmented by counties and states, so we summarized all the dataset into a single complete dataset of US COVID-19 data.

Additionally, dates of when the different epidemic waves happened were used. The exact timeline and dates are up for debate, as different countries and regions had different courses of the epidemic as different measures were implemented at distinct times. Different studies such as, Simona Iftimie et. al. (2021) [10], Valeria Caramello et. al. (2022) [2], and Somyanonthanakul, R. et al. (2022) [16], have different evaluations of the start and end of waves. For the purpose of this study, the different waves were segmented into similar waves shapes described on the papers. The third and fourth wave are heavily disputed in research relating to the COVID-19 epidemic, for this paper we decided to fuse both of them into a single split. Specifically, the dates used are described in the following table 1

Table 1: Waves during the COVID-19 epidemic.

Wave	Start	End
1st Wave	2020-03-26	2020-08-16
2nd Wave	2020-10-15	2021-03-14
3rd+4th Wave	2021-05-31	2021-09-30

Finally, the total data points used for the Singapore dataset was 806 time steps (2020-01-23 to 2022-04-07), and the rest of the 216 time steps were only used in visualization and qualitative analysis of the predictions. The total data points for the US dataset was 1418 time steps, that correspond to dates between 2020-02-01 and 2023-12-19.

2.2 Statistics

Some preliminary data analysis was performed by using some significant statistics of the model. We used the rolling mean and standard deviation to see the fluctuation over time of the different datasets. Additionally, to test the stationarity of the datasets we used the Dickey-Fuller test [5], where we test for the existence of a unit root of $\rho = 1$ in a simple AR model $y_t = \rho y_{t-1} + u_t$, and set the p-value for rejection at 0.05 in 15 lags.

2.3 Models

2.3.1 Traditional Models

The autorregressive (AR) model: This simple model is used to model the predictive value at the time step t by the observed values at previous time steps $t-1, t-2, \cdots, t-p$. The weight of each of the previous observed values is modeled by a coefficient β_p, and the existence of an intercept β_0. This can be formulated as follows:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \epsilon_t \tag{1}$$

We also have to consider that there will be some residual error from the prediction of the model, this is modeled by the ϵ_t value.

The moving-average (MA) model: This model alleviates the impact of unexpected external factors such as noises by using a moving average. The predicted value is then calculated by the q lagged forecast errors e_i from an AR model. Formally,

$$y_t = \phi_0 + \phi_1 \epsilon_{t-1} + \cdots + \phi_q \epsilon_{t-q} + \epsilon_t \tag{2}$$

where we isolate the ϵ_i from previous time steps in an AR model 1, as follows

$$\begin{aligned} \epsilon_{t-1} &= y_{t-1} - (\beta_0 + \beta_1 y_{t-2} + \cdots + \beta_p y_{t-p-1}) \\ \epsilon_{t-2} &= y_{t-2} - (\beta_0 + \beta_1 y_{t-3} + \cdots + \beta_p y_{t-p-2}) \\ &\vdots \\ \epsilon_{t-q} &= y_{t-q} - (\beta_0 + \beta_1 y_{t-q-1} + \cdots + \beta_p y_{t-p-q}) \end{aligned} \tag{3}$$

So the residual errors are modeled into the MA model by using the information of several forecasts lagged q steps behind the predicted value y_t.

The autorregressive moving average (ARMA) model: Combining the previous two models we can get model with more representative power. The previous lags are smoothed with the average, but now they are also directly considered for the future values of the time series. Formally we can represent this as

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \phi_0 + \phi_1 \epsilon_{t-1} + \cdots + \phi_q \epsilon_{t-q} + \epsilon_t \tag{4}$$

In this model, we can denote the order using p and q, where the ARMA model would be defined as ARMA(p, q).

The autorregressive integrated moving average (ARIMA) model: This expansion of the ARMA model uses differencing to exploit the stationarity of the series. We can manipulate the differentiation factor d on different time lags to alleviate noise and decompose the series into trends, seasonal, and residual components. We can define the ARIMA model with $d = 1$ as follows:

$$y'_t = \beta_0 + \beta_1 y'_{t-1} + \cdots + \beta_p y'_{t-p} + \phi_0 + \phi_1 \epsilon_{t-1} + \cdots + \phi_q \epsilon_{t-q} + \epsilon_t \tag{5}$$

with d the degree of the differentiation, y'_i is then defined by the subtraction of previous time steps of the series data. We can compute then

$$y'_i = y_i - y_{i-1} \tag{6}$$

in the case of $d = 1$. We can then further generalize continuing the pattern for a second order differentiation as such

$$y''_i = y'_i - y'_{i-1} y'' = (y_i - y_{i-1}) - (y_i - y_{i-2}) y'_i = y_i - 2y_i + y_{i-2} \tag{7}$$

The autorregressive integrated moving average with exogenous covariates (ARIMAX) model: We can assume then that the data is also affected by other external variables. We can integrate this covariates into the forecasting model and improve its prediction accuracy. We can do that by defining $(X_i)_t$ as the exogenous variable i at time t, and define a corresponding coefficient θ for the variable i. Formally:

$$y'_t = \beta_0 + \beta_1 y'_{t-1} + \cdots + \beta_p y'_{t-p} + \phi_0 + \phi_1 \epsilon_{t-1} + \cdots + \phi_q \epsilon_{t-q} + \theta_1 (X_1)_t + \cdots + \theta_m (X_m)_t + \epsilon_t \quad (8)$$

where m defines the number of exogenous covariables to be considered.

2.3.2 Deep Learning Models

As a small experiment, a simple deep learning model was implemented using Long Short-term Memory (LSTM) networks. These LSTMs are based on the Recurrent Neural Networks (RNN) models used to extract features and future states of sequences. The main problem with these RNNs were the vanishing gradients and exploding gradients. These limitations can be addressed by the LSTMs with better long-term dependencies using memory cells in the hidden layers [9].

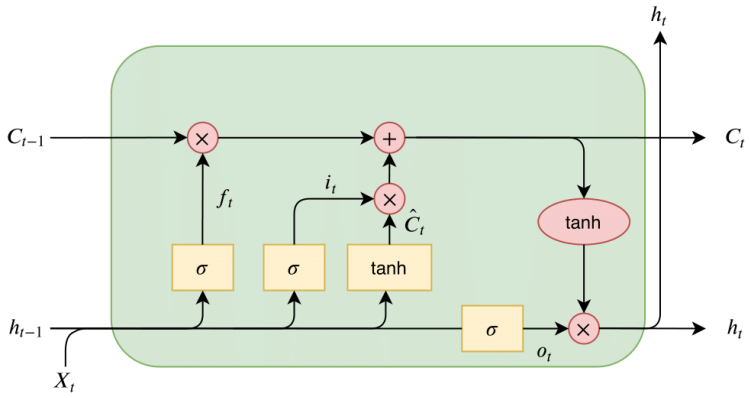

Figure 1: Scheme of an LSTM cell. Source: `https://thorirmar.com/post/insight_into_lstm/`.

This LSTM cell model (Figure 1) calculates a hidden state output h_t and the subsequent intermediate state C_t by

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \\ o_t &= \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \\ \hat{C}_t &= \tanh(W_C \cdot [h_{t-1}, X_t] + b_C) \\ C_t &= i_t \cdot \hat{C}_t + f_t \cdot C_{t-1} \\ h_t &= o_t \times \tanh(C_t) \end{aligned} \quad (9)$$

With $\hat{C}_t$ being the internal state of the cell, the forget gate f_t, the input gate i_t and the output gate o_t. All gates have their own weights and biases ($W_f, b_f, W_i, b_i, W_o, b_o$ and W_C, b_C) that can be optimized via gradient optimization to learn from data.

We can then chain multiple LSTM cells together to have more representative power at the cost of more parameters to train. This is what allows the LSTM-based networks to retain information from past events of the input sequence without losing gradients with the vanishing gradient problem that affects the vanilla RNNs.

Lastly, the used LSTM network in this research was the LSTM layer implemented in Keras. We used 50 units of output space with 2 layers of LSTM and a final dense layer to get the predictions of the next timestep.

2.4 Performance Metrics

To analyse the validity of the residuals and test the datasets we used the autocorrelation (ACF) and partial autocorrelation (PACF) functions implemented in the Python module statsmodel. Furthermore, to test the properties of the residuals we used the Ljung-Box [11] test with p-value of 0.05. This test can be performed with a simple statistic as follows:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k} \tag{10}$$

where $\hat{\rho}_k$ is the auto-correlation in the lag k, and h is the number of lags being tested. This test, the ACF and the PACF helps us validate if the given model is a good fit for the data.

Additionally, we used several performance measures to compare between the different models. Firstly, we used the coefficient of determination ($\mathbf{R}^2$), which represents the proportion of the variation in the dependent variable that can be predicted from the independent variable. Formally,

$$\begin{aligned}
\mathbf{R}^2 =& 1 - \frac{SS_r}{SS_t} \\
SS_r =& \sum_i (y_i - \hat{y}_i)^2 = \sum_i \epsilon_i^2 \\
SS_t =& \sum_i (y_i - \overline{y})^2 \\
\overline{y} =& \frac{1}{n}\sum_i y_i
\end{aligned} \tag{11}$$

where $\hat{y}_i$ is the predicted value at time step i, SS_r is the sum of squared residuals, SS_t is the sum of total squares proportional to the variance of the data, and $\overline{y}$ is the mean of the observed data. This $\mathbf{R}^2$ helps us identify when the model is a good fit for the data as it models how good the fit is of the predictions by the model and the observed values.

Finally, we used the root of mean squared error (RMSE) and the mean absolute difference (MAE) to measure the distances between data and the regression line. Specifically, RMSE was also used as the loss function for the LSTM model. Formally we can define both as:

$$\begin{aligned}
RMSE =& \sqrt{SS_r} = \sqrt{\frac{1}{2}\sum_i (y_i - \hat{y}_i)^2} \\
MAE =& \frac{1}{n}\sum_i |y_i - \hat{y}_i|
\end{aligned} \tag{12}$$

These two measures were chosen because the RMSE gives larger significance to large error, and MAE being an straightforward way to assess the fit of the models.

3 Results

3.1 Experimental Setup

3.1.1 Singapore Dataset

This dataset was used tentatively to test the different methods presented in Section 2. We can see the relevant stastics of the dataset from Singapore (until the first missing range in 2022-04-08) in Figure 2. The results for the stationarity of the data indicated that it was stationary, as the Dickey-Fuller test showed a p-value of 7.6×10^{-5}, way less than the needed p-value of 0.05.

Then, an ARMA(4,2) model was tested since the data was stationary. This order was found with Python's statsmodels *armar_order_select_ic* function to find the best minimum order of an ARMA

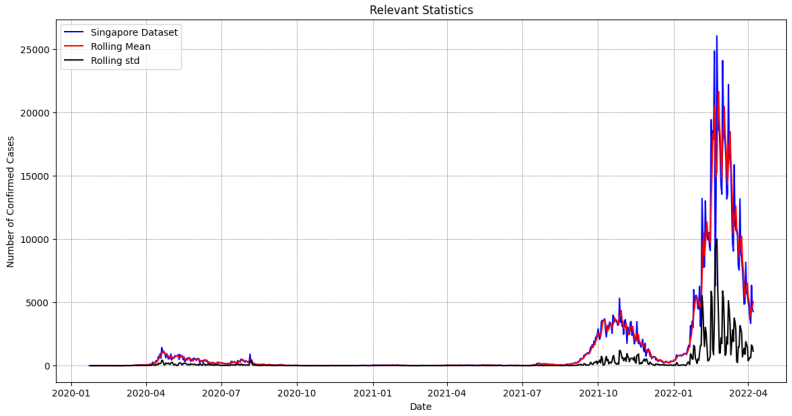

Figure 2: Plot of the daily confirmed cases in Singapore (until 2022-04-08) and the relevant rolling statistics.

model that fits the data. Nevertheless, the model resulted in a poor performance, as the residuals were not independent identically distributed noise. This was confirmed by the residuals ACF/PACF (Figure 3) and the Ljung-Box test. This last test scored a p-value of 6.06×10^{-53}, indicating that there's still significant correlation between the residuals.

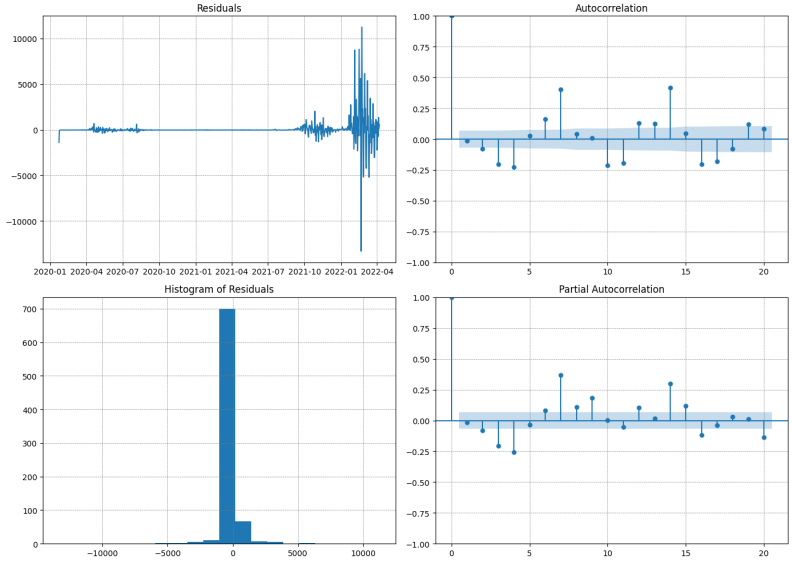

Figure 3: Plot of (a) Residuals, (b) Residuals Histogram, (c) Auto-correlation Function on 20 lags, (d) Partial Auto-correlation Function on 20 lags

Even with a not too good model, we performed some predictions (Figure 4) to see how confident it would be filling the missing range of values that the Singapore data had. It seems that it would eventually connect with the rest of the data, but it wouldn't predict the next wave. Nonetheless, on a real-time use, it may be able to predict up to a week ahead to know how the pandemic curve is behaving.

Lastly, the 2-layered LSTM network with 50 units was tested on the data, and it seemed to have a similar performance to the ARMA model. The data wasn't enough for it to properly predict the next time steps, but it was able to adjust the predictions to the model with less auto-correlated residuals. Additionally, last-7 days smoothing was tried, and residuals were smoother, but the prediction in the missing range still tended to the low values.

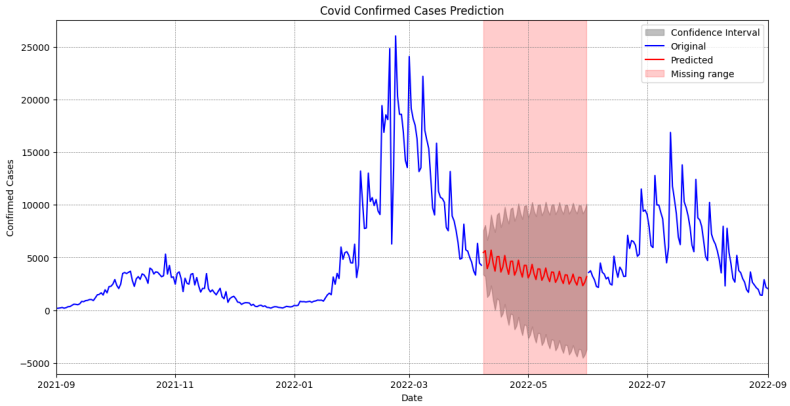

Figure 4: Plot of the Singapore daily confirmed cases with predicted values on the missing range.

3.1.2 United States Dataset

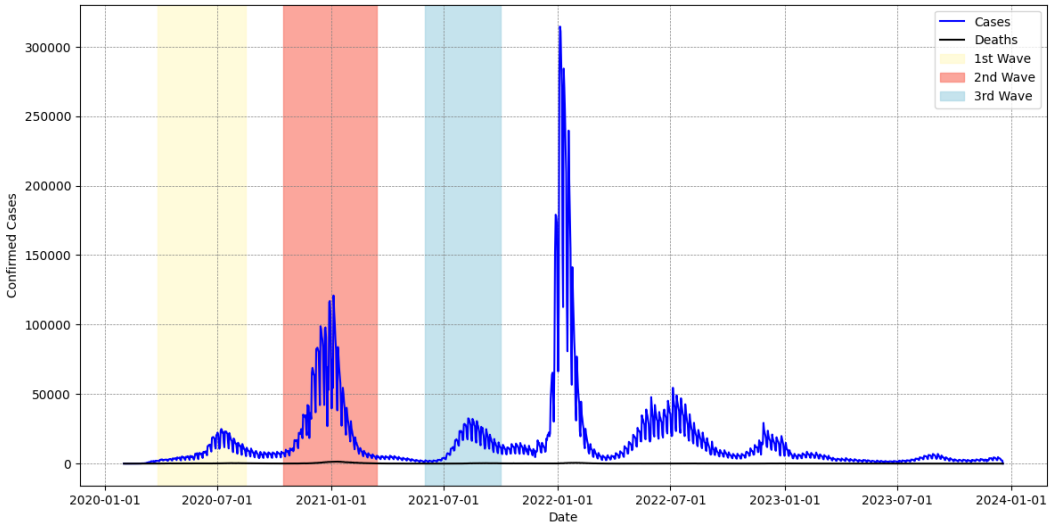

Figure 5: Plot of the daily confirmed cases in the United States.

In this dataset we performed more throughout modeling and testing. We segmented the tests into last 7-days average, weekly sum reports, and the full unprocessed dataset. Additionally, we segmented these into several splits: 1st Wave, 2nd Wave, 3rd+4th Wave, and the complete range of data (Table 1). A quick look into the data (Figure 5) had let us see some repetitive behaviours due to some states reporting more frequently than others.

Similarly to the previous dataset, we computed the rolling statistics and the Dickey-Fuller test. The statistic did not show any relevant information, but the test showed that the data was stationary as it got a p-value of 6.05×10^{-4}, less than the critical value $p < 0.05$. Due to this we continued with the data unprocessed as we did not have to convert it from non-stationary to stationary.

Unprocessed Dataset: We then fit AR, I, MA, ARMA, ARIMA and ARIMAX models with parameters $p \in \{0, 1, 2, 3, 4\}$, $d \in \{0, 1, 2\}$, $q \in \{0, 1, 2, 3, 4\}$ and exogenous variables that is X_1 the deaths per state. Then, we saved the model with most R2, and a complete table of statistics can be found in the annexed *performances.csv*. A summary of the best models per data split can be found in the following table:

Table 2: Summary of results found in *performances.csv*.

		Best Model	R2	RMSE	MAE
Data Split	**1st Wave**	ARIMAX(4,1,4)+X1	0.9371	14057	1149
	2nd Wave	ARIMAX(4,1,4)+X1	0.8843	91486	6715
	3rd+4th Wave	ARIMA(4,1,4)	0.9409	18195	1681
	All data	ARIMA(4,1,4)	0.9360	212021	2817

We then fit the previously described LSTM model to the dataset and found poor p-values (around 10^{-35}), meaning that the model wasn't fitting properly to the data. Qualitatively, it seemed more like the average of the predictions, rather than the unstable peaks seen in the unprocessed dataset.

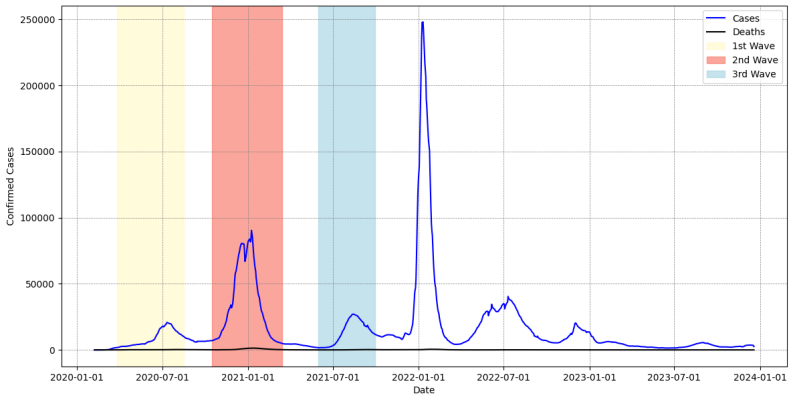

Figure 6: Plot of the average cases in the last 7 days in the United States.

Average Cases Last 7-days Dataset: Similarly, we fit the same models ARMA models to this averaged dataset and found the results in Table 3. In this case, fits were way better as the p-value could always be found over the 0.05 critical value, except for the complete dataset. Specifically, p-values for the 1st Wave, 2nd Wave, 3rd+4th Wave and all data ARIMA(X) models fits were 0.972, 0.968, 0.979, and 1×10^{-18}. This shows that most of the fits are very good for the data they are predicting, allowing for accurate short term forecasts.

Table 3: Summary of results found in *performances.csv*.

		Best Model	R2	RMSE	MAE
Data Split	**1st Wave**	ARIMAX(4,1,4)+X1	0.9988	1764	116
	2nd Wave	ARIMAX(4,1,4)+X1	0.9977	11425	799
	3rd+4th Wave	ARIMA(4,1,4)	0.9987	2305	166
	All data	ARIMA(4,1,2)	0.9987	28183	337

However, the LSTMs performed worse than in the case of the full dataset. This may be due to the LSTM not being complex enough, or not having enough data to work with. None of the residuals passed p-value of 0.05 critical value, denoting that the LSTM had not fitted properly to the series.

Sum of Cases per Week Dataset: Lastly, we summarized the weeks of the data set into 204 weeks, the cases per week that were found were summed up. We applied the same models to this dataset and got similar results to the previous dataset. Concretely, all p-values exceeded the 0.05 critical value with 0.966, 0.504, 0.183 and 1.0 respectively. These also indicate the the models were able to fit to the data properly, except on the 3rd+4th wave that has a lower p-value that is still inside the null hypothesis. When looking at the ACF and PACF, the auto-correlation is indeed IID noise.

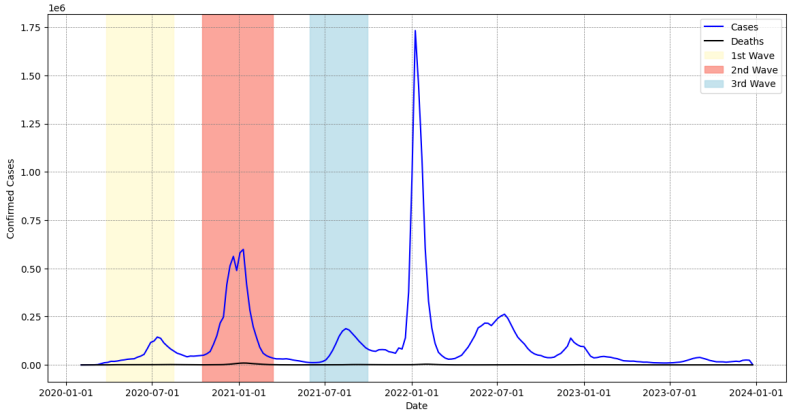

Figure 7: Plot of the cases per week in the United States.

Table 4: Summary of results found in *performances_weekly.csv*.

		Best Model	R2	RMSE	MAE
Data Split	**1st Wave**	ARIMAX(4,1,4)+X1	0.9662	26363	6361
	2nd Wave	ARMA(4,3)+X1	0.9440	150818	35847
	3rd+4th Wave	ARIMA(4,1,3)	0.9938	15132	3819
	All data	ARMA(4,4)+X1	0.8984	664903	22724

Likewise to the previous case, the LSTM wasn't able to fit properly to these cases. The residuals were not IID noise, and the qualitative fit was way off the data line.

Lastly, a qualitative comparison between the predictions of the unprocessed dataset and the average last-7-days dataset was performed. The forecast was set to 14 steps forward to see how useful it would have been on a real-life scenario. We can see the results in figure 8.

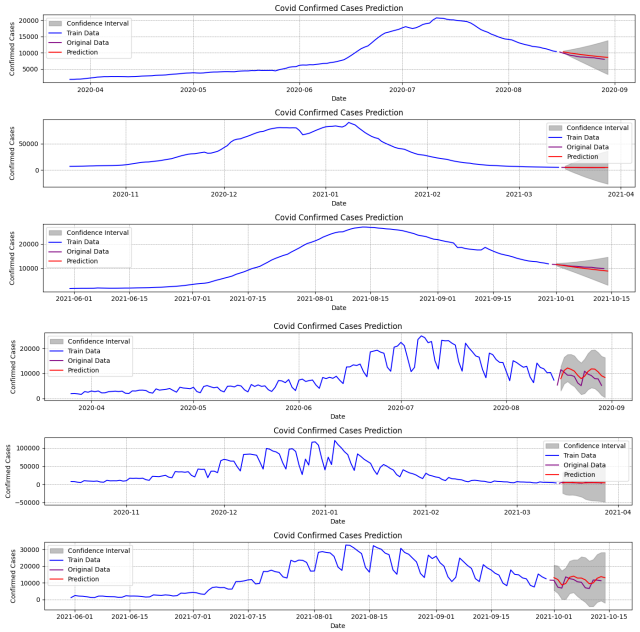

Figure 8: First 3 graphs represent the forecast performed in the last-7-days dataset, 1st, 2nd and 3rd+4th wave. The last 3 graph correspond to the forecast performed in the unprocessed dataset, 1st, 2nd and 3rd+4th wave respectively.

4 Conclusions

The COVID-19 pandemic represented a new paradigm on time series analysis as the need for forecasting new cases was needed. In previous work, ARIMA models have been successfully applied for prediction of the pandemic dynamics. Several studies have applied these ARIMA models in different countries and scenarios. This particular research has focused mainly in the United States complete summary of cases for the insights.

We evaluated the different models available for predictions and found interesting insights on the nature of the data. Specifically, we noticed that smoothing the data via last-7-days averages was particularly useful for getting better models and better adjusted predictions. Nonetheless, the use of these models should be limited to 2-3 weeks at most, since it won't be able to capture further waves. This behaviour is still useful to model the subsequent weeks in a real life scenario, and give a confidence range on where the curve of the pandemic will lie in the future.

On the other hand, the LSTM model didn't prove too useful compared to the traditional methods. Perhaps this is due to lack of data, or not a complex enough network. However, previous studies have shown that LSTM networks can be used for pandemic dynamics predictions.

Lastly, this work could be expanded by using more data and combining more meaningful exogenous values. This would allow for more interesting and accurate predictions, where the model would be able to give further specification on the next few steps of the pandemic.

References

[1] Mohamed R Abonazel and Nesma M Darwish. Forecasting confirmed and recovered covid-19 cases and deaths in egypt after the genetic mutation of the virus: Arima box-jenkins approach. *Commun. Math. Biol. Neurosci.*, 2022:Article–ID, 2022.

[2] Valeria Caramello, Alberto Catalano, Alessandra Macciotta, Lucia Dansero, Carlotta Sacerdote, Giuseppe Costa, Franco Aprà, Aldo Tua, Adriana Boccuzzi, and Fulvio Ricceri. Improvements throughout the three waves of covid-19 pandemic: Results from 4 million inhabitants of north-west italy. *Journal of Clinical Medicine*, 11(15), 2022.

[3] Rohitash Chandra, Ayush Jain, and Divyanshu Chauhan. Deep learning via lstm models for covid-19 infection forecasting in india. *PlosOne*, 17, 01 2022.

[4] Hui Xiang Chua. Covid-19 singapore. Available from `https://data.world/hxchua/covid-19-singapore`.

[5] D. Dickey and Wayne Fuller. Distribution of the estimators for autoregressive time series with a unit root. *JASA. Journal of the American Statistical Association*, 74, 06 1979.

[6] Emrah Gecili, Assem Ziady, and Rhonda D Szczesniak. Forecasting covid-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the usa and italy. *PloS one*, 16(1):e0244173, 2021.

[7] US Government. Covid-19 time-series metrics by county and state. Available from `https://healthdata.gov/State/COVID-19-Time-Series-Metrics-by-County-and-State-A/cr6j-rwfz/about_data`.

[8] Simon H Heisterkamp, Arnold LM Dekkers, and Janneke CM Heijne. Automated detection of infectious disease outbreaks: hierarchical time series models. *Statistics in Medicine*, 25(24):4179–4196, 2006.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Simona et al. Iftimie. First and second waves of coronavirus disease-19: A comparative study in hospitalized patients in reus, spain. *PLOS ONE*, 16(3):1–13, 03 2021.

[11] G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303, 1978.

[12] MD Prof Chaolin Huan, MD Yeming Wang, MD Prof Xingwang Li, PhD Prof Lili Ren, MD Prof Jiaping Zhao, MD Yi Hu, and et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The Lancet*, 395:497–506, 01 2020.

[13] Tanzila Saba, Ibrahim Abunadi, Mirza Naveed Shahzad, and Amjad Rehman Khan. Machine learning techniques to detect and forecast the daily total covid-19 infected and deaths cases under different lockdown types. *Microscopy Research and Technique*, 84(7):1462–1474, 2021.

[14] Christophorus Beneditto Aditya Satrio, William Darmawan, Bellatasya Unrica Nadia, and Novita Hanafiah. Time series analysis and forecasting of coronavirus disease in indonesia using arima model and prophet. *Procedia Computer Science*, 179:524–532, 2021.

[15] Sarbjit Singh, Kulwinder Singh Parmar, Sidhu Jitendra Singh Makkhan, Jatinder Kaur, Shruti Peshoria, and Jatinder Kumar. Study of arima and least square support vector machine (ls-svm) models for the prediction of sars-cov-2 confirmed cases in the most affected countries. *Chaos, Solitons & Fractals*, 139:110086, 2020.

[16] R. Somyanonthanakul, K. Warin, W. Amasiri, and et al. Forecasting covid-19 cases using time series modeling and association rule mining. *BMC Med Res Methodol*, 22(281), 2022.

[17] Kaiming Tao, Philip L Tzou, Janin Nouhin, Ravindra K Gupta, Tulio de Oliveira, Sergei L Kosakovsky Pond, Daniela Fera, and Robert W Shafer. The biological and clinical significance of emerging sars-cov-2 variants. *Nature Reviews Genetics*, 22(12):757–773, 2021.

[18] G Peter Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.