2022 International Conference on Big Data, Information and Computer Network (BDICN) | 978-1-6654-8476-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/BDICN55575.2022.00022

COVID-19 Pandemic Trend Prediction in America Using ARIMA Model

Yunhao Shi[1, *, †], Kailiang Wu[2, †] and Miao Zhang[3, †]
[1]Faculty of Arts, University of Sydney, Sydney, New South Wales, 2006(post), Australia;
[2]School of Electrical Engineering and Electronics, University of Liverpool, Liverpool, L693BX(post), The UK;
[3]School of Optoelectronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing, 400065(post), China
*Corresponding author' e-mail: yshi5771@uni.sydney.edu.au
[†]These authors contributed equally.

Abstract. **COVID-19 trend prediction helps policymakers to handle disease situations. Therefore, it is necessary to predict the pandemic spread trend for prevention and control. The traditional infectious disease model is established according to the transmission characteristics of the disease. However, the trend prediction method of the traditional infectious disease model ignores considering the actual prevention and control situation, resulting in inaccurate models. To address this problem, this paper uses the ARIMA model to predict the spreading trend. First, we download the pandemic data from the website, compare the pandemic situation in different countries and select the United States as the research object. Second, the time series forecasting method is used to analyze the characteristics of the experimental data set. Finally, we use the ARIMA model to analyze the confirmed cases of COVID-19 in the United States and predict the spreading trend. To verify the effectiveness of the ARIMA model, we compare it with the prophet model and random forest model, evaluate the model performance with mean absolute scaled error, symmetric mean absolute percentage error, and root mean squared error. The experimental results illustrate that the ARIMA model significantly outperforms baselines by obtaining the three values of 0.14, 9.97, 22316.57, respectively. The empirical results based on the pandemic spreading prediction in the United States show that the model has good applicability and accuracy.**

Keywords: COVID-19 ,Pandemic Trend Prediction, America ,ARIMA Model

I.Introduction

New Coronavirus pneumonia (COVID-19) has been spreading rapidly worldwide since 2020 [1, 2]. Until now, the pandemic situation in some countries is still unable to be effectively controlled [3]. Therefore, the trend prediction of New Coronavirus pneumonia has become a significant research focus. So reasonable pandemic spreading prediction has essential reference significance for pandemic prevention and control. Pandemic spreading trend prediction predicts the growth of COVID-19 concerning the number of infected individuals, the number of deaths, and the number of recovered cases.

Traditional trend prediction methods mainly include infectious disease prediction models, such as the SIR model [4], SEIR model [5], etc. Huppert et al. [4] introduced the classical SIR model and its many assumptions that are not realistic, including well-mixed population, homogeneity of the population, exponentially distributed duration of infection, and large population. Piccirillo [5] introduced that the SEIR model with a restriction parameter was used to explore the dynamic of the COVID-19 pandemic. However, the traditional trend prediction model ignores considering the actual situation, and the establishment of the model is imperfect, leading to inaccurate results. Many studies based on deep learning have attempted to predict COVID-19 trends to improve the prediction accuracy[6]. Devaraj et al.[7] performed a comparative analysis on the prediction models such as ARIMA, LSTM, Stacked LSTM, and Prophet approaches, and used multivariate LSTM models to forecast long-term COVID-19 cases. Shastri et al.[8] proposed deep learning based comparative analysis of COVID-19 cases in India and the USA. Convolution LSTM was designed to predict the COVID-19 cases with high accuracy and minor error for all four datasets of both countries. Mohammed et al.[9] implemented six different deep learning approaches on time series to compare the values associated in datasets and predict various affected aspects in the near future. These models effectively assist medical experts and scientific research institutions in the efficient prediction of COVID-19. However, prediction models have timeliness. As influence factors change, the effectiveness of the model has reduced.

In this paper, we use the ARIMA model to forecast the spreading trend of SARS-CoV-2 in the United States. First, we download the pandemic data from the American PCR testing website[10] and compare the pandemic situation in different countries. Among these countries, America tends to show a large proportion of COVID-19 cases, so that we select the United States as the research object. Second, we use cross-validation to scientifically classify the data, making the model results more accurate. Then, the time series forecasting method is used to analyze the characteristics of the experimental data set. We use the ARIMA model to analyze the data of the confirmed cases of COVID-19 in the United States and predict the spreading trend. We compare it with two baseline models to verify the effectiveness, the prophet model and the random forest model. Then, we calculate the mean absolute scaled error (MASE), symmetric mean absolute percentage error (SMAPE), and root mean square error (RMSE) of each model to evaluate the model performance. The MASE and RMSE values of the prophet model are much higher than the ARIMA model, so we focus on the other two models in the following work. The experimental results show that ARIMA achieves 0.14, 9.97, and 22316.57 on MASE, SMAPE, and RMSE, while the random

978-1-6654-8476-3/22/$31.00 ©2022 IEEE
DOI 10.1109/BDICN55575.2022.00022

forest separately achieves 0.51, 41.77, and 71579.50, respectively. As a result, the performance of the ARIMA model is better than the other two baseline models. The results show that the ARIMA model has good applicability and can accurately predict the pandemic spreading trend.

II.Method

This section describes the proposed method. Our method first preprocesses the data (Sec. 2.1). Then we construct the ARIMA model (Sec. 2.2).

A.Data Preprocessing

The data set of COVID-19 has been downloaded from the American PCR testing

and accessible on Aug

series and virus cases number in each state and union territory. Then, we analyze the macro situation of the pandemic and choose the United States as the research goal (Sec. 2.1.1).

1)Data analysis

Different countries' situations can be compared simultaneously to highlight the country which is poorly influenced. Figure 1 shows confirmed cases in different countries, collected from all the countries with new reported covid cases based on the data acquired in 2021. Top countries such as the United States, India, Brazil, and Turkey are highlighted with deep blue. These countries' death number is shown in figure 2.

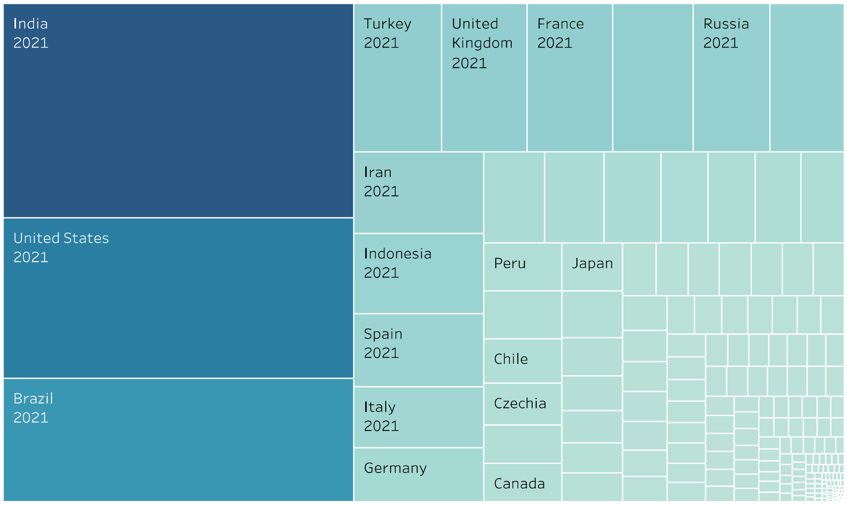


Figure 1. Confirmed Cases in Different Countries.

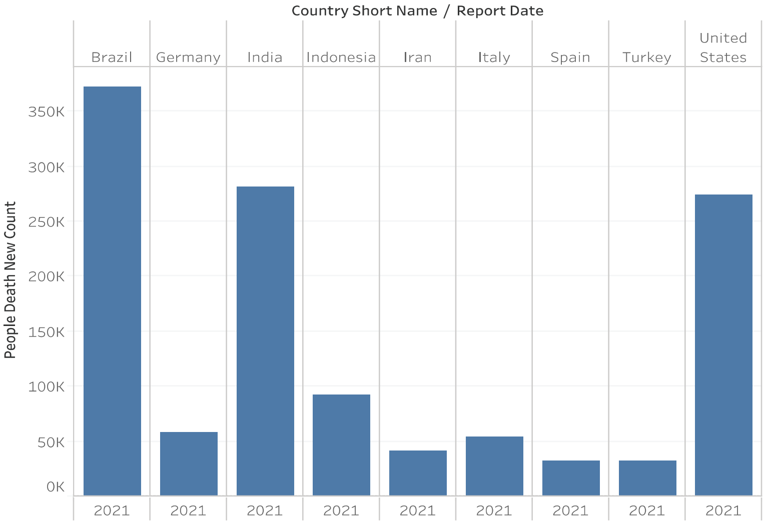


Figure 2. Death Number in Top Countries.

Figure 3 shows the new reported cases number in provinces around the world, which furtherly reflects that many provinces with a serious covid issue are in the United States. Considering that America has apparent death numbers and new cases pattern, it is chosen as the analyzed target highly impacted by COVID-19.

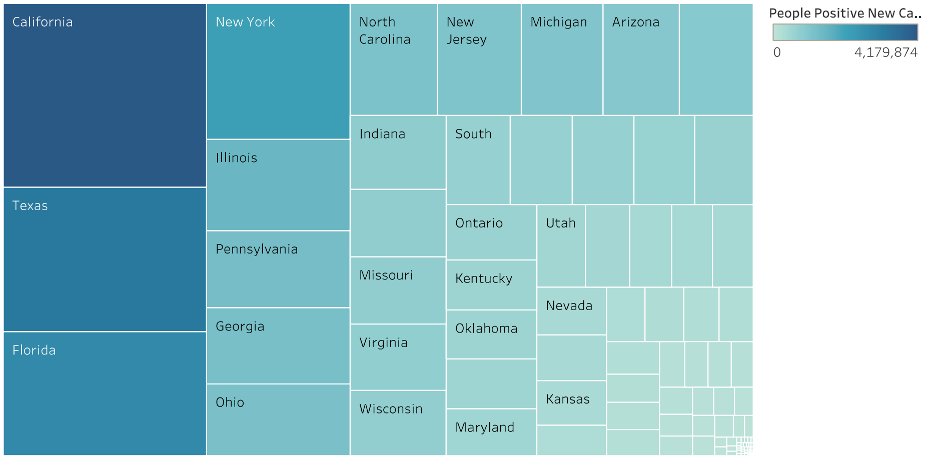


Figure 3. Confirmed Cases in Different Provinces.

B. ARIMA Model Construction

Autoregressive Integrated Moving Average (ARIMA)[12] can predict and analyze time series. In $ARIMA(p,d,q)$, AR stands for “autoregressive”, and p shows how many numbers of autoregressive terms are in this model; MA is the “moving average”, q means the total number of moving average terms, while d illustrates the order which could help to keep sequence stationery. The difference is the main step in this equation, although it does not appear in its name. The equation for $ARIMA(p,d,q)$ is shown below:

$$(1-\sum_{i=1}^{p}\varphi_i L^i)\,(1-L)^d X_t \;=\; (1+\sum_{j=1}^{q}\theta_j L^j)\,\varepsilon_t \qquad (1)$$

In this equation, L can be seen as a lag operator, which can be represented as $(1-L)X_t = X_{t-1}$. Additionally, p and q are the orders of this model, representing the end of the number sequence. For sequence i, it has a range between 1 and p. Similarly, sequence j exists in the range between 1 and q. The stationary of this series is decided by d, which is the degree of the ordinary differencing[13].

Before using the ARIMA model, the stability of the time series should be first judged. If it is unstable, the Stabilization treatment is required. The stationarity of time series is generally tested by the time series diagram and autocorrelation diagram. The characteristic of the sequence diagram is intuitive and straightforward, but the error is large. The auto-correlation diagram, including autocorrelation and partial autocorrelation function diagram, is relatively complex, but the result is more accurate.

III.Experimental Settings

This section describes the experimental settings in our experiment. First, we describe the dataset used in our experiment (Sec. 3.1). Next, we use cross-validation to scientifically classify the data, making the model results more accurate (Sec. 3.2). Then, we introduce the baselines in the experiment (Sec. 3.3). Finally, we use three Evaluation Metrics to evaluate our model (Sec. 3.4).

A. Dataset

Due to the significance of analyzing covid situation in America, the paper imports American data from website[14] to complete the time series forecasting. The Time series forecasting combine various observations from the previous study through a random variable model. We analyze the underlying relationship between data collected and predict the future values through patterns such as downward, increasing, or fluctuating curve trends.

B.Cross-Validation

Cross-validation[11] is a statistical analysis method used to verify the performance of classifiers. The basic idea is to group the original data sets falling in a certain sense, one part is used as the training set, and the other part is used as the validation set. The training set is used to train the classifier. Then, the verification set is used to verify the model, and the final classification accuracy is recorded as the performance index of the classifier.

C.Baselines

To verify the effectiveness and accuracy of the ARIMA model, it is compared with the prophet model and random forecast model.

1) Prophet Model

Prophet model[15] is an open-source data prediction tool of Facebook based on Python and R language. It is a program used to predict the data in time series. The additive model as its core function, in which nonlinear trend is used to predict the seasonality of annual, week, and day including the holiday effect. During testing, this model was proved to have a remarkable ability to forecast the output in the timeline. It fits for predicting the time series data, which has a strong seasonal effect. Prophet is very sensitive to the difference of data and trends, which results in its strong ability to deal with outliers.

The model requires its input to have two variables: ds and y. The format of ds used in this program is related to the time series marked with $YYYY - MM - DD$. Another factor is y which is the forecast output. The paper mainly inputs collected data and calls the method package after the output can then be plotted.

2) Random Forest

Random forest[16] comes from the basic idea in machine learning, the decision tree. The random forecast can be seen as a group of decision trees that uses the average to low down the variance of the output. Training the random forecast model involves the bagging algorithm. In determining data set X and target Y, the bagging algorithm can get the forecast of X, which is listed below:

$$f = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \qquad (2)$$

Therefore, the random forecast model can get many f from many decision trees. This output will then be used to get the final forecast result. Additionally, to get a higher accuracy answer, random forecast uses the random subsets of features which will help to low down the strong correlation between different decision trees.

D.Evaluation Metrics

We use mean absolute scaled error (MASE), Symmetric mean absolute percentage error (SMAPE), and root mean square error (RMSE) to evaluate the model performance.

1)MASE

The mean absolute scaled error (MASE)[17] is a standard to judge the accuracy of forecasting. Its value comes from the division of two results. The molecular is the mean absolute error of one forecast value, while the denominator is the mean absolute error of the sum from 2 to T, adding with the naive forecast. When the result equals 1, this illustrates that the model can be seen as a perfect example to trust its forecast accuracy. If the result equals 0.5, this means the accuracy of model prediction is doubled. The lower the value is being calculated, the more the model can be trusted. If the result is larger than 1, this still needs to be significantly improved before the next prediction.

$$MASE = mean\left(\frac{|e_j|}{\frac{1}{T-1}\sum_{t=2}^{T}|Y_t - Y_{t-1}|}\right) \qquad (3)$$

In this equation, j represents the number of forecasting times, and the numerator e_j is the predicted error in that period. Its value is calculated from the actual value, Y_j minus the forecast value, which is F_j. Therefore, it can then be represented in the equation: $e_j = Y_j - F_j$.

The naive forecast method can be understood as the previous value minus the current. In the equation, the domain is a range from 2 to T, which equation is: $Y_t - Y_{t-1}$. Hence the denominator is the mean absolute error for this predict method.

2)SMAPE

Symmetric mean absolute percentage error (SMAPE)[18] is an accurate measurement. Its basic calculation theory is to calculate the percentage errors. It can usually be defined as equation shown follows:

$$SMAPE = \frac{100\%}{n}\sum_{t}^{n}\frac{|F_t - A_t|}{(|F_t| + |A_t|)/2} \qquad (4)$$

In this equation, n indicates sample size, A_t stands for the actual value from the collection, and F_t represents the prediction value. The lower the SMAPE value of a forecast, the higher its accuracy.

3) RMSE

Root mean square error (RMSE)[19] is a frequently used measurement of prediction. It can calculate the differences between the predicted values through one model and the actual values collected through observation. We compare the predicted results with the actual values and calculate the RMSE of each model. The lower the RMSE value of a forecast, the higher its accuracy. RMSE can be calculated by the following formula:

$$RMSE(x, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2} \qquad (5)$$

IV.Results and Discussion

This section gives detailed experiments to study ARIMA performance compared with baseline models (Sec. 4.1). Then, we discuss the improvement direction of the model (Sec. 4.2).

A.Overall Performance

Firstly, we analyze and compare the performance of the three models with three parameters, RMSE MASE, and SMAPE. Then, we look at the prediction results of prophet. However, due to the unsatisfactory performance of prophet, we choose the random forest model for progressive comparison. Table 1 shows the detailed comparison results of related methods. From Table 1, we can observe the following results.

RMSE Performance. The number for the prophet model is 1760138.05, which is a vast number and creates doubt about the result. Random forest and Arima models are then imported to make a comparison. By importing the accuracy table, the RMSE number for ARIMA and random forest model is 22316.57 and 71579.50, respectively. Both are lower than the prophet model, and the ARIMA model tends to perform the best.

MASE Performance. We find that the ARIMA model's value is 0.14, lower than 0.51 for the random forest model and 1249.80 for the prophet model. That means the ARIMA model gives the most accurate forecast result, the random forest model has the doubted accuracy, and the prophet model needs improvement.

SMAPE Performance. Surprisingly, the prophet SMAPE value is 1.99, lower than ARIMA (9.97) and random forest model (41.77). However, considering the ARIMA model's previous performance and a correspondingly low SMAPE value, it is still considered the preferred model in the paper.

Table 1. The performances of ARIMA and baseline models.

Model	MASE	SMAPE	RMSE
ARIMA	**0.14**	9.97	**22316.57**
Random Forest	0.51	41.77	71579.50
Prophet	1249.81	1.99	1760138.05

For further comparison, we visualize the output of the prophet model in figure 4. We can see that three lines are formed similarly in the table. The X label of the table is ds which is time, and the Y label is total infected people in the US. The thin blue line in the middle is the ideal forecast line. However, there will be errors and differences from the ideal result. Therefore, the coarse blue line represents the possible lower forecasting result, and the black line represents the possible higher one. The actual value should be in the range of these two lines and close to the ideal one.

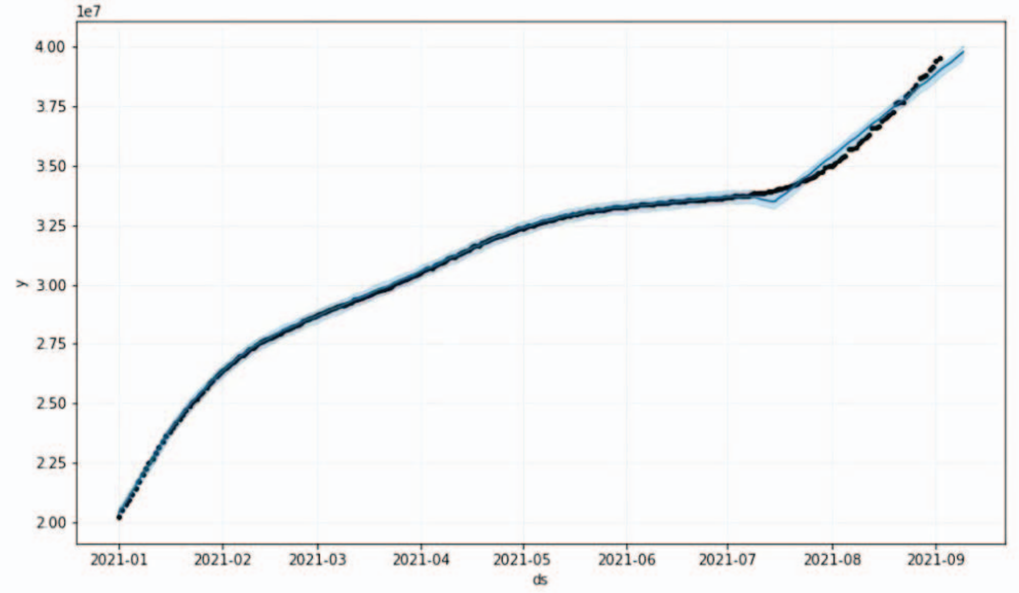


Figure 4. The Fitting Output of Prophet Model.

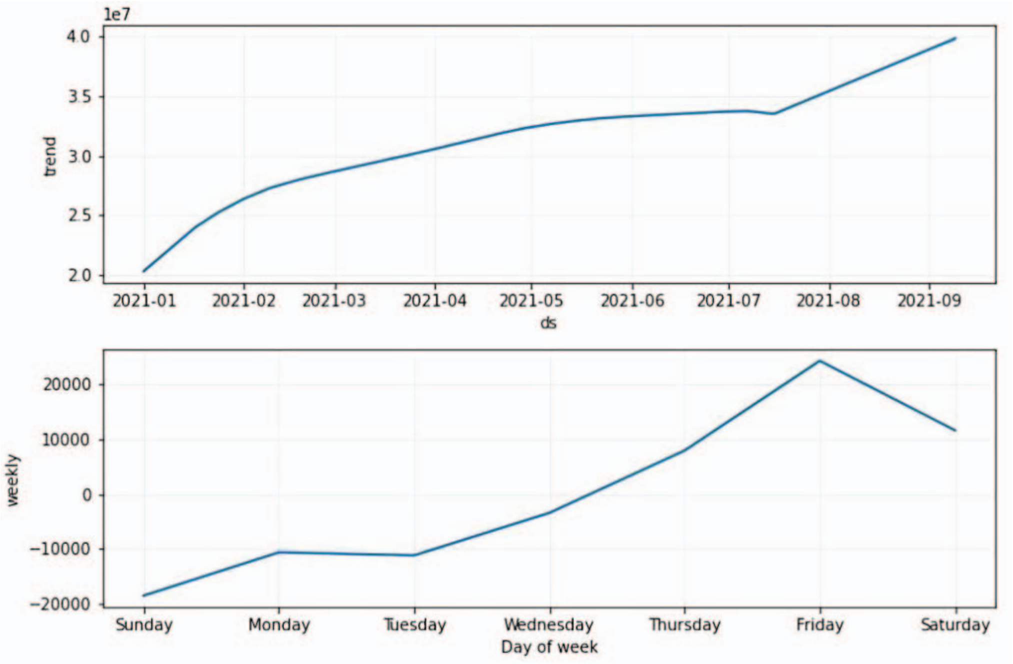


Figure 5. The Spread Output of Prophet Model.

Figure 5 indicates how will the COVID-19 cases increase in the coming months. It is almost the same as the ideal line in figure 4. The picture at the bottom of figure 5 shows a forecast of the covid-19 weekly trend and shows a peak value on Friday.

Since the MASE and RMSE values used to evaluate the model performance of prophet are not ideal, we choose the random forest model with better performance as a comparison with ARIMA. Figure 6-8 give the Times Series results, cross-validation and forecast, respectively. It can be seen from figure 6-8 find the declining trend between March to July is more conspicuous than the prophet model. The diagram suggests that covid-19 cases are controlled during that period. However, the cases begin to increase again after July. The Cross-Validation method is used to verify this phenomenon, which is shown in figure 7. It starts with a group of data for training purpose, and then forecast the later data points in another data group called validation set. The 245 instances from data, 240 of them are considered the training set, and leftover five are considered validation set by keeping seven days of windows space for evaluating the predictive model's performance.

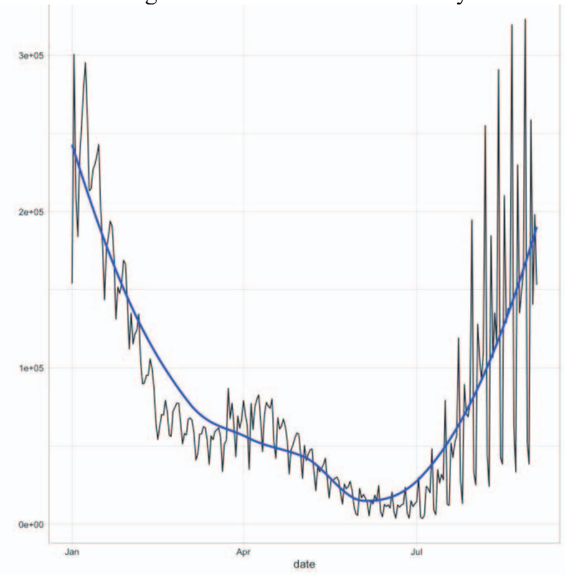


Figure 6. Time Series.

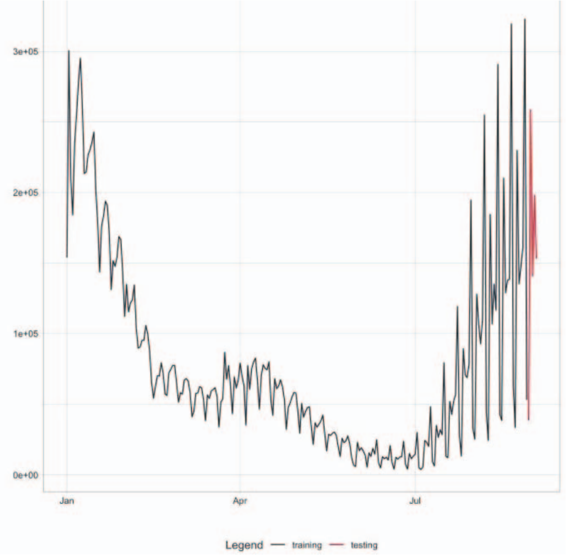


Figure 7. Cross-Validation.

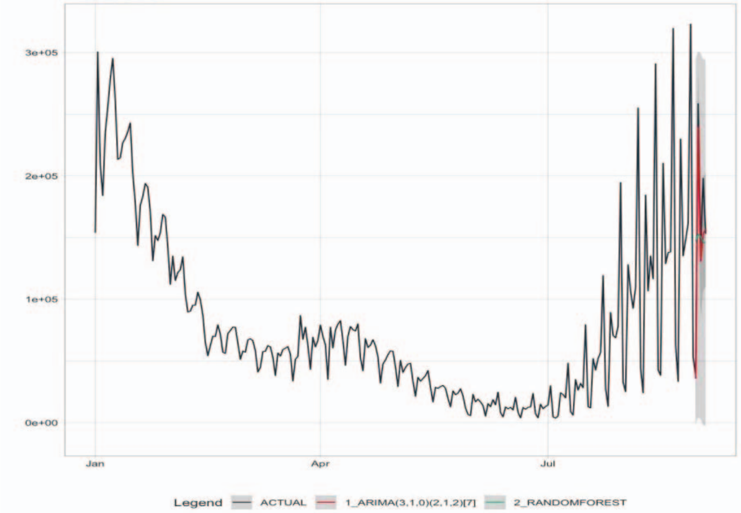
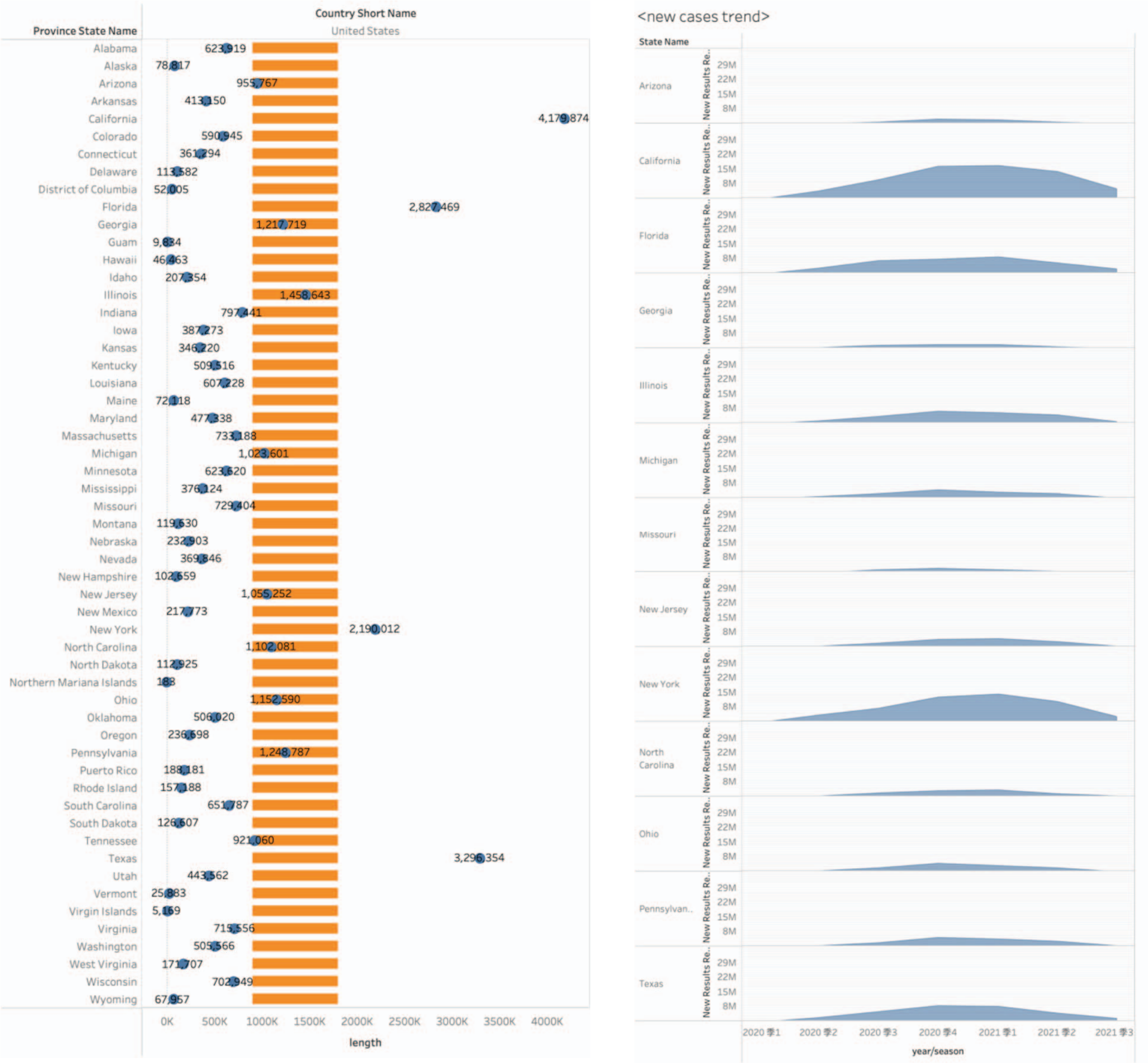


Figure 8. Forecast Plot.

After fitting and cross-validation, random forest and ARIMA models are implemented. Observing the forecast in figure 8 shows that the ARIMA model tends to forecast more precisely than the random forest model.

B.Further Discussion

Data surveys on different provinces should help locate a severely influenced area to control the epidemic situation further.



(a)

(b)

Figure 9. Provinces Extremely Influenced by COVID-19. (a) is the Gantt chart, in which the baseline number of total severe cases is set to 900000. (b) is several provinces' new Covid-19 cases' data from 2020 to 2021 in seasons.)

Figure 9 (a) shows provinces with blue points lying in the orange area, which are the severe area observed. While those blue points lie on the right-hand side of the orange area, they represent provinces that Covid-19 immensely influences. Figure 9 (b) shows that new cases in most areas have declined since 2021, which means the situation is eased. However, the Covid-19 new case number is still huge. To make a further cases forecasting, factors like facial mask-wearing measures[20], social distance restriction order's time duration[21], and geographic characteristics[22] for different provinces should be considered as new variables.

V.Conclusion

In this paper, we build an ARIMA prediction model to predict the pandemic spread trend in America. We also build baselines with random forest model and prophet model as comparative models. By comparing each model's MASE, SMAPE, and RMSE value, we prove that the ARIMA model is more suitable to forecast pandemic trends than the other two models. We further explore the disease situation in different regions, and severely influenced areas are visualized. Our model can be applied to these areas to control the spread of the pandemic better.

In the future, we will focus on further improving these models by adding more variables. To make a more accurate forecasting, facial mask wearing measures, social distance restriction order's time duration, geographic characteristics, and other factors should be considered new variables. Also, more state-of-the-art deep learning methods will be tested on this dataset.

References

[1] World Health Organization. (2020) Naming the coronavirus disease (COVID-19) and the virus that causes it. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(COVID-2019)-and-the-virus-that-causes-it

[2] World Health Organization. (2021) Coronavirus Disease (COVID-19) Situation Reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[3] World Health Organization. (2021) WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. https://COVID19.who.int/

[4] A. Huppert, G. Katriel. (2013) Mathematical modeling and prediction in infectious disease epidemiology. Clinical Microbiology and Infection, 19: 999-1005.

[5] Piccirillo, V. (2021) Nonlinear control of infection spread based on a deterministic SEIR model. Chaos, Solitons & Fractals, 149: 111051.

[6] Muzammil Khan, Muhammad Taqi Mehran, Zeeshan Ul Haq, Zahid Ullah, Salman Raza Naqvi, Mehreen Ihsan, Haider Abbass, (2021) Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review. Expert Systems with Applications, 185: 115695.

[7] Jayanthi Devaraj, Rajvikram Madurai Elavarasan, Rishi Pugazhendhi, G.M. Shafiullah, Sumathi Ganesan, Ajay Kaarthic Jeysree, Irfan Ahmad Khan, Eklas Hossain. (2021) Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?. Results in Physics, 21: 103817.

[8] Sourabh Shastri, Kuljeet Singh, Sachin Kumar, Paramjit Kour, Vibhakar Mansotra. (2020) Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study. Chaos, Solitons & Fractals, 140: 110227.

[9] Shaik M A, Verma D. (2020) Deep learning time series to forecast COVID-19 active cases in INDIA: a comparative study. IOP Conference Series: Materials Science and Engineering, 981(2): 022041.

[10] COVID-19 Diagnostic Laboratory Testing (PCR Testing) Time Series | HealthData.gov. https://healthdata.gov/dataset/COVID-19-Diagnostic-Laboratory-Testing-PCR-Testing/j8mb-icvb

[11] Stone, M. (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society, Series B (Methodological), 36 (2): 111–147.

[12] Sun, J. (2021) Forecasting COVID-19 pandemic in Alberta, Canada using modified ARIMA models. Computer Methods and Programs in Biomedicine Update, 100029.

[13] G.Peter Zhang. (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50: 159-175.

[14] Our World in Data. (2021) Coronavirus (COVID-19) Vaccinations - Statistics and Research. https://ourworldindata.org/covid-vaccinations?country=DEU

[15] Christophorus Beneditto Aditya Satrio, William Darmawan, Bellatasya Unrica Nadia, Novita Hanafiah. (2021) Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. Procedia Computer Science, 179: 524-532.

[16] Yeşilkanat, C.M. (2020) Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos, Solitons & Fractals, 140: 110210.

[17] Franses, Philip Hans. (2016) A note on the Mean Absolute Scaled Error. International Journal of Forecasting. 32 (1): 20–22.

[18] Tofallis, C. (2015) A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation. Journal of the Operational Research Society, 66(8): 1352-1362.

[19] Rob J. Hyndman, Anne B. Koehler. (2006) Another look at measures of forecast accuracy. International Journal of Forecasting. 22 (4): 679–688.

[20] John. (2020) Facial mask: A necessity to beat COVID-19. Build Environ. 175: 106827.

[21] Brown, R., Cowling, M. (2021) The geographical impact of the Covid-19 crisis on precautionary savings, firm survival and jobs: Evidence from the United Kingdom's 100 largest towns and cities. International Small Business Journal: Researching Entrepreneurship, 39: 319-329.

[22] Wellenius, G.A., Vispute, S., Espinosa, V. et al. (2021) Impacts of social distancing policies on mobility and COVID-19 case growth in the US. Nat Commun, 12: 3118.