Multiple Models for Forecasting
Case Rate and Death Rate of COVID-19

By

Rui Li

Senior Honors Thesis
Department of Statistics and
Operations Research
University of North Carolina at Chapel Hill

April 20, 2021

Approved:

Mario Giacomazzo, Thesis
Advisor

Yao Li, Reader

Abstract

Within the last two decades, coronaviruses have generated devastating effects on humans being. Especially in 2020, the appearance of new coronaviruses, COVID-19, makes researchers aware of the importance of forecasting and predicting the COIVD-19 spreading patterns. This research utilizes four different time series models - Naïve model, ARIMA model, Horizon-Specific Model without Difference, and Horizon-Specific Model with Difference – to predict the case and death rate with a horizon time of 1 day, 3 days, and 5 days ahead for 173 different countries. We identify the best model for each country based on the minimization of out-of-sample root mean squared error (RMSE). The results show that when forecasting the case rate, ARIMA models are the best fit for around 54% of countries with a horizon time of 1 day ahead, while the horizon-specific models without difference are the most appropriate for about 50% and 49% of countries with a horizon time of 3 days and 5 days ahead, respectively. When forecasting the death rate, the ARIMA models significantly outperform the others and are suitable for 59%, 39%, and 40% of countries with a horizon of 1 day, 3 days, and 5 days ahead, respectively.

1. Introduction

Coronaviruses are a large family of viruses that infects humans and leads to an upper respiratory infection. There have been two major outbreaks of coronaviruses for the last two decades, resulting in severe diseases and side effects after recovery [1-2]. SARS, severe acute respiratory syndrome, occurred in Southern China from 2002 to 2003. The disease rapidly spread from Hong Kong to most Asian countries, ultimately causing 8422 cases with 916 deaths, a case fatality rate of 11% [3]. Later in 2012, MERS-CoV, Middle East respiratory syndrome coronavirus had been identified in the Middle East, Africa, and South Asia, resulting in 858 known deaths due to the infection and related complications [4]. In 2020, a novel coronavirus named COVID-19 ravaged the world. The first case of COVID-19 was reported on December 27, 2019, in Wuhan, China, and was recognized as a pandemic in March 2020 by the World Health Organization [5]. COVID-19 has a devastating effect on human health and has caused 1.47 million deaths till November 30, 2020. Compared to SARS and MERS-CoV, COVID-19 has brought an unprecedented disaster to the globe. Therefore, it is essential to understand COVID19's trending.

Time series analysis is very effective for data gathered and indexed by time, especially for epidemiological problems like the West Nile virus (WNV), infections' SARS rate, and MERS's outbreak prediction [6]. With global researchers having employed different time series models to forecast the COVID-19's trend of the world [7], we are interested in setting up a data algorithm that can forecast future case and death rates with different models for each country. This project predicts the COVID-19's case and death rate of 173 countries, with time horizons ahead of 1 day, 3 days, and 5 days by utilizing four different time series models. We find the best model for each country by comparing their out-of-sample RMSE (Root Means Square Error). We aim to set up a system that could help each country find its own COVID-19's spreading pattern.

2. Data

a. Data Source

This research mainly utilizes three different datasets – *jhu_full_data*, *jhu_population* from the GitHub of Our World in Data [8], and *Land_Area* from the World Bank [9]. Thanks to Our World in Data, an online scientific publication focusing on global disease problems [10] and the World Bank, we can research with these valuable resources. The *jhu_full_data* records the COVID-19 infection situation of 199 different countries in the world. For each country, the dataset gathers the information of COVID-19's new cases and deaths number, total cases and deaths number, and weekly cases and deaths number, from the first day of infection to February 21, 2021. *Figure 2.1* and *Figure 2.2* are the visualizations of the global COVID-19's distribution.

Figure 2.1 World COVID-19 Cases Number

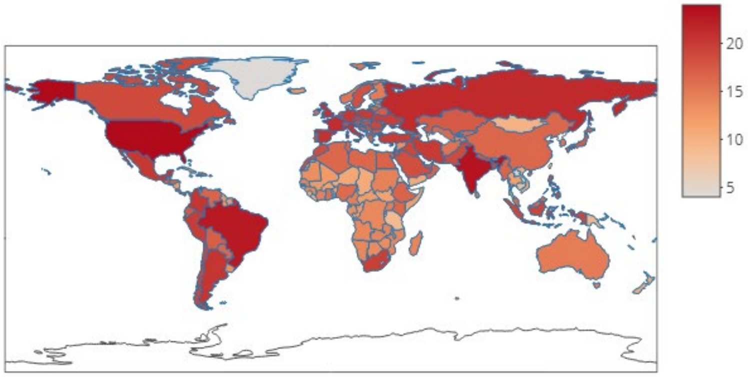


Figure 2.2 World COVID-19 Deaths Number

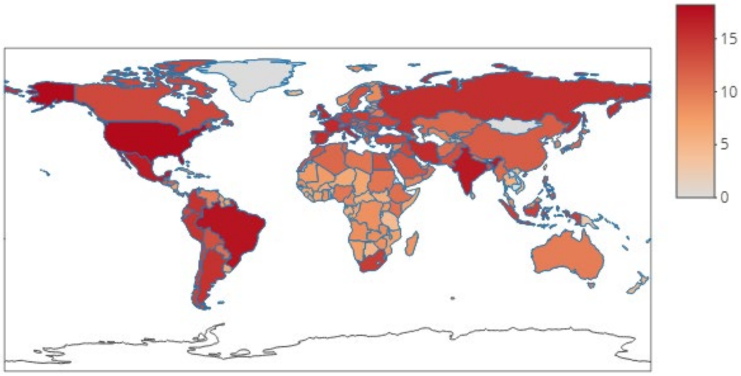


Figure 2.1 and *Figure 2.2* have very similar distributions, which record the total cases number (or deaths number) until February 21, 2021. We can notice that the Americas and tropical regions have more severe epidemics, while Asian countries have better control over the disease.

The *jhu_population* gathers the population data of 199 countries in early 2020, while *Land_Area* describes the land area (square miles) of 201 countries in 2018. *Figure 2.3* and *Figure 2.4* present an approximation of the world population and the world population density in 2020.

Figure 2.3 World Population in 2020

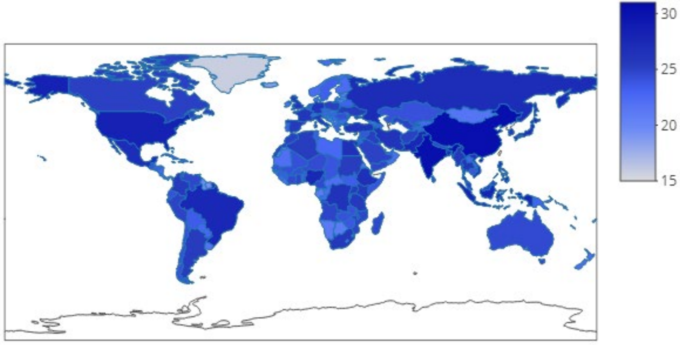
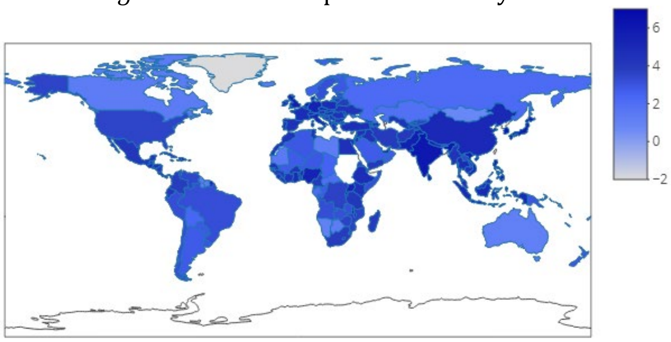


Figure 2.4 World Population Density in 2020



b. Data Cleaning

After summarizing the *jhu_full_data*, we discover 26 small countries and regions that miss daily new cases and new death numbers. Therefore, we remove those observations from the data and select the necessary variables: Date, Country, New cases, and New deaths. After that, we combine the *jhu_full_data* with *jhu_population* by the name of the country and create two variable $case_{rate}$ and $death_{rate}$.

$$case_{rate} = \frac{cases_{new}}{population} * 100000,\ death_{rate} = \frac{deaths_{new}}{population} * 100000 \quad (2.1)$$

As shown by equation (2.1), $case_{rate}$ is equal to the number of new cases per one hundred thousand persons, while $deaths_{rate}$ is equal to the number of new deaths per one hundred thousand persons. *Figure 2.5* and *Figure 2.6* respectively describe the case rate and death rate of the world.

Figure 2.5 World COVID-19 Case Rate

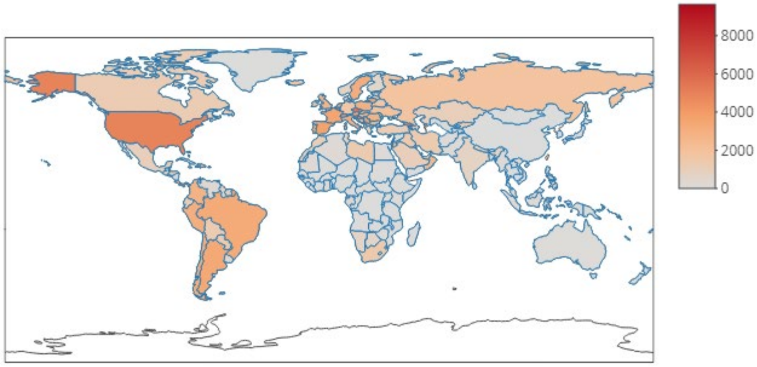
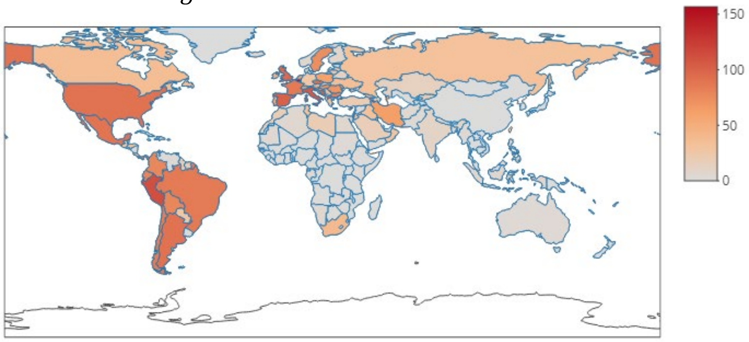


Figure 2.6 World COVID-19 Death Rate



Compared to *Figure 2.1-2*, *Figure 2.5-6* has a significant difference between rates value. While China and Russia have large population densities but low COVID-19's case and death rate, countries like the US and Brazil have both high value in population density and COVID-19's statistic data. The difference between countries inspires us to cluster different countries into various types and conduct specific forecasts.

c. Model Preparation

We split the dataset into training (the year 2020) and testing (the year 2021) data to conduct time series analysis. As the dataset gathers daily information, we are able to visualize the changes in case rate and death rate by day, and we define the forecasting variables with the following symbols:

$$C_t = case_{rate}\ at\ time\ t, \qquad D_t = death_{rate}\ at\ time\ t$$

Figure 2.7-10 shows four typical patterns of COVID-19 spreading. As the first country to report the COVID-19's infection, China's COVID-19 severity reached the peak value in March and quickly under control by April 2020, as shown in *Figure 2.7*. Brazil (*Figure 2.8*) represents the countries that have the constant but severe situation all the time, while Canada (Figure 2.9) and France (Figure 2.10) are typical for most counties in the world, whose have two peak values, one around April 2020, and the other around the winter of 2021.

Figure 2.7 COVID-19 Spreading Curve of China

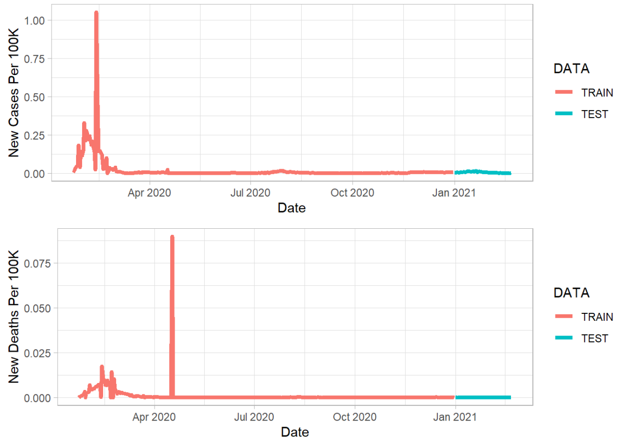


Figure 2.8 COVID-19 Spreading Curve of Brazil

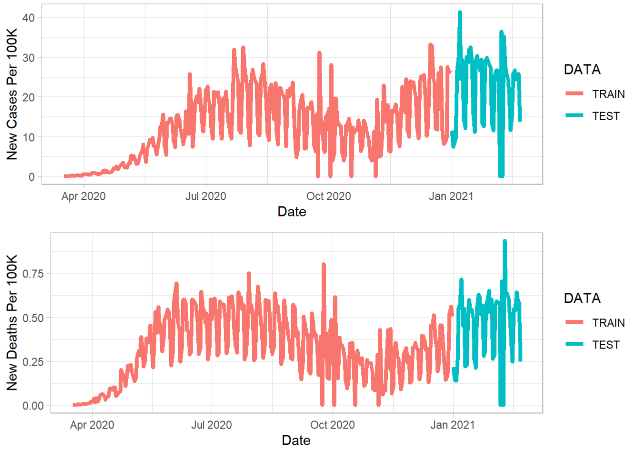


Figure 2.9 COVID-19 Spreading Curve of Canada

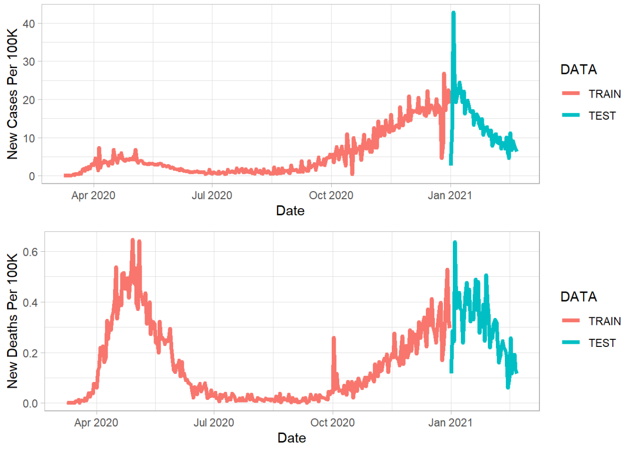
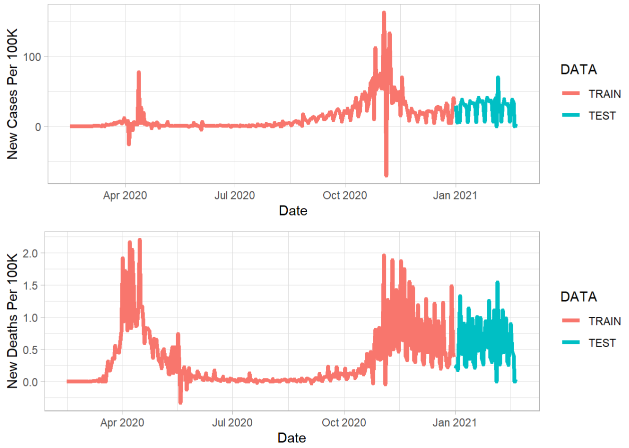


Figure 2.10 COVID-19 Spreading Curve of France



The red curve visualizes the training data, and the green curve represents the testing data. We use different models to analyze each country's pattern and determine the best model by comparing the closeness between the test data's curve and the other forecasting models' curve.

3. Methodology

We build four different time series models to forecast each country's case rate and death rate at a horizon time of 1 day, 3 days and 5 days. For each model and each horizon, we calculate the RMSE during the testing period, figuring out the best model for each country. We store the models' coefficients in a dataset, which helps us set up data algorithms for prediction in the future. Details of the four time series models that we used are found below.

a. Naïve Model

Economists first suggested the Naïve Model as benchmarks of forecasting accuracy in the 1940s. Since the model simply predicts the next period's value by setting it to be that of the preceding value, it is considered the most straightforward time series model, and any other forecasting models that failed to perform better than the Naïve model should be disqualified [11]. A naïve forecast is optimal when data follow a random walk. In this project, the Naïve method forecasts the case and death rate of COVID-19 to be the last observation value [12].

$$C_{t+h} = C_t + \varepsilon_t, where\ h = 1, 3, 5 \quad (3.1)$$

$$D_{t+h} = D_t + \varepsilon_t, where\ h = 1, 3, 5 \quad (3.2)$$

Equations (3.1) and (3.2) show the relationship between case rate (and death rate) at time t and the predicting values of 1 day, 3 days, and 5 days ahead. We estimate the white noise ε_t by calculating the Root Mean Square Root (RMSE) between the current and predicted value. Then, we run a loop through 173 countries and store the RMSE values of case rate and death rate in a larger table for results analysis.

b. ARIMA Model

The ARIMA(p, d, q) Model (Autoregressive Integrated Moving Average) is a widely used model for time series forecasting and aims to describe the autocorrelations in the data [7]. It combines the Autoregressive model AR(p) and Moving average model MA(q). The entire models of COVID-19's cases and death rate prediction can be written as

$$C'_t = c + \emptyset_1 C'_{t-1} + \cdots + \emptyset_p C'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (3.3)$$

$$D'_t = c + \emptyset_1 D'_{t-1} + \cdots + \emptyset_p D'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (3.4)$$

Where p = order of the autoregressive, d = degree of first differencing involved, q = order of the moving average part, ε_t is the white noise. Specifically,

$$C'_t = (1-L)^d C_t, \qquad LC_t = C_{t-1}$$

$$D'_t = (1-L)^d D_t, \qquad LD_t = D_{t-1}$$

Here "L" is the backward shift operator to move value one day backward. We build the ARIMA model for each country with *auto.arima()* function in R, and save the p, d, and q value for the best classical model, chosen using stepwise selection and AIC, for later analysis. Instead of building separate models for each horizon time, we employ rolling forecasting to set up ARMIA models.

c. Horizon-Specific without Differencing

The horizon-specific model without differencing is to conduct time series regression models over forecast variables. To stabilize the variance of a time series C_t and D_t, we make a transformation on the forecast variables and predictors by taking the values of the logarithms. Equations (3.5) and (3.6) are the models of forecasting COVID-19's case and death rate of 1 day, 3 days, and 5 days ahead.

$$Log(C_{t+h} + 1) = \beta_0 + \beta_1 Log(C_t + 1) + \beta_2 Log(C_{t-1} + 1) + \cdots + \beta_5 Log(C_{t-5} + 1) \quad (3.5)$$

$$Log(D_{t+h} + 1) = \beta_0 + \beta_1 Log(D_t + 1) + \beta_2 Log(D_{t-1} + 1) + \cdots + \beta_5 Log(D_{t-5} + 1) \quad (3.6)$$

Where h = 1, 3, 5. Since the case and death rate can be zero, we add 1 to the original value before doing the transformations. We conduct linear regression to forecast future rates by utilizing the rates from the past five days. We loop the models for 173 counties and save the coefficients. Besides, we transfer the predicted value back to raw data by taking exponential transformations,

$$\hat{C}_{t+h} = e^{Log(\widehat{C_{t+h}}+1)} - 1, \qquad \hat{D}_{t+h} = e^{Log(\widehat{D_{t+h}}+1)} - 1$$

and calculate the RMSE for each country.

d. Horizon-Specific with Differencing

The horizon-specific with differencing model computes the difference between consecutive observations, stabilizing the mean of a time series by removing changes in the level of a time series, and therefore eliminating trend and seasonality [7]. Equations (3.7) and (3.8) are the models of forecasting COVID-19's cases and death rate of 1 day, 3 days, and 5 days ahead.

$$Log(C_{t+h} - C_t + 1) = \beta_0 + \beta_1[Log(C_t - C_{t-1} + 1)] + \cdots + \beta_5[Log(C_{t-5} - C_{t-6} + 1)] \quad (3.7)$$

$$Log(D_{t+h} - D_t + 1) = \beta_0 + \beta_1[Log(D_t - D_{t-1} + 1)] + \cdots + \beta_5[Log(D_{t-5} - D_{t-6} + 1)] \quad (3.8)$$

Where h = 1, 3, 5. After we conduct basic linear regression on each country's models and save the coefficients to the results table, we transfer the forecast variables back to calculate the RMSE. The transformations are

$$\hat{C}_{t+h} = e^{Log(\widehat{C_{t+h}-C_t+1})} - 1 + C_t, \qquad \hat{D}_{t+h} = e^{Log(\widehat{D_{t+h}-D_t+1})} - 1 + D_t$$

4. Results

a. Case Rate

Table 4.1 Percentage of Countries that a Model outperform based on Smallest RMSE-Case Rate

		Models			
		Naive	ARIMA	HS w/o Diff	HS w Diff
Horizon	1	0.028902	0.543353	0.231214	0.196532
	3	0.034682	0.33526	0.508671	0.121387
	5	0.034682	0.346821	0.491329	0.127168

From *Table 4.1*, horizontally, we can see that relatively few countries have the smallest RMSE value of the Naïve model. When horizon =1, the ARIMA model is the best fit for 54% of counties. When horizon =3 and 5, horizon-specific models without difference are the best fit for 50% and 49% countries, respectively. Generally, the horizon-specific models without difference are suitable for more countries than those with difference, which may be caused by two reasons. In 2021, some countries like the US and Brazil have a worse COVID-19 situation because of the continuous outbreaks, while others, especially Asian Countries-China, Japan, and Korea-have better control over the disease with timely government regulation.

Vertically, the horizon-specific models with difference have similar percentage values,12%, at the three horizon time, indicating that around 12% of countries may have the best performance with this model. When we verify the explanation by researching the original dataset, we confirm that horizon-specific models with difference are the best fit for countries such as Japan, South Sudan, and Zambia at three different time predictions.

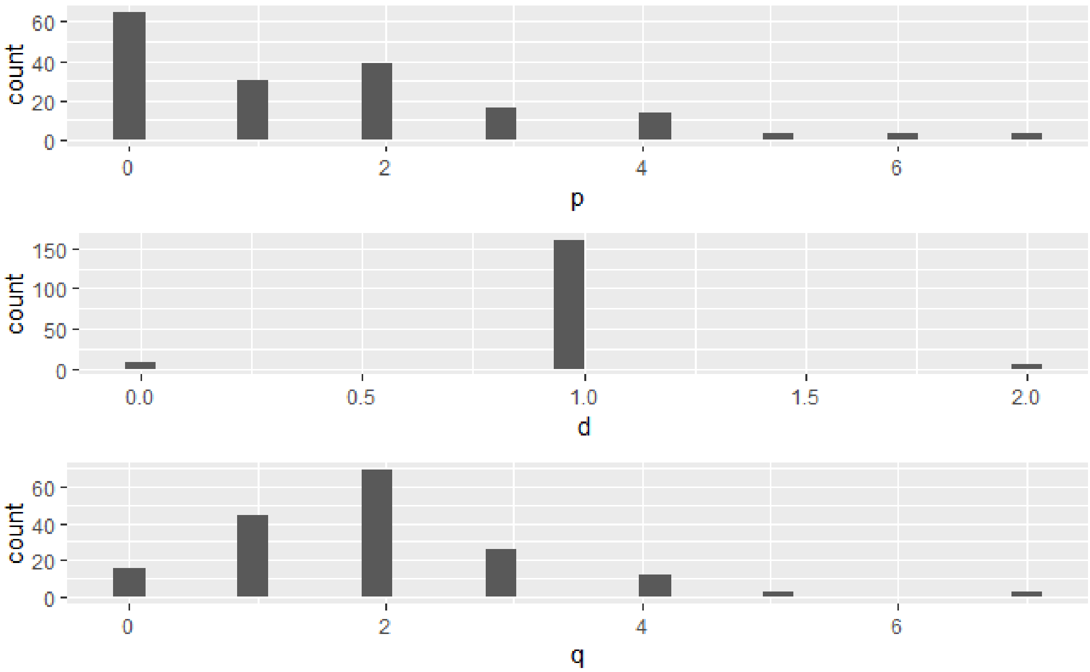
Table 4.2 Summary of Case Rate's RMSE for Different Models

	Horizon								
	1			3			5		
	Percentile of RMSE								
	5%	50%	95%	5%	50%	95%	5%	50%	95%
ARIMA	0.053482	2.364728	28.2511	0.061499	2.769217	30.13886	0.066402	3.101633	32.77622
HS w/o Diff	0.057362	2.41878	27.46566	0.07031	2.621284	27.21092	0.077105	2.853925	28.41588
HS w Diff	0.057261	2.702174	28.55484	0.072697	3.427373	34.71377	0.083387	3.465854	38.64386

Table 4.2 shows the 5%, 50%, and 95% percentile case rate's RMSE for different models. ARIMA has the smallest median RMSE with a horizon of 1 day ahead. However, the value does not significantly vary from the other models. Horizon-specific models without difference have the smallest median value at 3-days and 5-days ahead prediction. Horizontally, the RMSE of all three different models increases, showing that the errors are compounding at every prediction round.

When we have a closer look at the countries with extreme RMSE, we get some interesting results. For instance, Tanzania has a zero RMSE for the ARIMA model, which may be caused by the few sample data gathered from this country. On the other hand, Andorra has RMSE > 50 at all three models. As an inland European Country, Andorra is surrounded by France and Spain, indicating that its COVID-19's trending is much more complicated and can be significantly affected by its nearby countries. These results alert us that we may need to insert other new variables or pick up new models to conduct the prediction for countries like Andorra and Tanzania.

Figure 4.1 Summary of ARIMA's p, d, and q Values for Case Rate



When forecasting case rate's 1-day ahead, ARIMA models have dominant advantages over the other models. *Figure 4.1* summarizes the p, d, and q values of 173 countries, and the model ARIMA(0,1,2) is the best fit for most countries.

$$C_t - C_{t-1} = c + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t \quad (4.1)$$

Equation (4.1) describes the ARIMA(0,1,2) model, which is similar to a random walk model but with complex white noise calculation.

Figure 4.2 Predicting Case Rate of France

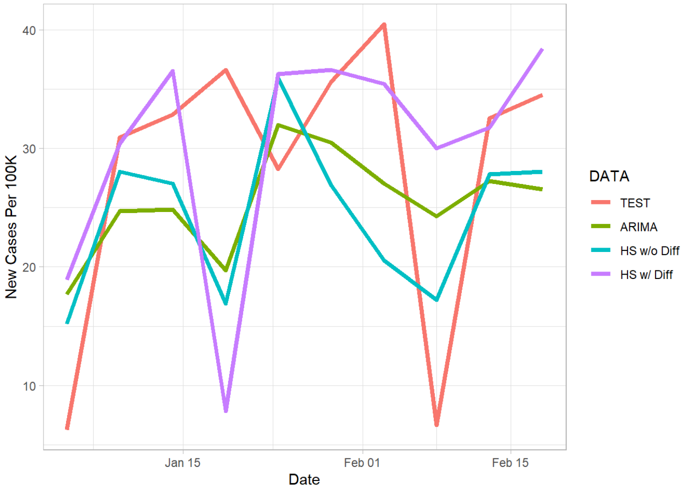


Figure 4.3 Predicting Case Rate of China

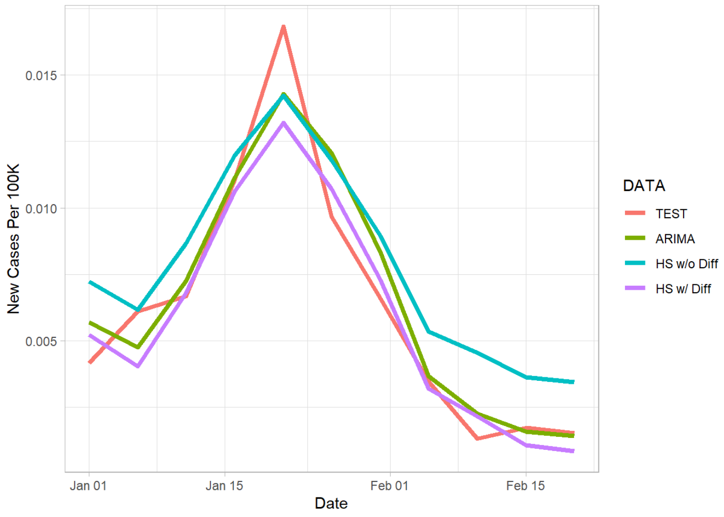


Figure 4.4 Predicting Case Rate of Brazil

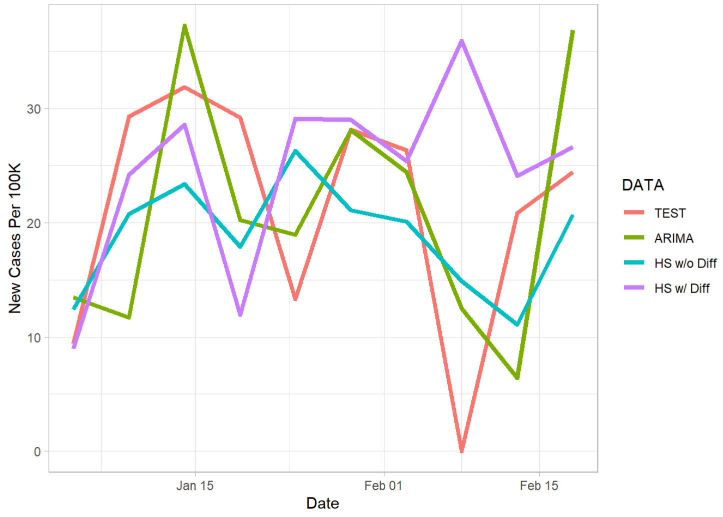
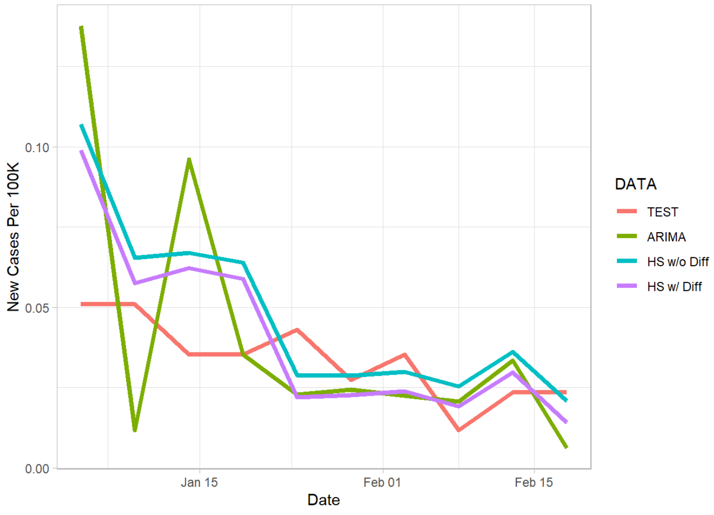


Figure 4.5 Predicting Case Rate of Australia



Figures 4.2-5 are the results of some typical countries. For France (*Figure 4.2)*, the curve of ARIMA is closer to the test curve than the other models. However, both China and Brazil (*Figures 4.3-4*) are more applicable to horizon-specific models than the ARIMA models. As for Australia (*Figure 4.5*), the horizon-specific model with the difference is the best fit.

b. Death Rate

Table 4.3 Percentage of Countries that a Model outperform based on Smallest RMSE-Death Rate

		Models			
		Naive	ARIMA	HS w/o Diff	HS w Diff
Horizon	1	0.069364	0.595376	0.127168	0.208092
	3	0.057803	0.398844	0.387283	0.156069
	5	0.046243	0.404624	0.34104	0.208092

From *Table 4.3*, we can see that relatively few countries have the smallest RMSE value of the Naïve model. Compared to the case rate's results, when horizon =1, 3, and 5, the ARIMA model outperforms the others and is the best fit for 60%, 39%, and 40% of counties.

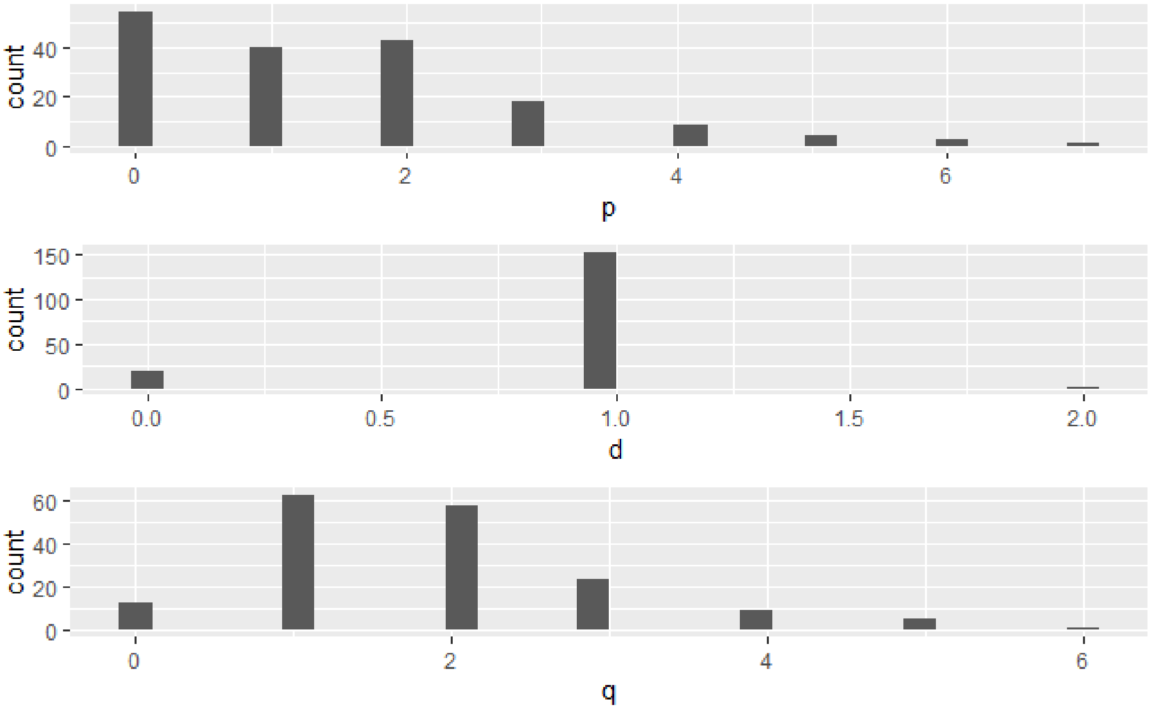
Table 4.4 Summary of Death Rate's RMSE for Different Models

	Horizon								
	1			3			5		
	Percentile of RMSE								
	5%	50%	95%	5%	50%	95%	5%	50%	95%
ARIMA	0.000338	0.05381	0.571891	0.000311	0.05633	0.605386	0.000106	0.056317	0.585087
HS w/o Diff	0.001122	0.056031	0.609481	0.001057	0.055335	0.542959	0.001093	0.057108	0.615005
HS w Diff	0.000263	0.056429	0.561369	0.000353	0.058027	0.563637	0.000559	0.063115	0.65123

Table 4.4 shows the 5%, 50%, and 95% percentile of death rate's RMSE for different models. If we have a closer look at the ARIMA, horizon-specific models with and without difference, it is apparent that ARIMA has the smallest mean RMSE at a horizon of 1 day and 5 days ahead. When forecasting the death rate at a horizon of 3 days ahead, ARIMA models and horizon-specific models without difference have similar RMSE values.

Noticeably, the RMSE of death rate is relatively smaller than the case rate, showing that many patients have been recovered from the infection.

Figure 4.6 Summary of ARIMA's p, d, and q Values for Death Rate



From *Tables 4.3-4*, we can see that the ARIMA models have dominant advantages over the other models at all horizons. *Figure 4.6* summarizes the p, d, and q values of 173 countries, and the model ARIMA(0,1,1) is the best fit for most countries.

$$D_t - D_{t-1} = \alpha \varepsilon_{t-1} \quad (4.2)$$

Equation (4.2) describes the ARIMA(0,1,1) model, which correcting auto-correlated errors in a random walk model by adding the simple exponential smoothing model [13].

Figure 4.7 Predicting Death Rate of France

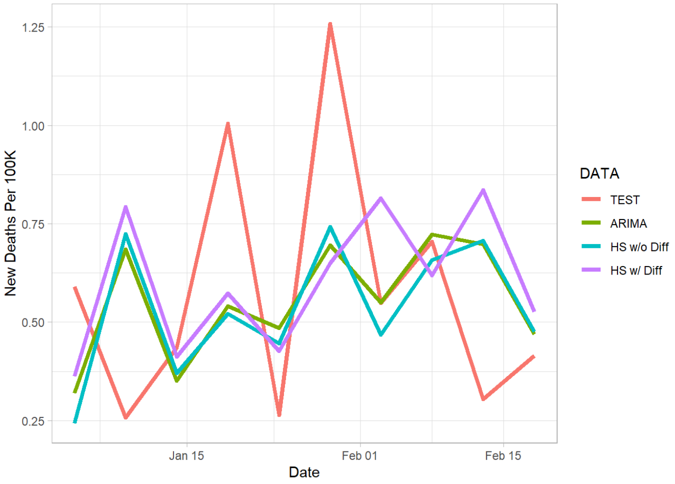


Figure 4.8 Predicting Death Rate of China

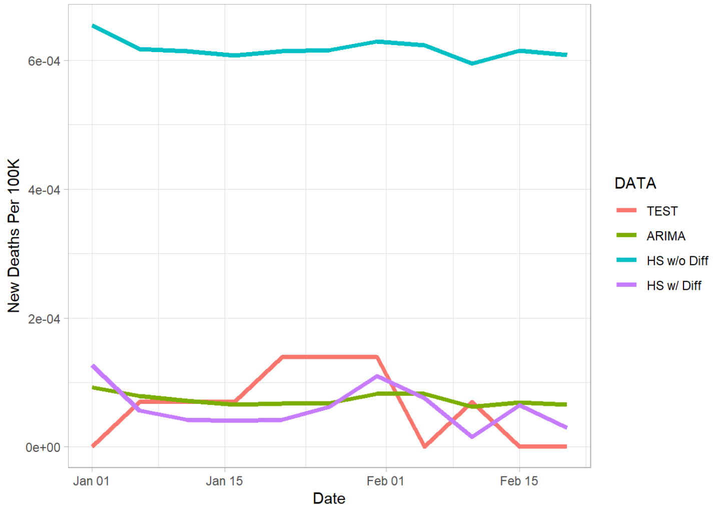


Figure 4.9 Predicting Death Rate of Brazil

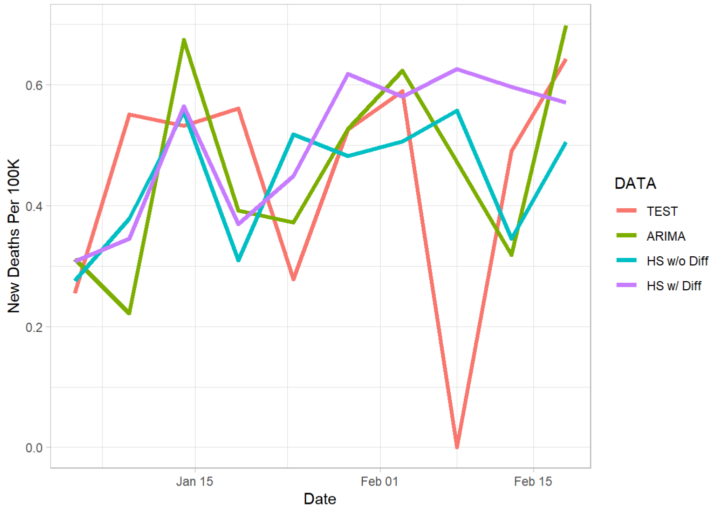
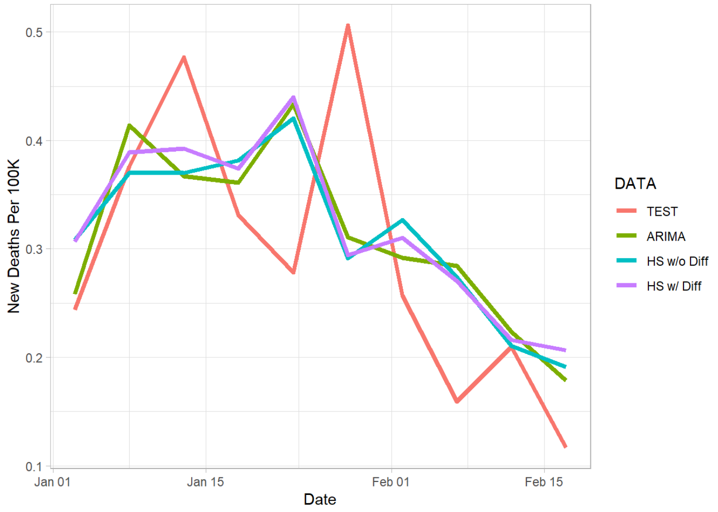


Figure 4.10 Predicting Death Rate of Canada



Figures 4.7-10 are the results of some typical countries. For France (*Figure 4.7*) and Brazil (*Figure 4.9*), the curve of ARIMA is closer to the test curve than the other models. On the other hand, the horizon-specific model with the difference is more applicable to China (*Figure 4.8*) and Canada (*Figure 4.10*).

5. Conclusion

Conclusively, the Naïve model is the least effective, and the ARIMA model applies to most countries. However, by viewing the graph above, it is evident that all models' curves are still having some difference with one of the test data. One major problem of the results is that we only conduct basic linear regression models while conducting horizon-specific models. To solve this problem, we could try running ridge regressions to improve the models' performance. Moreover, test data trends are dissimilar to train data, which may be affected by other factors, such as government regulations, vaccine inoculation, and other medical resources. We should insert more predictors into our models.

For the next step, we are going to explore clustering techniques to group countries according to the similarity in their ARIMA model orders and coefficients. Then create an R Shiny App where a user picks the country and horizon, and it outputs a table of the 1 day, 3 days, and 5 days forecast from all of the models in a table.

References

[1] Sauer, L. (2020). What is Coronavirus? Retrieved April 14, 2021, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

[2] Singhal, T. A Review of Coronavirus Disease-2019 (COVID-19). *Indian J Pediatr* 87**,** 281–286 (2020). https://doi.org/10.1007/s12098-020-03263-6

[3] Chan-Yeung M, Xu RH. SARS: epidemiology. Respirology. 2003;8: S9–14.

[4] Middle East Respiratory Syndrome Coronavirus. Available at: https://www.who.int/emergencies/mers-cov/en/. Accessed February 16, 2020.

[5] Qu, Jie-Ming, Bin Cao, and Rong-Chang Chen. 2020. *COVID-19: The Essentials of Prevention and Treatment*. Edited by Jie-Ming Qu, Bin Cao, and Rong-Chang Chen. Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-824003-8.09994-0.

[6] Maleki, M., Mahmoudi, M. R., Heydari, M. H., & Pho, K. (2020). Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using

time series models. *Chaos, Solitons & Fractals, 140*, 110151. DOI:10.1016/j.chaos.2020.110151

[7] Khan, F., Saeed, A., & Ali, S. (2020). Modeling and forecasting of new cases, deaths and recover cases of COVID-19 by using VECTOR Autoregressive model in Pakistan. *Chaos, Solitons & Fractals, 140*, 110189. DOI:10.1016/j.chaos.2020.110189

[8] GitHub/Owid. (2020). Owid/covid-19-data. Retrieved April 15, 2021, from https://github.com/owid/covid-19-data/tree/master/public/data/jhu

[9] WB. 2018. "Land Area." The World Bank. https://data.worldbank.org/indicator/AG.LND.TOTL.K2.

[10] Our World in Data. (2020). Retrieved April 15, 2021, from https://en.wikipedia.org/wiki/Our_World_in_Data

[11] McLaughlin, R. L. (1983). Forecasting models: Sophisticated or naive? *Journal of Forecasting (Pre-1986), 2*(3), 274. Retrieved from http://libproxy.lib.unc.edu/login?url=https://www.proquest.com/scholarly-journals/forecasting-models-sophisticated-naive/docview/224797219/se-2?accountid=14244

[12] Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on April 14, 2021.

[13] Business, F. (n.d.). Introduction to ARIMA models. Retrieved April 17, 2021, from https://people.duke.edu/~rnau/411arim.htm