Forecasting COVID-19 vaccine distribution in the United States, Japan, Taiwan, and China using the Auto-Regressive Integrated Moving Average (ARIMA) model

Kenneth (Hsuan An) Chen *
Project Advisor: Michael Tsiang †

August 23rd, 2021

Abstract

Developed at unprecedented speeds, vaccines have thus far played a crucial role in slowing down the COVID-19 pandemic around the world. Therefore, it is an absolute necessity for countries to be able to accurately forecast the distribution of vaccines. This paper uses an Auto-Regressive Integrated Moving Average (ARIMA) model to analyze and forecast 30 days of COVID-19 vaccine distribution for the United States, Japan, Taiwan, and China. Specifically, for the United States and Japan, the predicted variable was the percent of the population that was fully vaccinated while the predicted variable for Taiwan and China was the total number of doses administered. The data used to fit our model was pulled from a publicly available dataset compiled from various sources around the world. For each country, the training data consisted of that country's vaccination data from whenever they first administered vaccines until July 19, 2021. After fitting the model on the training data, the model was then tested against 30 days of data from July 20, 2021 to August 18, 2021. The paper found that the univariate ARIMA model was able to, on average, forecast the distribution of COVID-19 vaccines within 5% of the actual value for each country.

1. Introduction

As of August 24, 2021, the COVID-19 pandemic, caused by the coronavirus SARS-CoV-2, has claimed the lives of over 4.4 million people worldwide according to Johns Hopkins University [1]. As the virus continues to spread around the world, the need for vaccines becomes more and more essential with each passing day. Vaccines made by Pfizer-BioNTech or Moderna are used by the United States, Japan, and Taiwan, and they are 94% effective against COVID-19 hospitalization among fully vaccinated adults and 64% effective among partially vaccinated adults aged 65 years and older [2]. The Sinovac vaccine, often used by China, has been shown to have an efficacy of 51% against symptomatic SARS-CoV-2 infection, 100% against severe COVID-19, and 100% against hospitalization [3]. By reducing risk of hospitalization and slowing the spread of the virus, the vaccines give countries control over the pandemic and protection for their people. Therefore, the ability to predict and forecast the distribution of vaccines is essential for healthcare systems to plan ahead and produce the most optimal strategy against the pandemic going forward.

*Department of Statistics, University of California, Los Angeles, kennethc22@g.ucla.edu
†Department of Statistics, University of California, Los Angeles, michael.tsiang@stat.ucla.edu

In this paper, we analyze and forecast the vaccine distribution for the United States, Japan, Taiwan, and China. For the United States, a country that initially struggled to contain the spread of the virus, the COVID-19 vaccines became a core part of their campaign against the pandemic. An aggressive push for the distribution of vaccines allowed for the country to administer over 200 million doses within the first 100 days of President Joe Biden's term [4]. Japan, on the other hand, experienced early success in terms of containing the spread of the virus [5].

However, with the Olympics held in Tokyo during the summer of 2021 and a recent surge in COVID-19 cases, vaccine distribution became a priority for the country [6]. Similarly, during the early months of the pandemic, Taiwan was one of the most successful countries in terms of suppressing the pandemic. However, as restrictions began to relax and the island nation struggled to obtain vaccines, there has been a resurgence of COVID-19 outbreaks [7]. Lastly, China, despite being ground zero for the COVID-19 pandemic, has been on the forefront of global vaccine distribution [8]. As of June 9, 2021, China has administered nearly 60% of all COVID-19 vaccine doses globally [9]. Figure 1 shows the daily COVID-19 cases for the United States, Japan, Taiwan, and China.

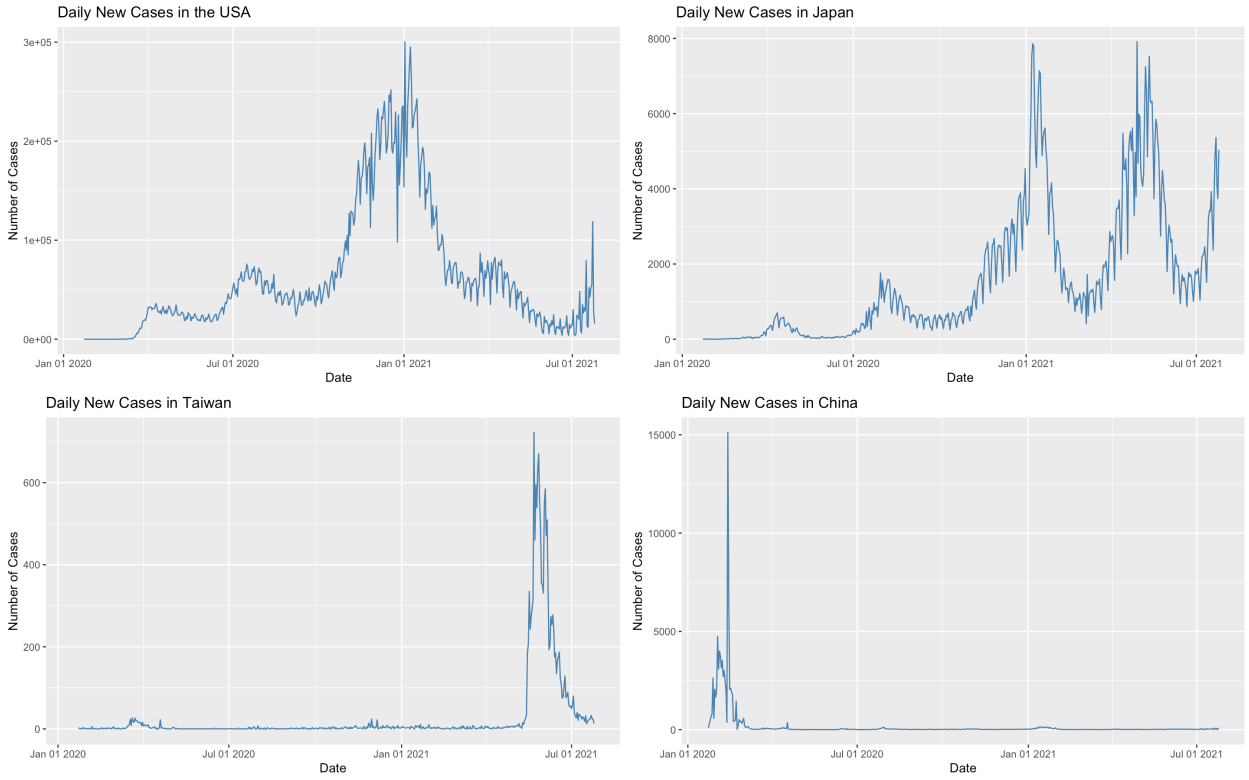

Figure 1: Daily COVID-19 cases for the United States, Japan, Taiwan, and China

In recent research, Cihan used Auto-Regressive Integrated Moving Average (ARIMA) models as well to predict COVID-19 vaccine distribution on the continental level [10]. Similarly, substantial work has been done with forecasting the spread of COVID-19 cases using the ARIMA model. Moftakhar and Seif used the ARIMA model to forecast COVID-19 cases in Iran for the next 30 days after March 21, 2020 [11]. Marbaniang predicted the trend of COVID-19 cases around India for the next 20 days after March 18, 2020 using the ARIMA model as well [12]. Perone utilized the ARIMA model to predict cumulative COVID-19 cases in Italy for a period after April 4, 2020 [13]. ArunKumar et al. used both ARIMA and SARIMA models to forecast the epidemiological trends of the COVID-19 pandemic for the 16 countries that account for most of the global cumulative cases [14].

The remainder of the paper is organized as follows. Section 2 will provide information on the dataset. Section

3 describes the methodology used in the paper including imputation, data analysis, model selection, and model accuracy. The results, which will include the final model for each country as well as its respective accuracy with the test data, will be provided in Section 4. Subsequent discussion will be in Section 5. Lastly, the conclusion will be in Section 6 where we discuss the implications and findings of this paper. The programming language used in this paper was R and we primarily utilized the following packages: ggplot2, car, forecast, astsa, Metrics, tseries, and imputeTS [15, 16, 17, 18, 19, 20, 21, 22].

2. Data

Table 1: Summary statistics of the predicted variables

Country	Min.	Q1	Median	Mean	Q3	Max.	SD
USA	0.00405	0.08386	0.25423	0.24803	0.41226	0.48717	0.1682881
Japan	0	0.00617	0.02190	0.05431	0.08399	0.23800	0.06883339
Taiwan	0	28,883	190,414	943,717	1,597,721	5,419,988	1,373,305
China	1,500,000	170,781,750	386,810,000	566,887,983	950,525,500	1,467,316,000	468,608,872

The data is obtained from Our World in Data, and the dataset is compiled from various sources depending on the country. The data for the United States is obtained from the CDC, the Prime Minister's Office for Japan, the National Health Commission for China, and the Taiwan CDC for Taiwan [23]. The data itself contains daily information regarding COVID-19 vaccinations for 234 countries around the world such as total doses administered, total number of people that are fully vaccinated, and the daily number of vaccinations. At the time of writing, the dataset consists of 14 columns and over 42,000 rows. Note that the dataset is updated daily with new data.

For the United States and Japan, the predicted variable is the percent of the population that is fully vaccinated. This variable was created by dividing the number of people that are fully vaccinated by the total population according to the 2020 Census for both the United States and Japan [24, 25]. For Taiwan and China, due to a lack of data, the predicted variable will be the total number of doses administered. The data for these variables shown over time is presented in Figure 2 below.

Missingness is often a term used to describe the manner in which data are missing from a sample of a population [26]. Specifically, with our dataset, we see that the missingness varies by country (see Table 2). For the United States and Japan, there are relatively few missing values compared to the total length of the time series with most of the missing values occurring during the earlier days. This missingness would most likely be due to the fact that during the early days of vaccine distribution, there may not be enough information to warrant a daily report. On the other hand, Taiwan and China have a comparatively larger number of missing values. For Taiwan, the missingness is distributed more evenly while China has all of their missing values occurring during the first half of the time series. Currently, Taiwan has been struggling to obtain vaccines, so this may be the most likely reason for the missingness in the data. With regards to China, the National Health Commission may have simply decided not to report the information daily for the first 108 days or so of the time series. Imputation for these values will be discussed in Section 3. Note that for the United States and Japan, the first 25 and 21 values, respectively, were missing because there had yet to be anybody that was fully vaccinated during those days.

Table 2: Summary statistics of missingness in the data

Country	Length	Missing Values	% Missing	Bin 1	Bin 2	Bin 3	Bin 4
USA	187	6	3.21%	4 NAs	0 NAs	1 NAs	1 NAs
Japan	132	10	7.58%	10 NAs	0 NAs	0 NAs	0 NAs
Taiwan	121	23	19%	2 NAs	5 NAs	12 NAs	4 NAs

Country	Length	Missing Values	% Missing	Bin 1	Bin 2	Bin 3	Bin 4
China	217	85	39.2%	47 NAs	38 NAs	0 NAs	0 NAs

3. Methodology

3.1 ARIMA

For this paper, we mainly utilized the Auto-Regressive Integrated Moving Average (ARIMA) model to analyze and forecast the data. First introduced by Box and Jenkins in 1976, the ARIMA(p, d, q) model is used to forecast time series data by taking into account the data and residual error at previous time values [27]. Further, it can incorporate possible non-stationarity of the data via differencing. Specifically, an ARIMA model is essentially broken down into three parts:

1. Auto-Regression (AR) with order p
2. Integrated (I) with order d
3. Moving Average (MA) with order q.

Let y_t be the time series data at time t and $\epsilon_t \sim wn(0, \sigma^2)$ is a white noise series at time t. The AR(p) model regresses the current variable at time t against the previous p time values as follows:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t$$

where α is the intercept and ϕ_i for $i \in \{1, 2, ..., p\}$ are the coefficients of the model. On the other hand, the MA(q) model regresses the current variable at time t against the residuals at the previous q time values as follows:

$$y_t = \alpha + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_p \epsilon_{t-p} + \epsilon_t$$

where α is the intercept and θ_i for $i \in \{1, 2, ..., q\}$ are the coefficients of the model. The ARMA(p, q) model then incorporates both AR(p) and MA(q) as follows:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_p \epsilon_{t-p} + \epsilon_t.$$

Finally, the ARIMA(p, d, q) model accounts for non-stationarity in the time series data by applying an ARMA(p, q) model to the time series differenced d times. This is essentially the "integrated" part of the ARIMA model. A time series value at time t that has been differenced once, $y_t^{(1)}$, is defined as follows:

$$y_t^{(1)} = y_t - y_{t-1},$$

and a time series value at time t that has been differenced twice, y_t'', is defined as follows:

$$y_t^{(2)} = y_t^{(1)} - y_{t-1}^{(1)} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}.$$

A time series with d-order differencing has been differenced d times. Note that although y_t itself need not be a stationary time series to fit an ARIMA(p, d, q) model, it is assumed that y_t is stationary after d-order differencing.

3.2 Imputation

Imputation, the acting of filling in missing values in the data, is necessary to reduce bias and improve the quality of our model. To impute missing values in the data, we utilized the package imputeTS created by Moritz and Bartz-Beielstein. For each imputation algorithm in imputeTS we did the following:

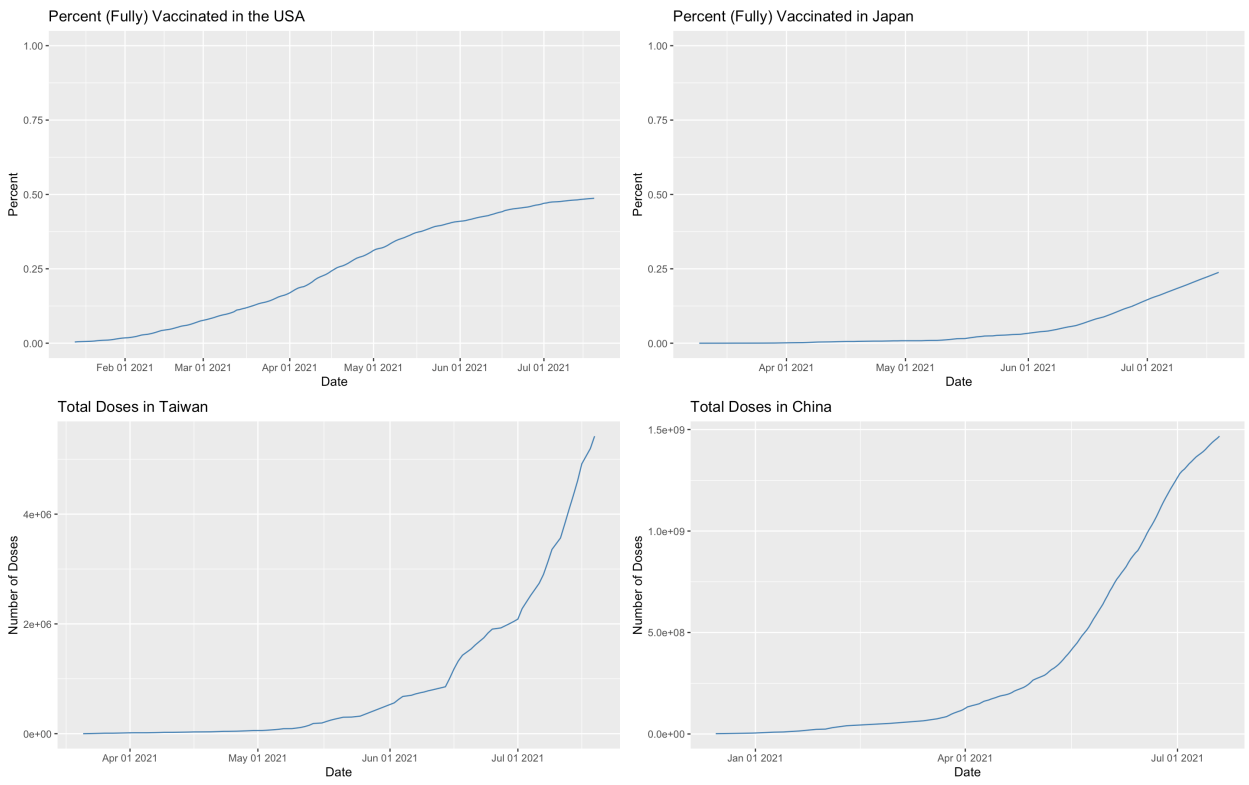

Figure 2: Time series plot of the predicted variables for the United States, Japan, Taiwan, and China

1. Impute the dataset (for a given country)
2. Perform an 80-20 split on the dataset for cross-validation
3. Fit the cross-validation training dataset with an ARIMA(p, d, q) model using the auto.arima[1] function
4. Test model accuracy with the cross-validation test dataset using MAE and MAPE as evaluation metrics

Note that for an 80-20 split, the training data will consist of the first 80% of days and the validation data will consist of the final 20% of days. For instance, if there are 100 days of data in total, the first 80 days of data will be the training data and the final 20 days will be the validation data. For each country, we used this procedure to pick the imputation algorithm that gave us the best test results from the cross-validation. After we select an imputation algorithm for a given country, we tune the parameters of the algorithm by applying the same procedure to a set of parameter candidates.

3.3 Exploratory Data Analysis

Before we fit an ARIMA model, we must first check that the data is either stationary or stationary after some level of differencing. To do so, we check the autocorrelation function (ACF) plots for each country (Figure 3).

Upon examination of Figure 3, it is clear that the time series data from each country is not stationary. Thus, we perform differencing for each dataset in order to produce a stationary time series. For each country, we tried first order, second order, and third order differencing to determine the order of differencing. Any order higher than three would cause overdifferencing and thus was not attempted. Second order differencing proved to be the most effective for each dataset in terms of producing a stationary time series, as evidenced by Figure

[1]The auto.arima function in the forecast package uses a variation of the Hyndman-Khandakar algorithm to fit an ARIMA model [18]

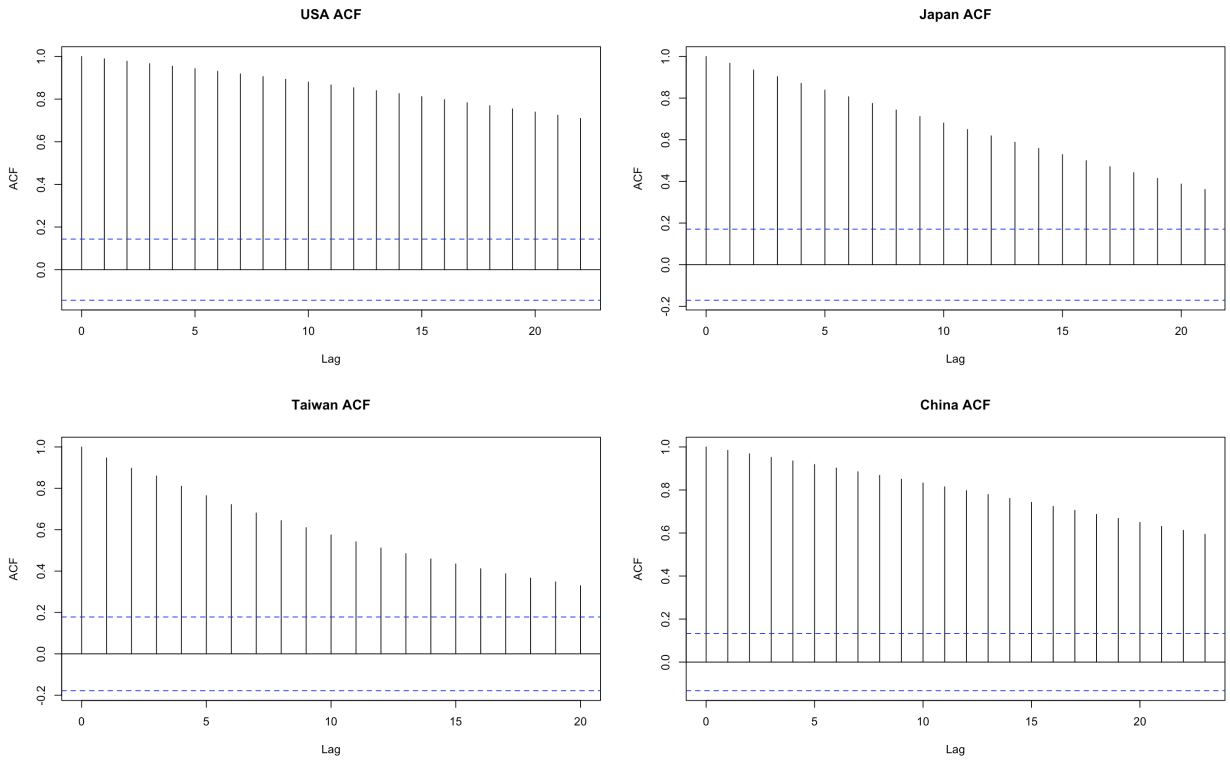

Figure 3: Autocorrelation Function (ACF) plots for the United States, Japan, Taiwan, and China. The blue dotted lines indicate the 95% confidence bands for autocorrelation. Values higher than the line indicate statistically significant autocorrelation.

4. Note that in the context of the vaccination data set, first order differencing produces the daily increase in the percentage of the population that is vaccinated (or total number of vaccine doses administered) while second order differencing produces the rate of increase. To further confirm stationarity of the time series, we examined the ACF plot and ran an Augmented Dickey-Fuller[2] (ADF) test as shown in Figure 5 and Table 3, respectively. The adf.test function in the tseries package was used to run the ADF test.

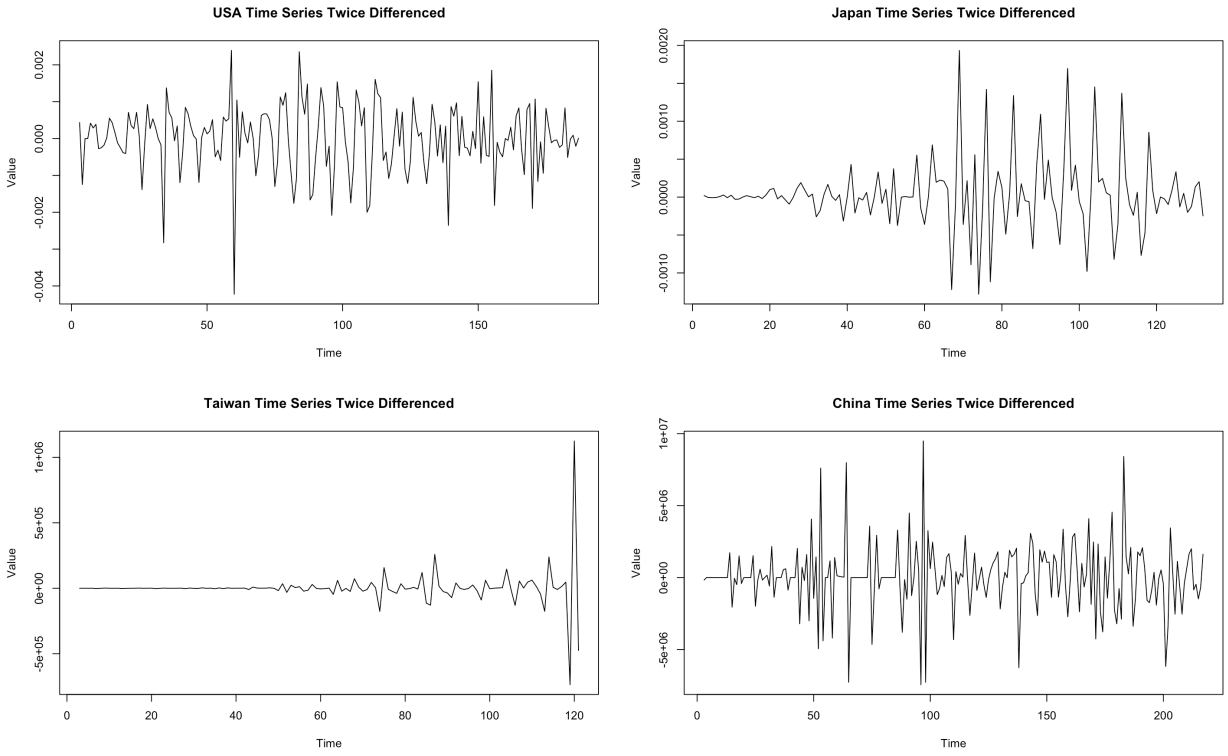

Figure 4: Twice differenced time series plot for the United States, Japan, Taiwan, and China

Table 3: ADF test results for twice differenced data

Country	Test Statistic	P-value
USA	-13.822	< 0.01
Japan	-12.559	< 0.01
Taiwan	-23.239	< 0.01
China	-20.799	< 0.01

It should be important to note that the ACF plot for the United States in Figure 5 exhibits a sinusoidal pattern which may be an indication of non-stationarity. This pattern continues even as we increase the order of differencing to three and four (any order higher than four was not attempted in order to avoid overdifferencing). However, since the differenced time series plot and the Augmented Dickey-Fuller test indicated strong evidence of stationarity after second order differencing, we deemed the time series data for the United States to be "stationary enough" for practical use. The reason being that real-world data can be messy at times, and thus it is unreasonable to expect perfect stationarity from real-world datasets. Similarly, with the differenced ACF plot for Japan in Figure 5, we see that there is statistically significant autocorrelation at lags that are multiples of 7 (lags 7, 14, and 21) which may also be an indication of non-stationary. However, like with the time series data for the United States, we deemed the Japanese time series

[2]The ADF test tests the null hypothesis that a unit root is present in a time series sample [28]

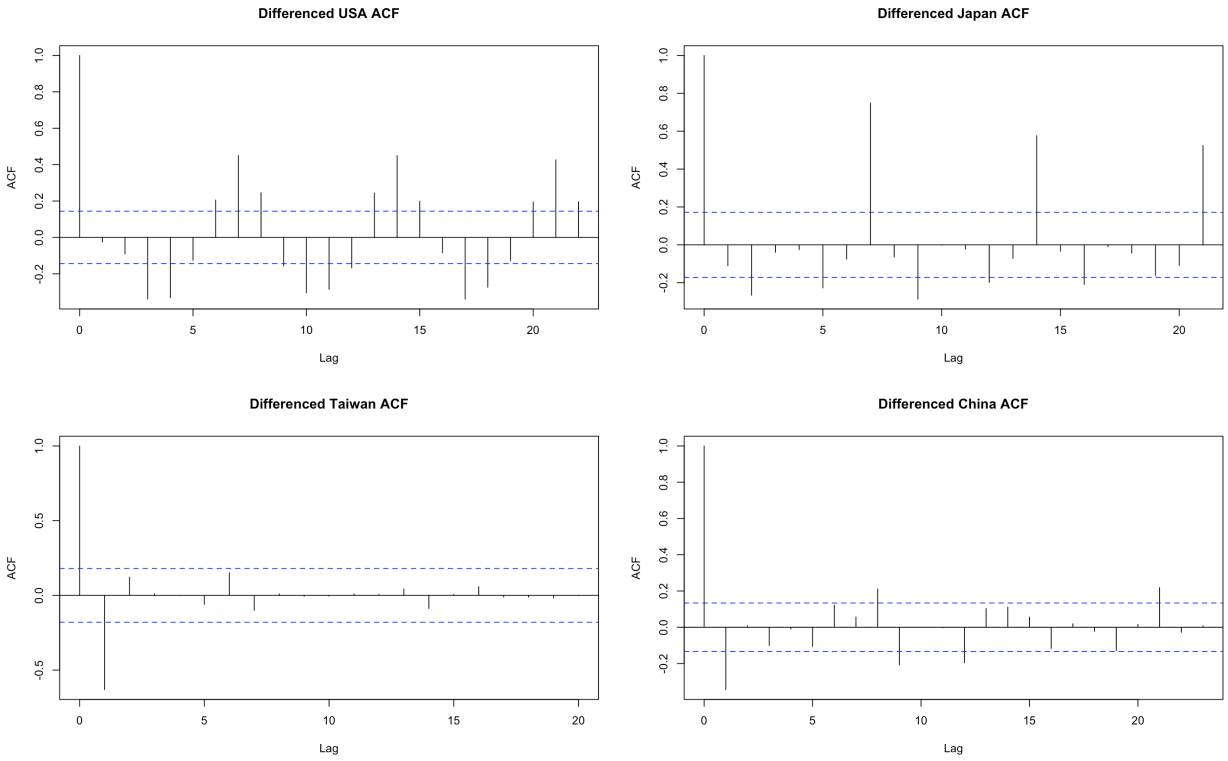

Figure 5: Autocorrelation Function (ACF) plot of twice differenced time series for the United States, Japan, Taiwan, and China

data to be "stationary enough" as well after examining the differenced time series plot and conducting the Augmented Dickey-Fuller test.

3.4 Model Selection

For each country, we first fit an ARIMA(p, d, q) model to the dataset using the auto.arima function and check diagnostics using the sarima function in the astsa package. To determine if the residuals are stationary, we mainly examine the ACF plot and Ljung-Box statistic plot to check for statistically significant autocorrelation. An example of satisfactory diagnostics plots is provided in Figure 6.

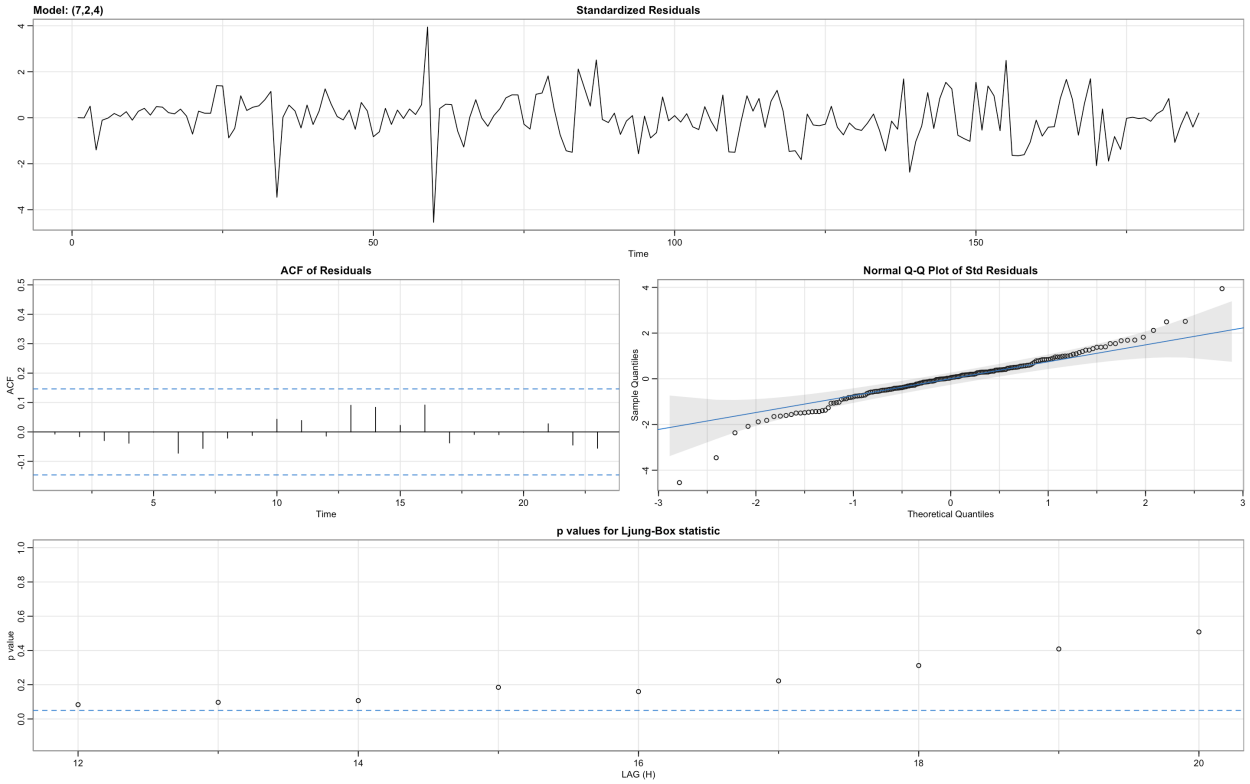

Figure 6: Diagnostics for an ARIMA(7, 2, 4) model fitted on the USA time series. The top graph is a plot of the standardized residuals to check the assumption that the residuals are stationary. The middle-left graph is an autocorrelation function (ACF) plot of the residuals to check if residuals are autocorrelated. The middle-right graph is a Normal Q-Q plot of the standardized residuals to check if the distribution of the residuals is normal. The bottom graph is a plot of the p-values for the Ljung-Box statistics. The Ljung-Box test checks for significant autocorrelation up to specified lags.

If we see that the residuals are not stationary, we then manually select parameters for our ARIMA(p, d, q) model by looking at the ACF and partial autocorrelation function (PACF) plots for the given country. That is, by examining the ACF and PACF plots and looking for lag values where there is statistically significant autocorrelation and partial autocorrelation, respectively, we come up with a set of candidate values for the orders p and q of the ARIMA(p, d, q) model ($d = 2$ here since our data is twice differenced). We then try every combination of $(p, 2, q)$ values until we get satisfactory diagnostics from the residual. If there is more than one combination that provides satisfactory diagnostics, we then run cross-validation with an 80-20 split for each of the combinations and pick the combination that provides the best test results in terms of MAPE and MAE. Finally, we run an ADF test on the residuals to further confirm stationarity. This process is summarized in Figure 8 in the form of a flowchart.

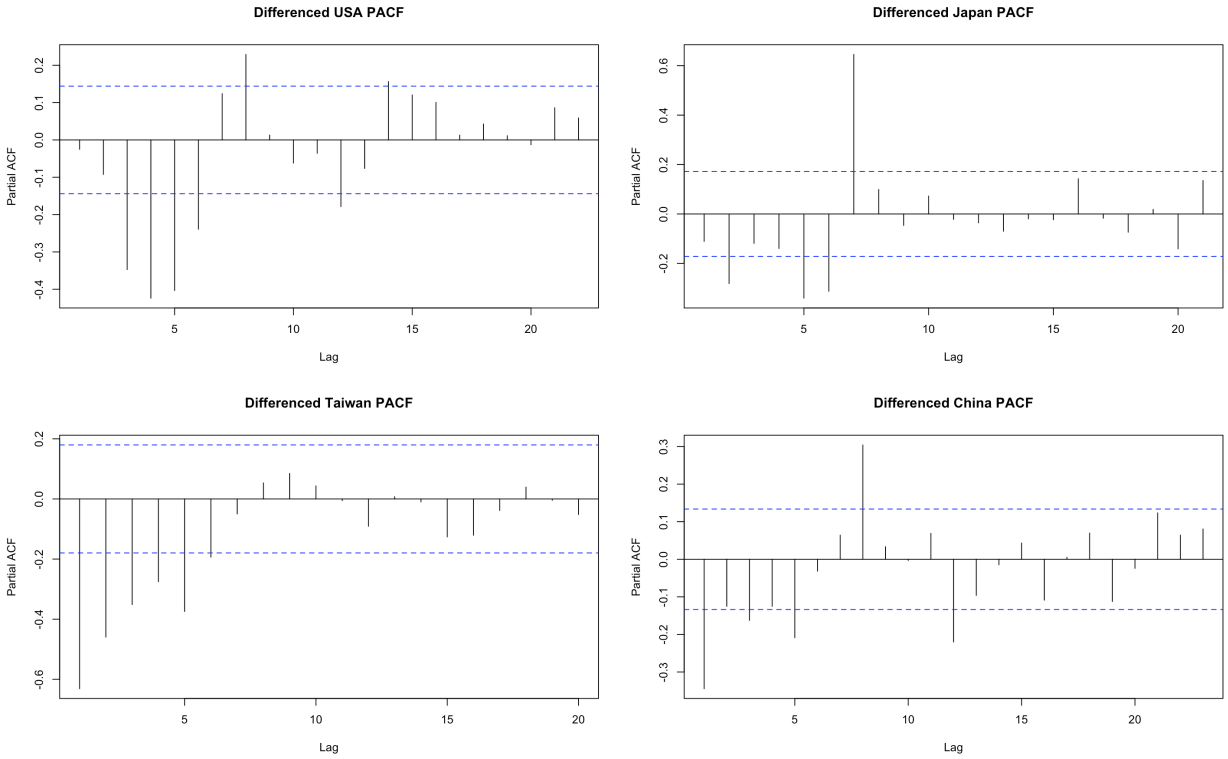

Figure 7: Partial autocorrelation function (PACF) plot of twice differenced time series for the United States, Japan, Taiwan, and China. Partial autocorrelation measures the autocorrelation of different lags conditioning on the values in between. The blue dotted lines indicate the 95% confidence bands for partial autocorrelation. Values higher than the line indicate statistically significant partial autocorrelation.

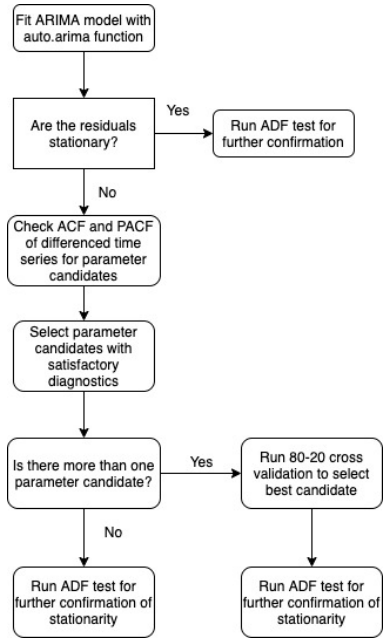

Figure 8: Flowchart of the model selection process

3.5 Final Accuracy & Forecast

The dataset was obtained on July 19, 2021 and the dataset is updated daily with new data for that day. The goal of this paper is to evaluate how well the ARIMA model can forecast and predict future values of vaccine distribution. Thus, to test the accuracy, the models will be fitted on all the data until July 19, 2021 and tested against the incoming data from the 30 days between July 20, 2021 and August 18, 2021. That is, the models will be tested against 30 "future" days of data in order to determine how well they can forecast 30 days into the future. Once we have tested the accuracy of the model, we fit the model again on all the data until August 18, 2021 in order to forecast the vaccine distribution for the next 30 days from August 19, 2021 to September 17, 2021.

4 Results

4.1 Imputation

We summarize the results for the imputation selection process in Table 4.

Table 4: Cross-validation results from imputation method selection

Country	Method	Training MAPE	Training MAE	Test MAPE	Test MAE
USA	Next Observation Carried Backward	0.63780%	0.0004656	1.32378%	0.0064919
Japan	Spline Interpolation	4.91300%	0.0002662	0.30161%	0.0004777
Taiwan	Simple Moving Average	5.02959%	15,491	18.03203%	569,333
China	Linear Moving Average	2.71791%	1,333,107	1.81786%	23,253,830

As the name suggests, Next Observation Carried Backward (NOCB) imputes the missing value at time t with the nearest non-missing value at time $t^* > t$. Since NOCB does not have any parameters, there was no need for additional tuning. For Japan, the main parameter that needed to be tuned for (cubic) spline interpolation was the specific method of spline interpolation that was used. The resulting method that was chosen from the procedure described in Section 3.2 was the Forsythe, Malcolm, and Moler (FMM) spline interpolation [29]. Note that spline interpolation imputes missing values by fitting multiple piecewise (cubic, in this case) polynomials to the data. For Taiwan and China, the main parameter to tune was the number of non-missing neighbors, k, to take into account when calculating the moving averages. For instance, a value of $k = 4$ means that the moving average is calculated with the two nearest non-missing neighbors on each side of a given missing value. Note that a simple moving average imputes a missing value with the arithmetic mean while a linear (weighted) moving average utilizes weights that decrease in arithmetical progression for its average value. The parameter candidates for both datasets were 2, 4, 6, 8 (1, 2, 3, 4 neighbors on each side, respectively). For Taiwan, the parameter selected was $k = 6$ and for China, the parameter selected was $k = 4$.

Using the graphing functions in the imputeTS package, we can visualize the imputations in the context of the original time series. Specifically, in Figure 9 and Figure 10 below, the original data values are shown with blue circles and the new imputed values are shown with red diamonds.

4.2 Model Selection

The results of the model selection process described in Section 3.4 are summarized below in Table 5.

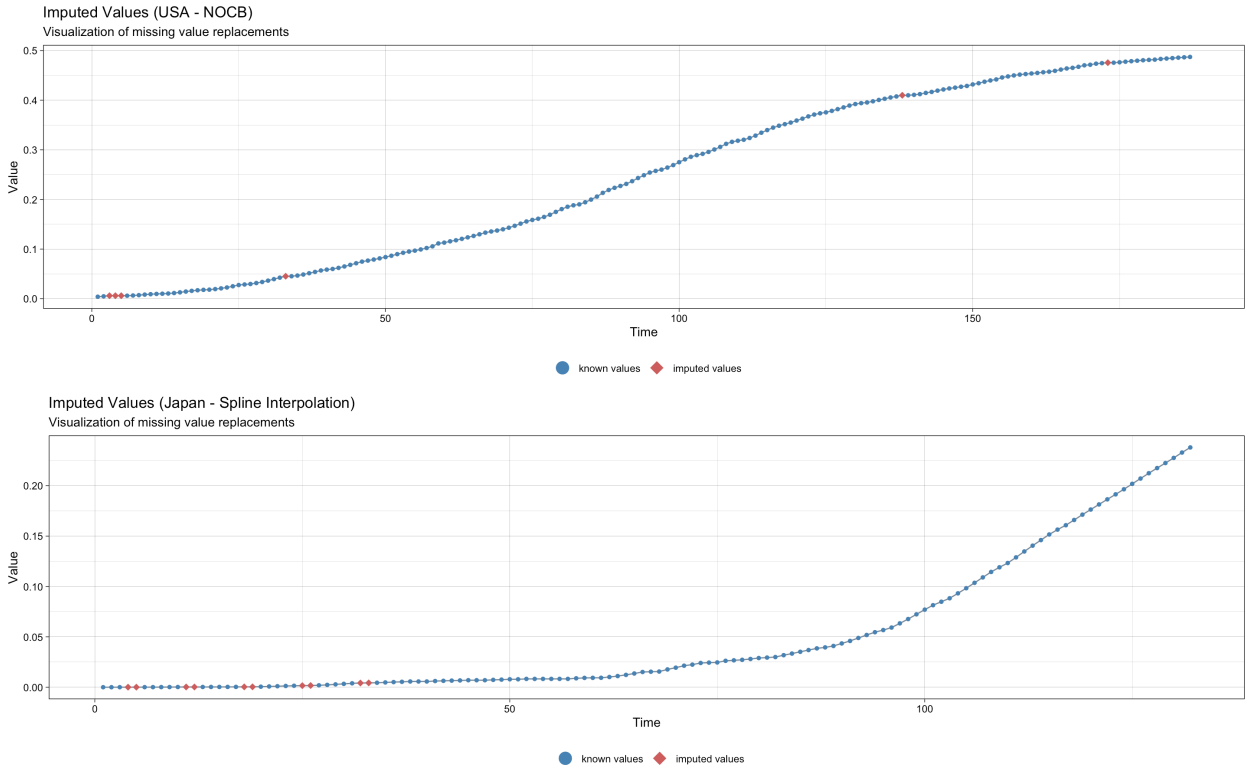

Figure 9: Time series plots with imputations shown for the United States and Japan

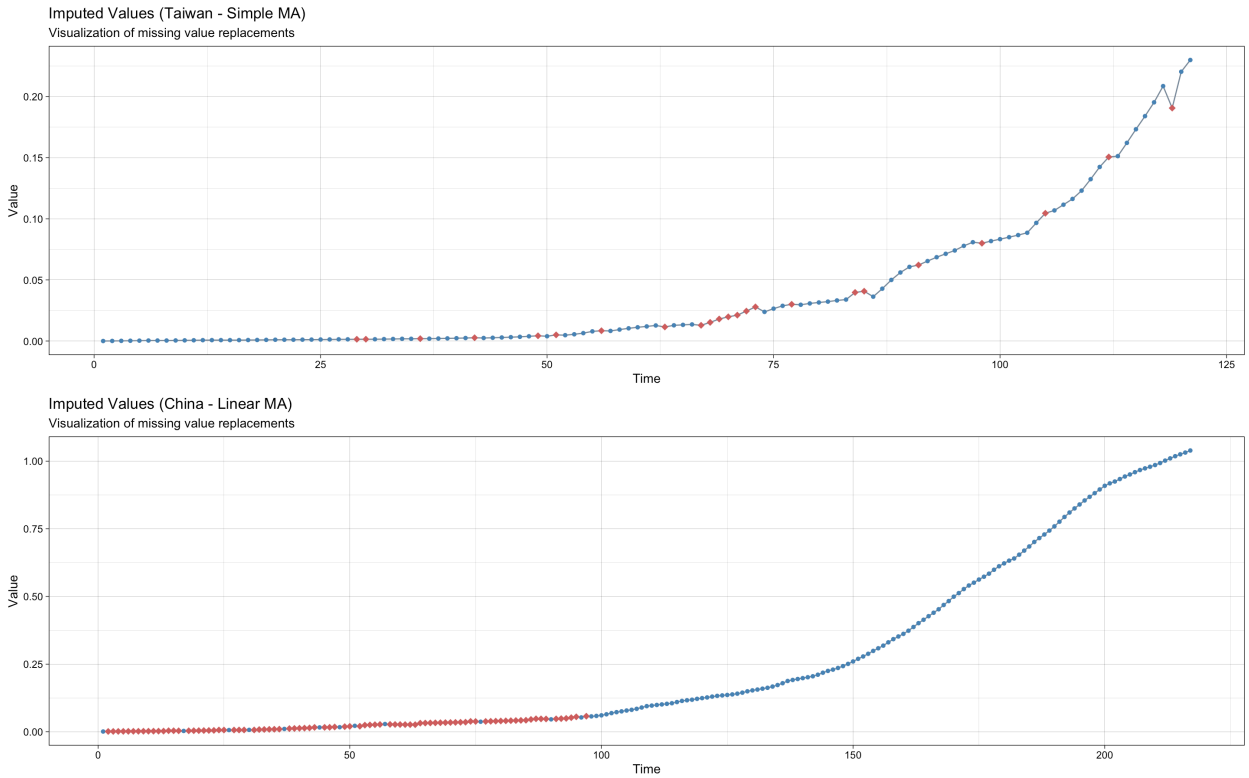

Figure 10: Time series plots with imputations shown for Taiwan and China

Table 5: Cross-validation results of model selection process

Country	Model	Training MAPE	Training MAE	Test MAPE	Test MAE
USA	ARIMA$(7, 2, 4)$	0.6170833%	0.0004236083	0.9249835%	0.004350111
Japan	ARIMA$(7, 2, 2)$	3.056263%	0.0001943532	1.746194%	0.003066149
Taiwan	ARIMA$(0, 2, 3)$	5.379693%	16,860	13.12162%	517,400
China	ARIMA$(0, 2, 9)$	2.755072%	1,304,621	1.879643%	23,465,862

4.3 Final Accuracy & Forecast

The results of the models tested against the 30 days of test data from July 20, 2021 to August 18, 2021 are shown in Table 8 below. We compare the scores of the ARIMA models with the scores of corresponding simple linear regression (SLR) models fitted to the same dataset with time as the independent variable. In order to increase the accuracy of the regression model for Japan, Taiwan, and China, the output variable is transformed using the inverseResponsePlot (IRP) function from the car package. The IRP function aims to improve the fit of a linear regression model by raising the response variable to the power of λ. Specifically, nonlinear least squares is used to estimate λ in the function $\hat{Y} = b_0 + b_1 Y^{\lambda}$ where Y is the response and $\hat{Y}$ are the fitted values. This relationship is then visualized by plotting Y on the vertical axis and $\hat{Y}$ on the horizontal axis [30].

Table 6: Comparison of final accuracy between ARIMA and SLR

Country	Model	Final MAPE	Final MAE
USA	ARIMA$(7, 2, 4)$	0.1315424%	0.0006598469
USA	SLR	14.56898%	0.08595986
Japan	ARIMA$(7, 2, 2)$	1.174482%	0.003836685
Japan	SLR with IRP ($\lambda = 0.23$)	8.064614%	0.03261154
Taiwan	ARIMA$(0, 2, 3)$	4.805561%	451,903
Taiwan	SLR with IRP ($\lambda = 0.08$)	30.36361%	4,733,186
China	ARIMA$(0, 2, 9)$	4.261691%	71,481,164
China	SLR with IRP ($\lambda = \frac{1}{8}$)	29.38651%	747,587,791

Once we have tested the accuracy of the model, we fit the model to the data up until August 18, 2021, and use the model to forecast a further 30 days into the future. That is, we predict what the vaccine distribution will look like for the United States, Japan, Taiwan, and China from August 19, 2021 to September 17, 2021. Figure 11 and 12 show the forecasted values of the ARIMA(p, d, q) model for each country with a blue line while the 80% and 95% confidence intervals are shown with the dark blue and gray shaded areas, respectively. For the United States and Japan, the models predict that around 54% and 53% of the respective populations will be fully vaccinated by September 17, 2021. For Taiwan and China, the models predict that 11,971,830 and 2,247,671,449 total vaccine doses, respectively, will be administered by September 17, 2021.

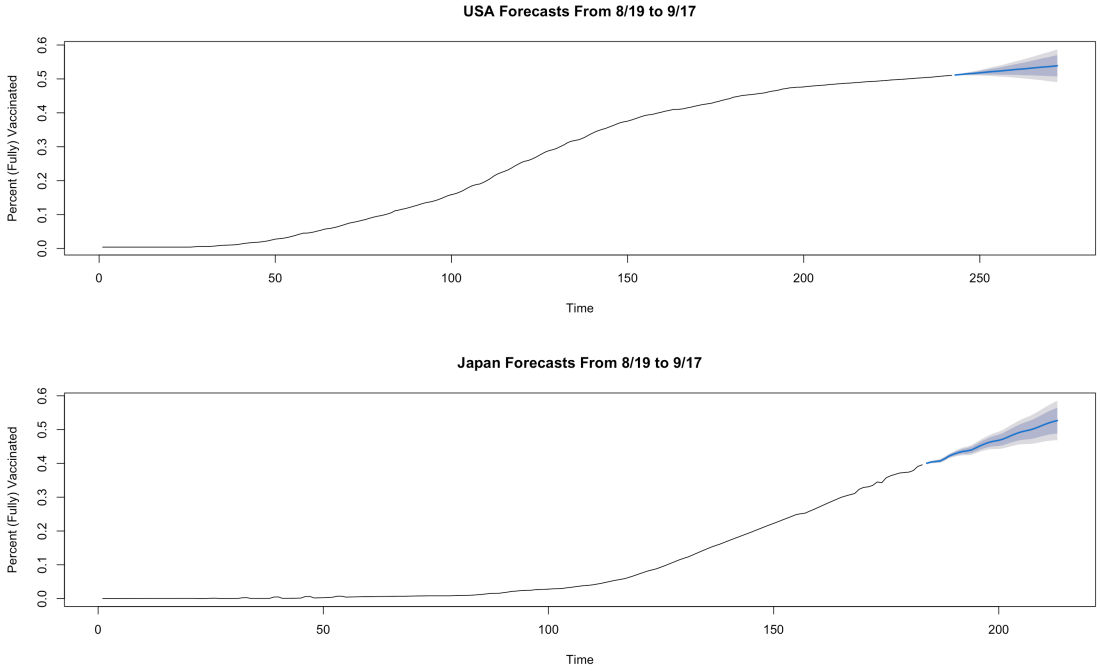

Figure 11: Forecasted values over time (in days) for the United States and Japan. The number of days was used instead of actual dates for clarity because each country started vaccinations at different dates.

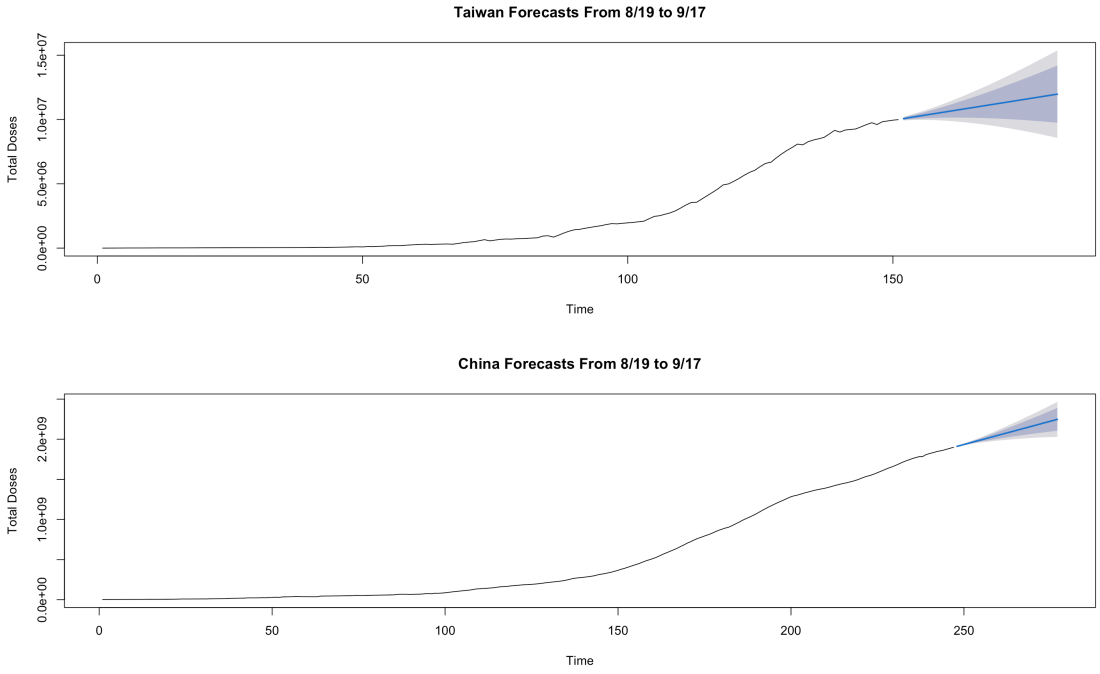

Figure 12: Forecasted values over time (in days) for the Taiwan and China

5 Discussion

5.1 Accuracy

For each country, the model is quite accurate in terms of minimizing MAE and MAPE. If we refer to the MAE column in Table 8, the ARIMA(p, d, q) model for the United States makes, on average, predictions that are about 0.066% off from the actual percentage of the population that is fully vaccinated. That is exceptionally accurate when compared to the standard deviation of around 16.8% for the time series from Table 1. From Figure 13, we see that the model tends to slightly underestimate the values. Similarly, for Japan with the MAE from Table 8, we see that the ARIMA(p, d, q) model, on average, makes predictions that are about 0.384% off from the actual percentage of the population that is fully vaccinated. Although not as accurate as the USA model, the Japan model is still incredibly accurate when compared with the MAE to the time series standard deviation of around 6.88% (Table 1). From Figure 13, we see that the model overestimates the values during the last few days. For Taiwan's ARIMA(p, d, q) model, we see that, on average, the predictions are off by about 451,903 doses from the actual total number of doses administered with the test data (Table 8). That gives an average absolute percent error of about 4.81% with the test data which is not as accurate as the USA and Japan models (Table 8), but still quite accurate when the MAE achieved is compared to the time series standard deviation of around 1,373,305 (Table 1). From Figure 13, we see that the model underestimates the values initially, but overestimates the values for the latter half of the test data. Finally, the ARIMA(p, d, q) model for China performs quite similarly to the Taiwan model, with an average percent error of about 4.26% (Table 8). Compared to the time series standard deviation of about 468,608,872 (Table 1), the model is quite accurate as well with an average absolute error of about 71,481,164 (Table 8). From Figure 13, we see that the model underestimates the values for most of the test data.

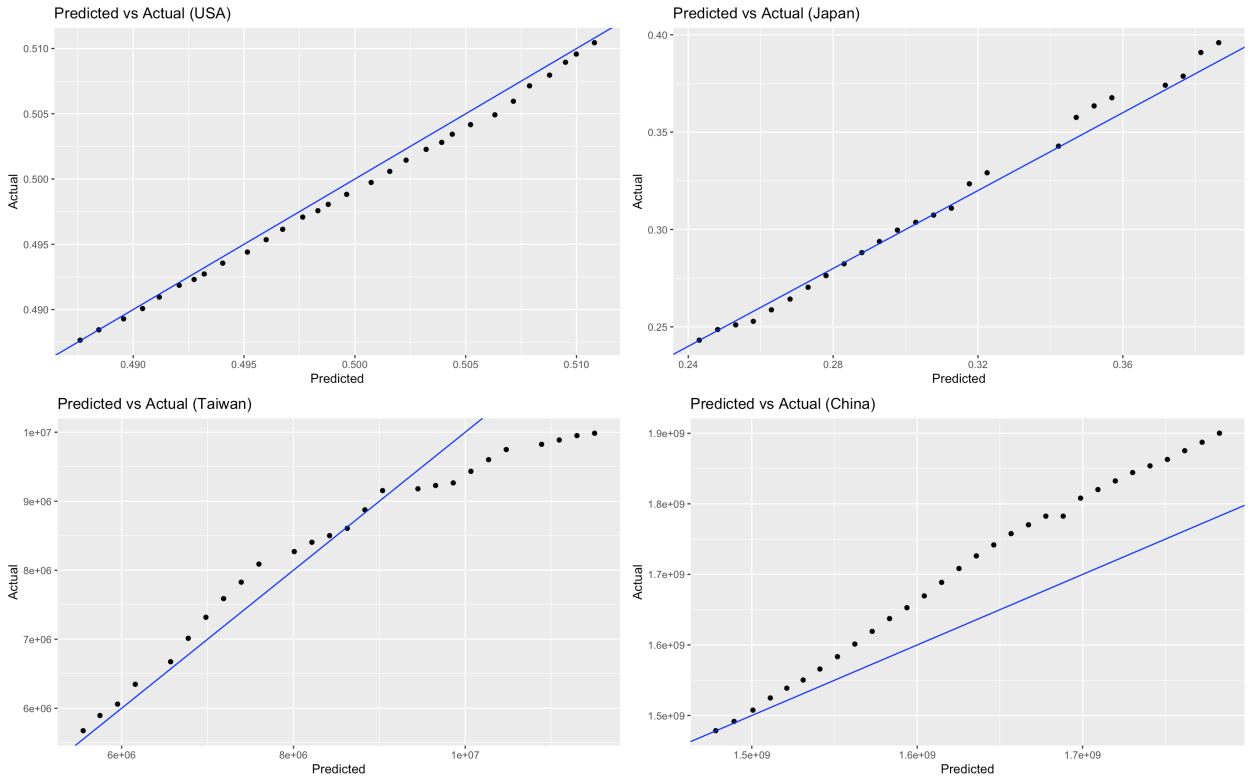

Figure 13: Predicted vs Actual values for the United States, Japan, Taiwan, and China from July 19, 2021 to August 18, 2021

5.2 Simple Linear Regression

We examine the accuracy of the ARIMA(p, d, q) models further by comparing them with the accuracy achieved by using simple linear regression (SLR) models. SLR models are often used in practice but they do not account for the data's dependence over time. Thus, SLR models provide good context and act as a baseline to understand the improvement in model accuracy and quality when using time series models like ARIMA. For each country from Table 8, we see that the ARIMA(p, d, q) models are much more accurate by examining and comparing the final MAPE and MAE scores for both models. Furthermore, Figure 14 gives a visualization of the comparison by plotting the predicted values from July 19, 2021 to August 18, 2021 for the ARIMA(p, d, q) models in blue, the SLR models in orange, and the actual values in red. Specifically, for each country, we see that the SLR model predictions become less and less accurate as the number of testing days increase while the ARIMA model predictions remain, for the most part, quite close to the actual values. It is important to note again that the SLR models are not statistically valid. Specifically, the time dependent nature of the data violates the independence assumption for simple linear regression.

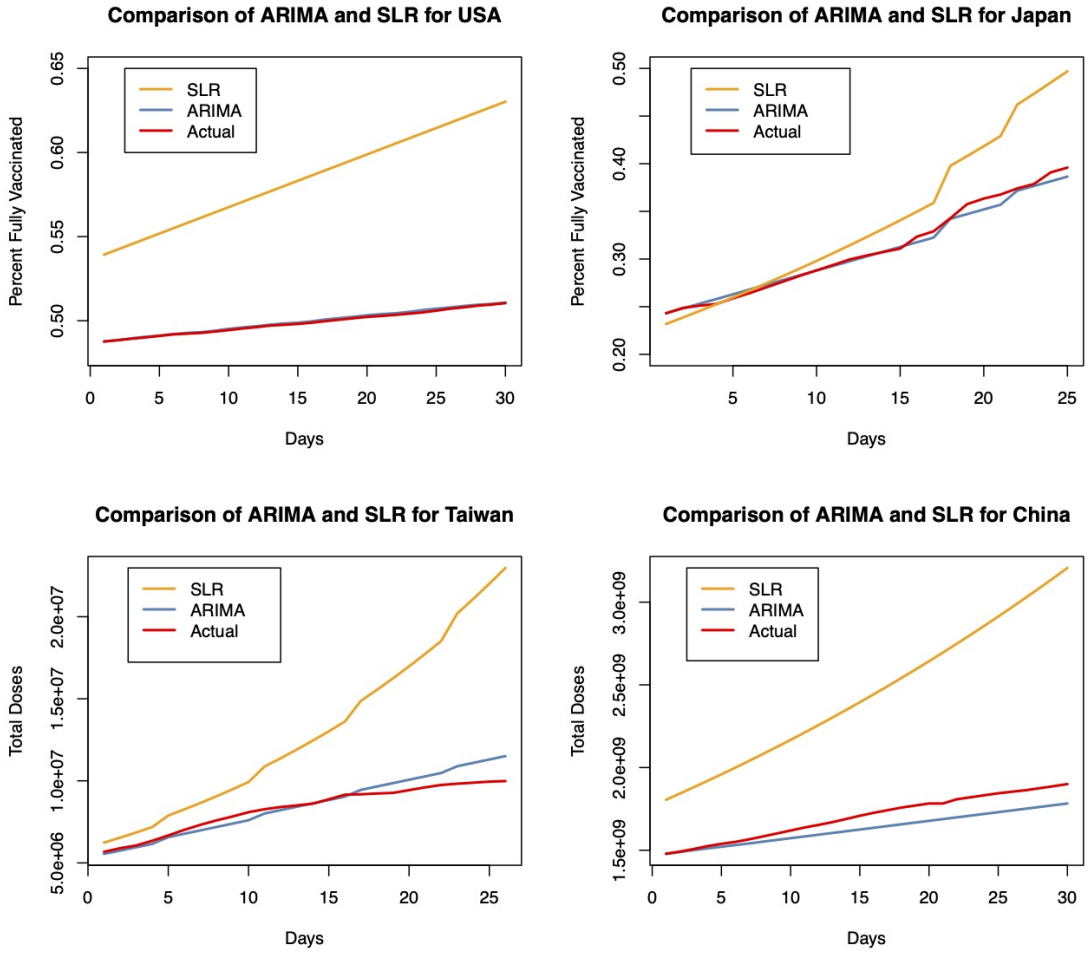

Figure 14: Comparison of ARIMA and SLR for the United States, Japan, Taiwan, and China from July 19, 2021 to August 18, 2021

5.3 Forecasts

For the United States, the model predicts that the country will have about 54% of its population fully vaccinated by September 17, 2021. This follows the recent trend of slowdown for the country in terms of vaccine distribution. Note that 51% of the population have been fully vaccinated as of August 18, 2021. While it is important that more than half of the population is fully vaccinated, this slowdown in vaccine distribution could be concerning as schools begin to open up around the country and the more contagious Delta variant continues to spread [31]. On the other hand, the ARIMA(p, d, q) model for Japan predicts that the country will have around 53% of the population fully vaccinated by September 17, 2021. Compared to the United States, this is incredible progress considering that about 39% of the Japanese population are fully vaccinated as of August 18, 2021 and that the country had a relatively late start in vaccine distribution. This aggressive push to distribute COVID-19 vaccines and have its citizens protected could be mostly attributed to the Olympics being held in Tokyo during the summer of 2021 in conjunction with a recent surge in COVID-19 cases around the country as mentioned in Section 1.

Forecasts from the ARIMA(p, d, q) for Taiwan show steady progress in the 30 days leading up to September 17, 2021 in terms of the total number of vaccine doses administered. With almost 12 million doses predicted to be administered by September 17, 2021, this progress seems to be mostly attributed to the consistent supply of vaccines that Taiwan has been receiving from foreign countries and the fact that Taiwan has begun to distribute their own vaccines as well [32]. Hopefully, this will help Taiwan reach its expectation of having 60% of the population receive at least one dose by October [33]. For China, the ARIMA(p, d, q) model forecasts that the country will make consistent progress in vaccine distribution with over 2.2 billion doses predicted to be administered by September 17, 2021. This is not surprising with China administering vaccines as early as July of 2020 [34]. Further, this indicates that China's contribution to the global vaccination efforts of COVID-19 will continue to be substantial.

5.4 Imputation

Certain assumptions must be made about the data set before imputation methods can be implemented. Specifically, the missing values in the data set must be either Missing Completely At Random (MCAR) or Missing At Random (MAR) [35]. Upon investigation of each data set and the corresponding vaccination distribution history for the United States, Japan, and Taiwan, we found that the missingness in each data set did not appear to be particularly correlated with any variables, observed or unobserved. That is, when we examined a missing value at a given time, t, we did not find any reason for the data to be missing at that time when we referred to the corresponding vaccination distribution history. Thus, we are then able to assume that the missing data for the United States, Japan, and Taiwan are MCAR.

However, it should be noted that the imputations using the simple moving average with the Taiwan data set failed to maintain the monotonic structure that is expected of the data. While a linear moving average would have helped maintain the monotonicity in the data, we opted to impute with the simple moving average instead because it performed better during the method selection phase of the imputation process. Furthermore, the resulting ARIMA model for the Taiwan data set was still able to produce monotonic forecasts with accuracy despite the non-monotonic imputations. That being said, it would be interesting and a possible next step in this research project to see how monotonic imputations using the linear moving average would affect the resulting forecasts of the ARIMA model.

Furthermore, there is quite a bit of uncertainty from the imputations of the China data set due to the fact that the entirety of its missingness occurs during the first half of the data set. This lack of data during the earlier periods of vaccination can be attributed to the fact that China started vaccinations much earlier than the rest of the world and did not have enough data to be reporting at a daily frequency. In other words, we can reasonably assume that the missingness is correlated with time, and hence missing at random (MAR) [35]. Furthermore, the data that was reported came in at a rather consistent frequency, and it did not exhibit a tremendous amount of activity. The inactivity, along with the MAR assumption and the fact that the data is monotonic, allows us to be confident in our decision to impute the missing values in the China data set. Thus, we opted to include the imputed data from the earlier time period in the ARIMA

model despite the uncertainty because we believed that the imputation could still provide information that can help model accuracy.

5.5 Model Assumptions

It is important to discuss the assumptions of ARIMA models, especially when applied to monotonic data as with the case here. Specifically, the ARIMA model does not assume the data to be monotonic, and thus the forecasted values are not guaranteed to be monotonic. Since the forecasted values produced from our models exhibited monotonicity and high accuracy, the lack of monotonicity in the ARIMA model assumptions did not present itself as a major concern with this particular data set. However, it is still worth exploring the possibilities with a model that explicitly assumes monotonicity. This could be a possible next step for this research project, and it may yield even better performance along with more robust assumptions.

6 Conclusion

In this paper, we have shown that the ARIMA(p, d, q) model can be exceptionally accurate when it comes to predicting COVID-19 vaccine distribution for the United States, Japan, Taiwan, and China. From the MAPE column in Table 8, we see that the ARIMA(p, d, q) models are, on average, able to make predictions within 5% of the actual value for each country. This forecasting power can be helpful for policy makers and healthcare workers to make decisions that would optimize COVID-19 vaccine distribution and protect people against this pandemic. For instance, by knowing how many vaccine doses will be administered in 30 days, Taiwan would be able to plan ahead and figure out how to properly distribute their limited supply of vaccines. Another example would be if Japan was able to know how much of their population would be fully vaccinated for the Olympics a month ahead of time. Having that prior knowledge perhaps would have led to better preparation for the quadrennial event.

Furthermore, the ARIMA(p, d, q) model, although a fundamental and time-tested model, is relatively simple compared to other tools in time series analysis. While incredibly accurate with predicting vaccine distribution, there is still plenty of room to improve. Specifically, the ARIMA(p, d, q) models for Taiwan and China are not quite as accurate as the models for the United States and Japan. From the MAPE column in Table 8, we see that the models for Taiwan and China make, on average, predictions within 4-5% of the actual value while the models for the United States and Japan make, on average, predictions within 0.15% and 1.5% of the actual value, respectively. This difference in accuracy can be quite large especially when considering that the total number of vaccine doses are in the millions for Taiwan and in the billions for China. Thus, a reasonable next step would be to try more time series models in order to see how well we can increase the accuracy of the predictions. For instance, the Long Short-Term Memory (LSTM) model is often used to forecast COVID-19 cases as well as other various time series data. The toolbox of time series analysis is quite vast, so there is potential for improving the accuracy of predictions for Taiwan and China.

Acknowledgements

We thank the anonymous reviewer for their time in reading our paper and providing insightful comments and feedback, particularly in clarifying the assumptions made from the imputation and model fitting.

References

[1] *Johns hopkins coronavirus resource center*, 2021, https://coronavirus.jhu.edu/.

[2] M. W. Tenforde, S. M. Olson, W. H. Self, et al., *Effectiveness of pfizer-BioNTech and moderna vaccines against COVID-19 among hospitalized adults aged ≥65 years — united states, january–march*

2021, MMWR. Morbidity and Mortality Weekly Report, 70 (2021), pp. 674–679, https://doi.org/10.15585/mmwr.mm7018e1.

[3] *The sinovac-coronavac covid-19 vaccine: What you need to know*, https://www.who.int/news-room/feature-stories/detail/the-sinovac-covid-19-vaccine-what-you-need-to-know.

[4] S. Pettypiece and R. Shabad, *'we did it': Biden celebrates u.s. hitting milestone of 200 million doses in his first 100 days*, Apr 2021, https://www.nbcnews.com/politics/white-house/biden-push-more-vaccinations-administration-reaches-200-million-dose-milestone-n1264782.

[5] U. B. Sayeed and A. Hossain, *How japan managed to curb the pandemic early on: Lessons learned from the first eight months of covid-19*, Journal of Global Health, 10 (2020), https://doi.org/10.7189/jogh.10.020390.

[6] A. Gunia, *Can japan protect itself from the olympics?*, Jul 2021, https://time.com/6079066/japan-olympics-covid-emergency/.

[7] Y. Tan, *Covid-19: What went wrong in singapore and taiwan?*, May 2021, https://www.bbc.com/news/world-asia-57153195.

[8] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, and W. Tan, *A novel coronavirus from patients with pneumonia in china, 2019*, New England Journal of Medicine, 382 (2020), pp. 727–733, https://doi.org/10.1056/nejmoa2001017.

[9] S. Mallapaty, *China is vaccinating a staggering 20 million people a day*, Jun 2021, https://www.nature.com/articles/d41586-021-01545-3.

[10] P. Cihan, *Forecasting fully vaccinated people against COVID-19 and examining future vaccination rate for herd immunity in the US, asia, europe, africa, south america, and the world*, 111 (2021), p. 107708, https://doi.org/10.1016/j.asoc.2021.107708.

[11] L. Moftakhar and M. Seif, *The exponentially increasing rate of patients infected with COVID-19 in iran*, Archives of Iranian Medicine, 23 (2020), pp. 235–238, https://doi.org/10.34172/aim.2020.03.

[12] S. P. Marbaniang, *Forecasting the prevalence of COVID-19 in maharashtra, delhi, kerala, and india using an ARIMA model*, (2020), https://doi.org/10.21203/rs.3.rs-34555/v1.

[13] G. Perone, *An ARIMA model to forecast the spread and the final size of COVID-2019 epidemic in italy*, (2020), https://doi.org/10.1101/2020.04.27.20081539.

[14] K. ArunKumar, D. V. Kalaga, C. M. S. Kumar, G. Chilkoor, M. Kawaji, and T. M. Brenza, *Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-regressive integrated moving average (ARIMA) and seasonal auto-regressive integrated moving average (SARIMA)*, Applied Soft Computing, 103 (2021), p. 107161, https://doi.org/10.1016/j.asoc.2021.107161.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020, https://www.R-project.org/.

[16] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, 2016, https://ggplot2.tidyverse.org.

[17] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, Sage, Thousand Oaks CA, third ed., 2019, https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

[18] R. J. Hyndman and Y. Khandakar, *Automatic time series forecasting: TheforecastPackage forR*, Journal of Statistical Software, 27 (2008), https://doi.org/10.18637/jss.v027.i03.

[19] D. Stoffer, *astsa: Applied Statistical Time Series Analysis*, 2021, https://CRAN.R-project.org/package=astsa. R package version 1.13.

[20] B. Hamner and M. Frasco, *Metrics: Evaluation Metrics for Machine Learning*, 2018, https://CRAN.R-project.org/package=Metrics. R package version 0.1.4.

[21] A. Trapletti and K. Hornik, *tseries: Time Series Analysis and Computational Finance*, 2020, https://CRAN.R-project.org/package=tseries. R package version 0.10-48.

[22] S. Moritz and T. Bartz-Beielstein, *imputeTS: Time Series Missing Value Imputation in R*, The R Journal, 9 (2017), pp. 207–218, https://doi.org/10.32614/RJ-2017-009.

[23] H. Ritchie, E. Mathieu, L. Rodés-Guirao, C. Appel, C. Giattino, E. Ortiz-Ospina, J. Hasell, B. Macdonald, D. Beltekian, and M. Roser, *Coronavirus pandemic (covid-19)*, Our World in Data, (2020), https://ourworldindata.org/coronavirus.

[24] *Population census 2020*, 2020, https://www.e-stat.go.jp/en/stat-search/files?page=1&layout=datalist&toukei=00200521&tstat=000001136464&cycle=0&year=20200&month=24101210&tclass1=000001136465&tclass2=000001154388&tclass3val=0.

[25] U. C. Bureau, *2020 census*, 2020, https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-main.html.

[26] M. Refaat, *Data preparation for data mining using SAS*, Elsevier, 2007.

[27] G. Box and G. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco, revised edition ed., 1976.

[28] S. E. Said and D. A. Dickey, *Testing for unit roots in autoregressive-moving average models of unknown order*, Biometrika, 71 (1984), pp. 599–607, https://doi.org/10.1093/biomet/71.3.599.

[29] F. G.E., M. Malcolm, and C. Moler, *Computer Methods for Mathematical Computations*, Wiley, 1977.

[30] S. Weisberg and J. Fox, *An R Companion to Applied Regression*, 01 2011.

[31] *Delta variant: What we know about the science*, 2021, https://www.cdc.gov/coronavirus/2019-ncov/variants/delta-variant.html.

[32] F. T. C. E. News, Aug 2021, https://focustaiwan.tw/society/202108020007.

[33] F. T. C. E. News, May 2021, https://focustaiwan.tw/society/202105280019.

[34] A. Woodyatt, *China says it's been vaccinating doctors and border workers since july*, Aug 2020, https://www.cnn.com/2020/08/24/asia/china-vaccine-doctors-workers-intl/index.html.

[35] D. B. RUBIN, *Inference and missing data*, Biometrika, 63 (1976), pp. 581–592, https://doi.org/10.1093/biomet/63.3.581, https://doi.org/10.1093/biomet/63.3.581, https://arxiv.org/abs/https://academic.oup.com/biomet/article-pdf/63/3/581/756166/63-3-581.pdf.