

Short Term Forecast of COVID-19 cases in Japan Using Time Series Analysis Models

Nor Azriani Mohamad Nor^{1*}, Nor Alyani Aziz², Azlan Abdul Aziz³, Wan Nurshazelin Wan Shahidan⁴, Siti Nor Nadrah Muhamad⁵

^{1,2,3,4,5} Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia.

Corresponding author: * norazriani@uitm.edu.my

Received Date: 20 July 2022

Accepted Date: 29 July 2022

Revised Date: 20 August 2022

Published Date: 1 September 2022

HIGHLIGHTS

- Forecast new daily confirmed cases of COVID-19 in Japan over a short-term period using Univariate Time Series Model.
 - The Univariate Model can very well capture the dependent structure of the daily new confirmed cases time series.
 - Four time series univariate models; Naïve, Mean, ARIMA and ETS models was compared to identify the most suitable model in forecasting the number of COVID-19 infected cases in Japan.
-

ABSTRACT

The new strain of coronavirus (COVID-19) was found to have started in Wuhan, China in late December 2019. The virus has spread to countries all over the world including Japan. The World Health Organization (WHO) declared COVID-19 as a pandemic on 11 March 2020 due to the increasing number of confirmed cases and deaths daily. The COVID-19 outbreak has impacted the nation of Japan adversely and the number of confirmed cases in Japan continues to increase day by day. On 7 April 2020, Japan declared a state of emergency to prevent the pandemic from worsening. This study is conducted to forecast new daily confirmed cases of COVID-19 in Japan over a short-term period. Four univariate time series models were applied: the Naïve Model, Mean Model, Autoregressive Integrated Moving Average (ARIMA) Model and Exponential State Space Model. This study analyses daily data from 22 January to 10 April 2020 collected from the Our World in Data website. The prediction involves five phases of data analysis and five different partitions of estimation and evaluation parts in every model to ensure the accuracy of forecast values. R and R Studio software were used in this study to analyze the data. The results reveal that Naïve model with 99 percent of estimation part and 1 percent evaluation part produces the lowest value of error measures for Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE).

Keywords: COVID-19, Naïve, Mean, ARIMA, Exponential State Space Model, R programming



INTRODUCTION

The new pandemic of Coronavirus disease (COVID-19) affects every country of the world. Starting in December 2019, the world witnessed the COVID-19 outbreak in a town in Wuhan, China. The outbreak became an epidemic, and the disease quickly spread beyond China's borders to the rest of the world. The Japanese Ministry of Health reported the first case of COVID-19 on 16 January 2020 in a person who returned from Wuhan in late December 2019. Japan declared a state of emergency on April 7 and as of May 27, 2020, Japan is the western Pacific's third-largest nation with a total of 16,651 confirmed cases involving 886 deaths.

Generally, the COVID-19 outbreak has negatively impacted Japan's society and economy since the number of confirmed cases in Japan increases daily. The economy of JAPAN shrank in the second quarter of the year by 7.8 percent, making it the worst recession in the history of the country. This was primarily due to the economic difficulties that were triggered by the pandemic of COVID-19 (Times, 2020). Many sectors are forced to be closed temporarily, the business performance was drop, events are being canceled and Summer Olympics scheduled to be held in Tokyo, are forced to be postponed due to the pandemic. Due to the loss of a significant number of profits, tourism and restaurant companies are suffering a big setback (Yamamura & Tsutsui, 2020).

A time series is a pattern that is recorded at regular time intervals. Much research uses the time series model to predict time series data. Chintalapudi et al. (2020) used the ARIMA model to predict registered and recovered cases after 60 days of incarceration in Italy. Aimran and Afthanorhan (2015) compared four exponential smoothing techniques to determine the best model for predicting the Malaysian population. Dehesh et al. (2020) used ARIMA models to predict new cases of COVID-19 in several countries, Italy, China, South Korea, Iran, and Thailand. Duan and Zhang (2020) use ARIMA model to forecast daily new confirmed cases of the COVID – 19 outbreaks in Japan and South Korea. They discovered that the estimated ARIMA model can very well capture the dependent structure of the daily new confirmed cases time series. Therefore, this paper compares four univariate time series models to identify the most suitable model in forecasting the number of COVID-19 infected cases in Japan. The findings will benefit for other parties such as the Government, the Ministry of Health and business operations in developing action plans to stop the spread of the COVID-19 outbreak.

METHODOLOGY

This study used secondary data of earlier of COVID-19 outbreak cases in Japan, from January 22 to April 10, 2020. It involves five phases as shown in Figure 1. The process starts with the data cleaning; to ensure the data is free from missing values or outliers that can affect the accuracy of the forecast values. Then the next stages were conducted until the final stage of this study, the best model will be used to predict short-term forecast of daily confirmed cases of COVID-19 in Japan. The forecast values will be compared to the actual data to see how well that model performs on unseen data and determine its accuracy.



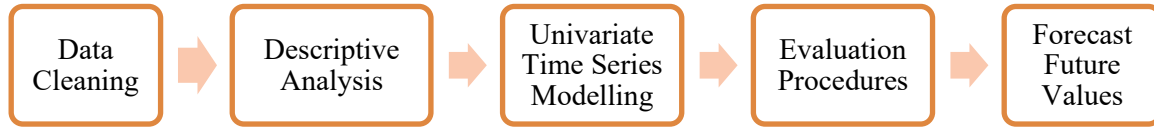


Figure 1: Phases of Data Analysis

Univariate Time Series Modelling

Four time series model were used in this study containing the Naive model, Mean model, the Box-Jenkins model and State Space Models for Exponential Smoothing (Hyndman et al., 2008). Naïve Forecast or Naïve Model was invented by Michael Gilliland (Snapp, 2012). The model is said to function successfully when there is no discernible pattern in historical data such as trend and fluctuation. Otherwise, the predictions value will be less accurate. The equation for Naïve Model is given in Equation 1.

$$F_{t+m} = y_t \quad (1)$$

where, F_{t+m} is the forecast values for m-step-ahead made at time t , m refers to the number of step-ahead forecast period ($m = 1, 2, 3, \dots$), and y_t is the last observation at time t .

The Mean Model assumes the expected value to be same with the average value of data series from the duration of data gathered. This model best performs when the data contains no discernible pattern, large fall, or growth. The general equation for the Mean Model is given by Equation 2.

$$F_{t+m} = \bar{y} \quad (2)$$

where F_{t+m} is the forecast value for m-step-ahead made at period t , \bar{y} refer to the mean of the actual historical time series.

Autoregressive Integrated Moving Average (ARIMA) Model also known as Box-Jenkins model, was developed by George Box and Gwilym Jenkins in 1976 (M.A. Lazim, 2018). The general term of ARIMA is written as ARIMA (p, d, q) where the terms p refers to the order of autoregressive, q for the order of moving average model and d for the number of differencing required to achieve stationary data. ARIMA Model is applied when the stationary assumption of variable is not met. The formula for ARIMA Model is shown in Equation (3)

$$y'_t = c + \phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (3)$$



where, y'_t is the differenced series, c is the intercept, $\phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p}$ are lagged values and $\theta_1 \varepsilon_{t-1} + \dots + \theta_p \varepsilon_{t-p} + \varepsilon_t$ are lagged errors.



Hyndman et al. (2002) proposed the State Space Models for Exponential Smoothing (ETS) framework, which included all 18 exponential smoothing models. This advancement simplifies forecasting by automatically generating prediction intervals, likelihood, and model selection criteria based on the model framework. Hyndman et al. (2008) depicts the model framework.

The common approach used in forecasting to evaluate forecast accuracy is splitting the data into two parts: estimation and evaluation (Lazim, 2018). Estimation data will be used to estimate the model parameters, while evaluation data will evaluate its accuracy. Five different sets of data partitioning will be used to ensure the accuracy of the forecast values.

Table 1: Five Sets of Data Partitioning for Estimation and Evaluation Part

Set 1	99% (n = 79)	1% (n = 1)
Set 2	95% (n = 76)	5% (n = 4)
Set 3	90% (n = 72)	10% (n = 8)
Set 4	85% (n = 68)	15% (n = 12)
Set 5	80% (n = 64)	20% (n = 16)

Data time frame: January 22 to April 10, 2020
 Total observations: 80

 Estimation part
 Evaluation part

Model Selection Criteria

The best model is based on a model that produce the smallest error measures calculated based on the out-of sample (evaluation part) forecast (M.A. Lazim, 2018). Five error measures were used are Mean Error (ME), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Scaled Error (MASE) as given in Equations (4), (5), (6), (7) and (8).

$$ME = \frac{\sum_{t=1}^n x_t}{n} \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^n e_t^2}{n}} \quad (5)$$



$$MAPE = \sum_{t=1}^n \frac{|(e_t / y_t * 100)|}{n} \tag{6}$$

$$MASE = \text{mean} \left(\frac{|e_j|}{\frac{1}{T-1} \sum_{t=m+1}^T |Y_t - Y_{t-m}|} \right) = \frac{\frac{1}{J} \sum_j |e_j|}{\frac{1}{T-m} \sum_{t=m+1}^T |Y_t - Y_{t-m}|} \tag{7}$$

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} = \frac{\sum_{i=1}^n |e_i|}{n} \tag{8}$$

FINDINGS AND DISCUSSIONS

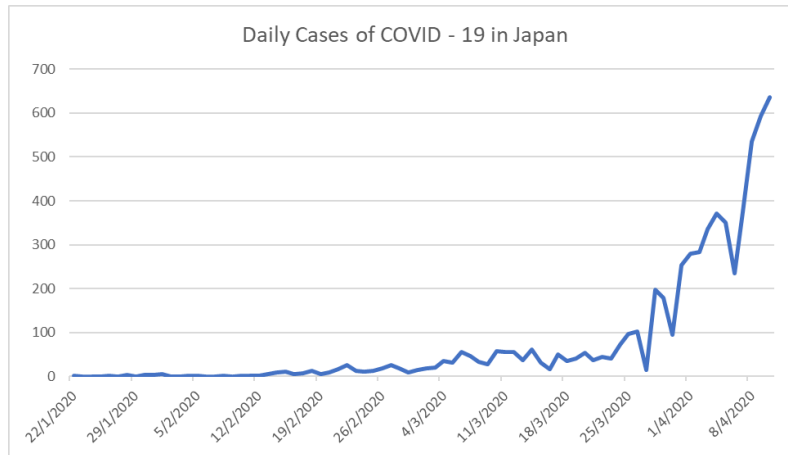


Figure 2: Daily cases COVID-19 case trends in Jepun

Figure 2 shows there is an upward trend with additive relationship of daily cases of COVID-19 in Japan. Data set from January 22 to April 10, 2020 with a total of 80 data demonstrate the maximum value is 636 while the minimum value is 0. This indicates that the highest new confirmed cases of COVID-19 in Japan (up to date data have been collected April 11, 2020) were 636 cases on 10 April 2020, and Japan once recorded no case of COVID-19 outbreak for the following date:

Table 2: Zero Recorded Case of COVID-19 in Japan

Month	Date
January	23/1/2020
	24/1/2020
	25/1/2020
	27/1/2020
	29/1/2020



February	2/2/2020
	3/2/2020
	6/2/2020
	7/2/2020
	9/2/2020

Four time series models of Naïve, Mean, ARIMA and ETS models are being analyzed into RStudio software. Table 3 and 4 present the model summary of 5 sets each model for estimation and evaluation part. Based on Table 3, all the models is well-fitted with the lowest error measure with Set 5 (80% for estimation and 20% of evaluation part). However, the “win” model or set of data partitioning will be selected based on the model that produces the smallest error measures in the evaluation part (M.A. Lazim, 2011). Therefore, based on Table 4, the best set for Naive model is Set 1, Mean model (Set 5), ARIMA model (Set 1) and Exponential State Space model (Set 4).

Table 3: Model Summary for Estimation Part

		Naïve Model					Mean Model					
	Data Partitioning	ME	RMS E	MA E	MAP E	MAS E	Data Partitioning	ME	RMS E	MA E	MAP E	MAS E
Estimation Part	Set 1	7.59	38.67	18.69	Inf	1.00	Set 1	0.00	122.91	82.58	Inf	4.42
	Set 2	3.09	30.04	14.64	Inf	1.00	Set 2	0.00	88.14	58.54	Inf	4.00
	Set 3	3.96	26.42	12.30	Inf	1.00	Set 3	0.00	61.67	39.25	Inf	3.19
	Set 4	2.64	15.48	8.94	Inf	1.00	Set 4	0.00	38.12	26.20	Inf	2.93
	Set 5	1.51	11.59	7.60	Inf	1.00	Set 5	0.00	21.90	18.06	Inf	2.37
		ARIMA Model					Exponential State Space Model					
	Data Partitioning	ME	RMS E	MA E	MAP E	MAS E	Data Partitioning	ME	RMS E	MA E	MAP E	MAS E
Estimation Part	Set 1	4.02	31.67	16.74	Inf	0.90	Set 1	5.65	37.13	18.78	Inf	1.00
	Set 2	3.05	29.84	14.45	Inf	0.99	Set 2	4.24	29.52	14.51	Inf	0.99
	Set 3	0.05	24.73	13.40	Inf	1.09	Set 3	2.68	21.12	11.48	Inf	0.93
	Set 4	0.01	14.86	8.77	Inf	0.98	Set 4	2.10	13.88	8.71	Inf	0.97
	Set 5	0.17	10.48	7.04	Inf	0.93	Set 5	2.13	10.97	7.07	Inf	0.93

Smallest error measure for each model, where Inf is denoted as Infinity.



Table 4: Model Summary for Evaluation Part

Naïve Model							Mean Model					
Data Partitioning	ME	RMS E	MAE	MAP E	MAS E		Data Partitioning	ME	RMS E	MAE	MAP E	MAS E
Evaluation Part	Set 1	42.00	42.00	42.00	6.60	2.25	Set 1	564.94	564.94	564.94	88.83	30.22
	Set 2	301.00	316.89	301.00	54.42	20.56	Set 2	480.92	491.03	480.92	89.47	32.85
	Set 3	146.00	197.13	158.25	32.41	12.87	Set 3	389.86	411.75	389.86	89.93	31.71
	Set 4	132.30	171.21	149.30	48.26	16.70	Set 4	283.26	303.39	283.26	89.06	31.68
	Set 5	124.10	155.62	124.70	45.77	16.40	Set 5	200.57	221.46	200.57	88.39	26.38
ARIMA Model						Exponential State Space Model						
Data Partitioning	ME	RMS E	MAE	MAP E	MAS E		Data Partitioning	ME	RMS E	MAE	MAP E	MAS E
Evaluation Part	Set 1	49.65	49.65	49.65	7.81	2.66	Set 1	105.55	105.55	105.55	16.60	5.65
	Set 2	301.00	316.89	301.00	54.42	20.56	Set 2	276.71	293.92	276.71	49.69	18.90
	Set 3	135.51	184.62	149.90	31.10	12.19	Set 3	5.14	84.68	70.67	19.48	5.75
	Set 4	161.81	215.77	30.97	19.81	0.54	Set 4	-50.76	89.30	59.53	27.38	6.66
	Set 5	234.60	282.66	234.60	67.77	30.86	Set 5	226.61	278.17	226.61	63.41	29.80

Smallest error measure for each model, where Inf is denoted as Infinity.

Then, a model that produces the lowest error measures will then be collected and compared to determine the best forecast, as shown in Table 5. The next step is to identify the best model out of four models used in this study. Again, the “win” model will be selected based on model that produce the lowest error measures values. Table 5 below is the comparison between four models that had been chosen as the ‘win’ models.

Table 5: Model Comparison Based on Four Models

Model	Naïve model	Mean model	ARIMA	Exponential State Space model
Evaluation Part	Set 1	Set 5	Set 1	Set 4
ME	42.0000	200.5687	49.6502	-50.7647
RMSE	42.0000	221.4609	49.6502	89.3001
MAE	42.0000	200.5687	49.6502	59.5317
MAPE	6.6038	88.3865	7.8066	27.3814
MASE	2.2469	26.3796	2.6562	6.6588




 Smallest error measure for each model

Table 5 show that the Naïve Model has the lowest error measures four out of five error measurements. Therefore, the Naïve Model has been selected as the best model and can be used to forecast future daily confirmed cases of COVID-19 in Japan.

Table 6: 3-steps ahead Forecast

Date	Forecast Values	95% CI	Actual new confirmed cases**	Forecast Accuracy
11 April	594	(486.8100, 701.1900)	701	84.74%
12 April	594	(462.7195, 725.2805)	522	86.21%
13 April	594	(442.4104, 745.5896)	300	32.89%

The predictions of 3-steps ahead forecast and the actual data of new confirmed cases in Japan from 11 to April 13, 2020, are shown in Table 6. The result produced using Naïve Model shows that the model is nearly identical to the actual data with predictive accuracy ranges from 32.89 percent to 86.21 percent. Therefore, this model is suitable for forecasting the new daily confirmed cases of COVID-19 in Japan in short term period.

CONCLUSION AND RECOMMENDATIONS

This study was conducted to predict the short term forecast for new daily confirmed cases of COVID-19 in Japan based on data starting from January 22, 2020 until April 10, 2020. We analyzed and generated the results of the Univariate Time Series Analysis models covering the Naïve Model, Mean Model, ARIMA Model and Exponential State Space Model.

For each model, five sets of data partitioning were used to ensure the accuracy of forecast values. Moreover, the calculation of five error measurements for ME, RMSE, MAE, MAPE, and MASE was crucial to determine the model performance, whereby the lower the value, the more efficient the forecasting model.

The Naive model is reliable for forecasting future new daily confirmed cases due to its high forecast accuracy. However, the forecast accuracy could change over time whenever there are additional or interruptions on the data. This implies that the model is sensitive to the fluctuations of the new daily confirmed cases of COVID-19. Therefore, we need more information to predict COVID – 19 cases over a long period, to increase the model’s accuracy.(Shaharudin et al., 2021)

This study only valid based on the dataset of COVID – 19 in Japan using the short-term data set (January 22, 2020 to April 10, 2020). Future studies should use long term data set (large sample size of data) and approach data partitioning using cross – validation technique to compare the performance and accuracy of the forecast model.



ACKNOWLEDGMENTS

The authors appreciate the reviewers for their contributions towards improving the quality of this research.

CONFLICT OF INTEREST DISCLOSURE

All authors declare that they have no conflicts of interest to disclose.

REFERENCES

- Aimran, A. N., & Afthanorhan, A. (2015). A comparison between single exponential smoothing (SES), double exponential smoothing (DES), holt ' s (brown) and adaptive response rate exponential smoothing (ARRES) techniques in forecasting Malaysia population. February. <https://doi.org/10.14419/gjma.v2i4.3253>
- Chintalapudi, N., Battineni, G., & Amenta, F. (2020). COVID-19 virus outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach. *Journal of Microbiology, Immunology and Infection*, xxx. <https://doi.org/10.1016/j.jmii.2020.04.004>
- Coronavirus Disease (COVID-19) - events as they happen.* (n.d.). World Health Organization. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen>
- Coronavirus disease (COVID-19) Highlights.* (n.d.). Retrieved July 10, 2020, from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200518-covid-19-sitrep-119.pdf?sfvrsn=4bd9de25_4
- Coronavirus disease (COVID-19) Situation Report -128 Highlights Situation in numbers (by WHO Region).* (n.d.). Retrieved July 10, 2020, from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200527-covid-19-sitrep-128.pdf?sfvrsn=11720c0a_2
- COVID-19 operations.* (n.d.). World Health Organization. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-operations>
- Dehesh, T., Mardani-Fard, H. A., & Dehesh, P. (2020). Forecasting of COVID-19 Confirmed Cases in Different Countries with ARIMA Models. *MedRxiv*, 2020.03.13.20035345. <https://doi.org/10.1101/2020.03.13.20035345>
- Duan, X., & Zhang, X. (2020). ARIMA modelling and forecasting of irregularly patterned COVID-19 outbreaks using Japanese and South Korean data. *Data in brief*, 31, 105779.
- Hyndman, R. J., & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, 27(3), 1–22. <https://doi.org/10.18637/jss.v027.i03>
- Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on 20 August 2022.



Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4), 679-688.

Japan: *WHO Coronavirus Disease (COVID-19) Dashboard*. (n.d.). Covid19. World Health Organization. Retrieved July 10, 2020, from <https://covid19.who.int/region/wpro/country/jp>

Novel Coronavirus (2019-nCoV) situation reports. (2019). World Health Organization. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/>

R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose (2002), “A state space framework for automatic forecasting using exponential smoothing methods,” *International Journal of Forecasting*, vol. 18, no. 3, pp. 439–454, 2002, doi: 10.1016/S0169-2070(01)00110-8.

R. J. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with Exponential Smoothing: The State Space Approach*. 2008. doi: 10.1007/978-3-540-71918-2.

Research & Analysis. <https://www.brightworkresearch.com/naive-forecast/>

Shaharudin, S. M., Ismail, S., Hassan, N. A., & Tan, M. L. (2021). *Short-Term Forecasting of Daily Confirmed COVID-19 Cases in Malaysia Using RF-SSA Model*. 9(June), 1–14. <https://doi.org/10.3389/fpubh.2021.604093>

Snapp, S. (2012, March 15). How to Best Understand the Naive Forecast. Brightwork

Times, N. S. (2020, August 18). *Japan suffers worst economic contraction in its history* | *New Straits Times*. NST Online. <https://www.nst.com.my/world/region/2020/08/617408/japan-suffers-worst-economic-contraction-its-history>

WHO,(2019). *Advice for public*. World Health Organization. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public>

World Health Organization. (2020, October 5). *COVID-19 disrupting mental health services in most countries, WHO survey*. [www.who.int](https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey). <https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey>

Yamamura, E., & Tsutsui, Y. (2020). The Impact of Postponing 2020 Tokyo Olympics on the Happiness of O-MO-TE-NA-SHI Workers in Tourism: A Consequence of COVID-19. *Sustainability*, 12(19), 8168. <https://doi.org/10.3390/su12198168>

