Predicting COVID-19 Confirmed Case in Surabaya using Autoregressive Integrated Moving Average, Bivariate and Multivariate Transfer Function

A Saikhu$^1$, C V Hudiyanti$^2$, J L Buliali$^3$, V Hariadi$^4$

Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

E-mail: 1saikhu@if.its.ac.id, 2cvairra@gmail.com, 3joko@if.its.ac.id, 4victor@if.its.ac.id

Abstract. In March 2020, the first case of Covid-19 was found in Indonesia. The increase of confirmed, suspected, and exposed in Surabaya has also significantly. Some studies show there is a relation among temperature, humidity, suspected, and exposed patients in an area with the number of confirmed COVID-19. Several statistical techniques that can be used to determine this relationship are to analyze and predict it using the ARIMA, bivariate, and multivariate transfer functions. The aim of this study is the performance of three models and determine the best model. The performance on the training data for ARIMA is 0.376, which shows that the accuracy of the model is 37.6%. The bivariate transfer function accuracy is 0.409, and the accuracy of the multivariate transfer function is 0.478. The result performance of ARIMA testing is 0.074, the bivariate transfer function is 0.055, and the multivariate transfer function is 0.108. The multivariate transfer function forecasting model is a technique in this case with the best performance.

1. Introduction
In recent, the Corona-virus disease (COVID-19) has become a serious concerned for public health around the world. Outbreaks of infectious diseases such as COVID-19 can also significantly affect many aspects of each region. March 2, 2020, the first case of COVID-19 was found in Indonesia, until now the number of confirmed, suspected, and exposed is still increasing. Since the beginning of the case, until the last data was obtained in August 2020, Surabaya is the second city with the highest number of COVID-19 confirmed, suspected, and exposed in Indonesia based on government data daily new cases of COVID-19 [1].

COVID-19 confirmed that patients are patients infected with COVID-19 with positive test results through the PCR Swab examination. A suspected is a person with an acute respiratory infection, i.e. fever (> 38°C) or history of fever, accompanied by any of the symptoms or signs of respiratory disease. It has are cough, shortness of breath, sore throat, runny nose, pneumonia mild to severe, and there were no other causes based on the conclusive clinical picture. In a suspected patient, 14 days before symptoms appear, patients have a travel history or live in a country or region that has a report of the occurrence of COVID-19 transmission. Exposed patients are people who have a fever (≥38°C) or a history of fever or respiratory system disorders. In the 14 days before symptoms appeared,
exposed patients had a travel history or lived in a country or region that had reports of COVID-19 transmission.

Based on government data, the number of confirmed, suspected, and exposed patients in the Surabaya has also increased significantly every day. Coronavirus transmission can be influenced by many factors, including climatic conditions, such as temperature (temp) and relative humidity (RH) [2].

Several previous research studies on the influence of climatic conditions and the spread of COVID-19 in a region, research by Fabiola is the influence of temperature, evaporation, rainfall, and regional climate on local transmission of the SARS-CoV-2 corona-virus in 31 states and Mexican capitals [3]. The statistical analysis obtained is the Local Transmission Ratio (LTR) calculated by region. Some days of infection effective since regional attacks in each state. The results obtained are the regional beginnings in dry climates appear earlier. It occurs because of lower temperatures than other regions and higher rainfall than tropical climates [3].

Research by Mazhar et al. analyzes the relation among COVID-19 and local climate parameters using existing global climate scales [4]. The results of the average day-time working hours correlate with the total cases of COVID-19 with a coefficient of determination of 0.42, and the average high temperature shows 0.59 and 0.42 with sum case of COVID-19 and case of deaths respectively [4]. Based on several previous research, it can be seen that using a variety of methods can be used to obtain the relation among regional climate variables and COVID-19 transmission in a region. In this research, we will predict the number of COVID-19 transmissions based on regional climate variables, i.e. temp and RH.

Previous research, predicting the transmission of COVID-19 using the LSTM network method in Canada, this research used Long Short Term Memory (LSTM) method which obtained the results that the transmission rate in Canada is a linear trend and in the USA it is exponential growth. One of these is another popular statistical technique and often used is Autoregressive Integrated Moving Model (ARIMA) [5]. Farajzadeh et al. are used the ARIMA model for forecasting monthly rainfall in the northwest from Iran [6]. ARIMA has become a valuable method for Multivariate phenomenon modelling, namely Vector Autoregression (VAR) rainfall forecast [6], Vector transfer function [7], and Spatio-temporal [8].

Based on the results of this research, this research compares the performance of three methods for predicting the number of COVID-19 confirmations, i.e. ARIMA, bivariate, and multivariate transfer functions method function which also in some research on other topics showed good performance in predicting time series data. We use Bivariate and multivariate transfer function methods to evaluate the effectiveness of adding external variables (independent variables) compared to ARIMA, which only uses confirmation cases variable. The dependent variable in this research is the confirmation case, and the independent variable is exposed COVID-19, suspect COVID-19, temp, and RH.

2. Data Set and Processing

2.1. Case Study
The observations daily are from March 23 until August 2, 2020. The number of data is 131 data. The distribution of the confirmed infected cases, suspect, and exposed COVID-19 by the COVID-19 in Surabaya was provided by COVID-19 Response Acceleration Task Force and available online https://covid19.go.id/peta-sebaran.

The climate data used in this research observation were obtained from the Meteorology, Climatology, and Geophysics Agency for the cities of Surabaya and Sidoarjo, Class I Juanda Meteorological Station, WMO-ID: 96935 (http://dataonline.bmkg.go.id/). The data used are daily data. This weather station observes a 30 km² area in Surabaya, Sidoarjo, and surrounding. The daily average temperature and humidity were used in our research to forecast confirmed cases.

The definition of these variables is: Confirmed is anyone who has tested positive for COVID-19 by the laboratory, suspected is anyone who has symptoms such as someone positive for COVID-19 but
not tested positive for Covid-19 by the laboratory, and exposed is anyone who has a history of contact with probable cases or confirmed cases of COVID-19 [9].

2.2. Data Preprocessing

Figure 1 shows the fluctuation in the number of COVID-19 confirmations. Based on the ACF and PACF correlograms in Figure 2 and Figure 3, it can be identified that the time series data already has a stationary-pattern. It is because the ACF chart has an exponentially decreasing pattern and the PACF chart has a 'cut-off' pattern after the 2nd lag [10].

The identification of stationary patterns can use testing, namely an augmented dickey-fuller test (adf.test). The test defines an alternative hypothesis that the time series data already has a stationary-pattern. The “adf.test” value is -3.5217 with a lag on order 15 and p-value 0.0432. The results show that the COVID-19 confirmation data has a stationary-pattern. The pre-processing required is to divide the total amount of data into two subsets, namely the training and the testing subset.set. The data will divide by a ratio of 70% for training and 30% for testing. The first subset is consists of 92 records and the second is consists of 39 records.

3. Forecasting Method

3.1. Autoregressive Integrated Moving Average

One of the statistical methods which are the most on forecasting or prediction research is the ARIMA model. Based on the research of Babu [11], the autoregressive model shows that there is a relation among a random value added by a and values in the past \((Z_{t-k})\) and value in the present \((Z_t)\). Moving Average model shows that there is a relation among a value in the present \((Z_t)\) and residuals in the past \((\epsilon_{t-k} = 1, 2, \ldots)\). ARIMA\((p,d,q)\) model is a mixture of AR\((p)\) and MA\((q)\) with \(d\) differencing order and a non-stationary data pattern. Equation (1) is the form of ARIMA\((p,d,q)\).

\[
\Phi_p(B)(1 - B)^d Z_t = \Theta_q(B) \epsilon_t
\]

(1)
where \( p \) is AR model order \( q \) is MA model order \( d \) is difference order and \( \Phi_p(B), \Theta_q(B) \) based on equation (2, 3).

\[
\Phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)
\]

\[
\Theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)
\]

Generalized ARIMA model for seasonal pattern data which is written as ARIMA \((p, d, q) (P D Q)s\) in equation (4).

\[
\Phi_p(B)\Phi_s(B^s)(1 - B)^d(1 - B^s)^DZ_t = \Theta_q(B)\Theta_s(B^s)a_t
\]

where \( s \) is a period of seasonal and \( \phi_p(B), \theta_q(B) \) based on equation (5, 6).

\[
\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)
\]

\[
\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)
\]

Flow-chart for the ARIMA model is described in Figure 4.

Figure 4. Flow-Chart of ARIMA modelling.

Figure 5. Flow-chart of transfer function modelling.

3.2. Bivariate Transfer Function (BTF)
The transfer function model is not the same method as the ARIMA model. In the time series model, the ARIMA model is a univariate model, but the transfer function is a multivariate model. So it can be concluded that the ARIMA model only connects the circuit with its past. In addition to the past series, the transfer function model also relates the series to other time series as described in Figure 5 and Figure 6. Other time series variables are used as predictor variables in the transfer function model.
Transfer function models can be used to model with one output and more than one output systems [6]. The first stage of pre-whitening is the dependent variable pre-accounting to eliminate the existing pattern. Dependent variable series so that there is only a white noise input series. Second, the independent variable pre-whitening to maintain the functional relationship on the transfer function and noise. The next stage is to detect and measure the power relation and by using the Cross-Correlation Function [12], input the value of $X_t$ into $Z_t$. However, the transformation on $Z_t$ does not have to convert $Z_t$ to white.

In the model with one output, only one equation is required to describe the model. It is showed to as a single-input transfer function model, as equation (7).

$$Z_t = C + \frac{\omega_s(B)}{\delta_s(B)} B^s X_t + N_t$$

where $N_t = \frac{\theta(B)}{\phi(B)} a^t$

$$\omega_s(B) = \omega_1 B + \omega_2 B^2 + ... + \omega_s B^s$$

$$\delta_s(B) = \delta_1 B + \delta_2 B^2 + ... + \delta_s B^s$$

$$\Phi_s(B) = \Phi_1 B + \Phi_2 B^2 + ... + \Phi_s B^s$$

with $b$ is a delay, $s$ is the numerator, and $r$ is the dominator.

### 3.3. Multivariate Transfer Function (MTF)

Like multiple regression models, transfer function models can also contain more than one input variable. In this paper, the single input and multi-input transfer functions are assumed to be stationary for the entire input and output series the general form of a single-input transfer function model in equation (9). We can apply the differencing technique if those variables are non-stationary.

$$Z_t = C + \frac{\omega_{s1}(B)}{\delta_{s1}(B)} B^{s1} X_{t1} + \frac{\omega_{s2}(B)}{\delta_{s2}(B)} B^{s2} X_{t2} + ... + \frac{\omega_{sm}(B)}{\delta_{sm}(B)} B^{sm} X_{mt} + N_t$$

### 3.4. Performance Model

The best model is evaluated using four performance measures, i.e. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), $R^2$ (R square) ($R_{sq}$) and Correlation ($r$) [13]. A good model is to have small RMSE and MAE and have a large $R_{sq}$ and $r$. The four performance measures are equation (10-13).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(Z_t - \hat{Z_t})^2}{n}}$$

$$MAE = \frac{\sum_{t=1}^{n}|Z_t - \hat{Z_t}|}{n}$$

$$R_{sq} = 1 - \frac{\sum_{t=1}^{n}(Z_t - \hat{Z_t})^2}{\sum_{t=1}^{n}(Z_t - \bar{Z})^2}$$

$$Correlation = r = \sqrt{R_{sq}}$$

### 4. Result and Discussion

#### 4.1. Autoregressive Integrated Moving Average (ARIMA)
Based on the flowchart of the ARIMA model (Figure 4), the confirmed COVID-19 data is identified, it has been stationary using time-series plots. The COVID-19 confirmation data is stationary according to figure 1 because the time-series plot shows that the pattern is a fluctuating in average-value.

Secondly, we identified the stationarity time-series based on ACF and PACF pattern. The ACF correlogram is shown in Figure 4. Events have an exponential decrease which is thought to be a non-seasonal model of AR. Figure 2 and Figure 3 shows that the confirmation COVID-19 data is stationary because PACF has been cut off after the 2nd lag. No transformation is needed nor difference.

Confirmed-data is identified as AR(2), ARMA(1,1), MA(2), etc. Based on the ACF and the PACF, several ARIMA models are proposed \((p d q)(P D Q)^d\) using training data. Table 1 shows the six significant models in which the best model is the third model as an equation (14):

\[
y_t = 6.693 + 0.297 y_{t-1} + 0.47 y_{t-2} + e_t
\]

This model has the least RMSE and the highest \(R_s^2\). Therefore it was selected as the best model.

The PACF correlogram of the residual model (2 0 0) has a white-noise pattern so that the model is acceptable. White noise means that the partial correlation value of the residuals is not significant; it can be seen from the PACF value not exceeding the UCL-LCL limit, according to figure 7. In figure 8, from the forecasting results, the graph that formed on the forecasting plot does not represent a graph similar to actual testing data; this results in low accuracy values. Table 2 shows that \(R_s^2\) value is a few.

![Figure 7. PACF Correlogam of COVID-19 confirmation cases of error of the third model.](image)

![Figure 8. Forecasting plot-based model 1.](image)

| Table 1. Performance of ARIMA based on training data. |
|-----------------------------------------------|
| Performance of ARIMA                          |
| Model  | RMSE    | MAE      | \(R_s^2\) | \(r\)   |
|--------|---------|----------|----------|--------|
| (2 0 0) | 49.884  | 32.499   | 0.376    | 0.613  |
| (1 0 2) | 51.009  | 32.856   | 0.355    | 0.596  |
| (1 0 1) | 51.775  | 29.887   | 0.330    | 0.574  |
| (1 0 0) | 58.132  | 39.693   | 0.146    | 0.382  |

| Table 2. Performance of ARIMA based on testing data. |
|-----------------------------------------------|
| Performance of ARIMA                          |
| Model  | RMSE    | MAE      | \(R_s^2\) | \(r\)   |
|--------|---------|----------|----------|--------|
| (2 0 0)(1 0 0)^6 | 61 | 47.8 | 0.074 | 0.272 |

4.2. Bivariate Transfer Function (BTF)

The prediction of the model is improved by adding one predictor variable. This model is called the bivariate input transfer function. The independent variables used by the predictors are carried out alternately to get the best BTF model.
According to figure 10, the number of exposed does not affect the number of confirmations, so it is not suitable for BTF. Cross-Correlation Function (CCF) from suspected variable to COVID-19 confirmed has a relationship. It is proven that the BTF (suspected and confirmed) model is significant. Partially, the other three variables influence the number of confirmed cases.

The best predictor is the suspected variable. The increase in the performance of BTF is better than ARIMA, so it can be concluded that the BTF model is better. From figure 9, plot “training – testing”: among actual and predicted for the BTF (suspected – confirmed) model is better than ARIMA, and then the graph can follow fluctuations in the observed data. Table 3 shows that the suspect variable is the best predictor for Covid-19 Confirmed in the BTF model. While the testing accuracy is still low, as shown in table 4.

![Figure 9. Forecasting plot Bivariate Transfer Function.](image)

![Figure 10. The cross-correlation function of exposed and confirmed.](image)

![Figure 11. The cross-correlation function of suspected and confirmed.](image)

| Table 3. Performance of BTF based on training data. |
|----------------------------------------------------|
| Performance of BTF |          |          |          |          |
| Input               | RMSE    | MAE      | R^2      | r        |
| Expose (0 0 0)      | -       | -        | -        | -        |
| Suspect (1 0 1)     | 50.309  | 31.562   | 0.409    | 0.639    |
| Temp (2 0 0)        | 49.407  | 34.491   | 0.397    | 0.630    |
| RH (1 0 1)          | 49.490  | 31.244   | 0.388    | 0.618    |

4.3. Multivariate Transfer Function (MTF)

The third model is MTF that the predictor variables as input are four variables. The results of the training process for the MTF model are shown in table 5. The exposed variable is not as significant as a predictor based on the BTF model.

Meanwhile, the suspected, temperature and humidity variables are significant as predictors. The training results show that the performance of the MTF model is quite good. In comparison, the performance of the testing process is shown in table 6 with a low R^2-value. Figure 12 shows that the training results are close to the actual value, but the test results still cannot follow the fluctuation of real data.
Table 4. Performance of BTF based on testing data.

| Performance of BTF | Input | RMSE | MAE  | $R_{sq}$ | $r$  |
|--------------------|-------|------|------|----------|------|
| Suspect            | 50    | 35.59| 0.055| 0.234    |      |

Table 7 is a summary of the performance comparison of the ARIMA, BTF, and MTF models. These results indicate that the MTF has better performance than the other two models.

Table 5. Performance of MTF based on training data.

| Performance of MTF | Input | RMSE | MAE  | $R_{sq}$ | $r$  |
|--------------------|-------|------|------|----------|------|
| Exposed            | 46.544| 32.333| 0.478| 0.691    |      |
| Suspect (0 0 2)    |       |      |      |          |      |
| Temperature (0 1 0)|       |      |      |          |      |
| Humidity (0 2 0)   |       |      |      |          |      |

Table 6. Performance of MTF based on testing data.

| Performance of MTF | Input | RMSE | MAE  | $R_{sq}$ | $r$  |
|--------------------|-------|------|------|----------|------|
| Suspect, Temp, RH  | 46    | 34.87| 0.108| 0.328    |      |

Table 7. Performance of 3 models based on testing data.

| Performance 3 Model | RMSE | MAE  | $R_{sq}$ | $r$  |
|---------------------|------|------|----------|------|
| Training            |      |      |          |      |
| ARIMA               | 49.884| 32.499| 0.376    | 0.613|
| BTF                 | 50.339| 31.560| 0.409    | 0.639|
| MTF                 | 46.544| 32.333| 0.478    | 0.691|
| Testing             |      |      |          |      |
| ARIMA               | 61   | 47.8 | 0.074    | 0.272|
| BTF                 | 50   | 35.59| 0.055    | 0.235|
| MTF                 | 46   | 34.87| 0.108    | 0.328|

5. Conclusion
The ARIMA model has the worst training and testing performance because it cannot follow the actual data pattern. The smaller accuracy value also evidences this than the other two methods, the $R_{sq}$ value of 0.376 for training and 0.074 for testing. BTF model has better accuracy than ARIMA because the relative test results can follow the real data pattern. However, it has a small correlation, the $R_{sq}$ value of 0.409 for training and 0.055 for testing. The best model is MTF because it has better accuracy than
the other two models for training and testing processes. The result of training is 0.478, and training is 0.108 for the $R^2_q$ value obtained. Statistical-based modeling of COVID-19 confirmation cases is recommended to use MTF if external variable observations are available.

Recommendations that can be given for future research are that researchers can add a lot of data and can try using other methods that can be used to predict such as Recurrent Neural Network (RNN) or Long Short-term Memory (LSTM).

Acknowledgements
This research and publication were funded by Penelitian Dana Departemen, a grant from Department of Informatics, ITS, Indonesia, the budget year 2020 (No: 850/PKS/ITS/2020).

References
[1] Gugus Tugas Percepatan Penanganan COVID-19 2020 Peta Sebaran COVID-19 Retrieved from https://covid19.go.id/peta-sebaran-covid19
[2] Wang J, Tang K, Feng K, and Lv W 2020 High temperature and high humidity reduce the transmission of covid-19 SSRN Electronic. J. 3551767
[3] Méndez-Arriaga F 2020 The temperature and regional climate effects on communitarian COVID-19 contagion in Mexico throughout phase 1 Sci. of the Total Environ. 735 139560
[4] Iqbal M M, Abid I, Hussain S, Shahzad N, Waqas M S, and Iqbal M J 2020 The effects of regional climatic condition on the spread of COVID-19 at global scale Sci. of the Total Environ. 739 140101
[5] Chimmula V K R and Zhang L 2020 Time series forecasting of COVID-19 transmission in Canada using LSTM networks Chaos, Solitons and Fractals 135 109864
[6] Farajzadeh J, Fakheri Fard A, and Lotfi S 2014 Modeling of monthly rainfall and runoff of Urmia lake basin using “feed-forward neural network” and “time series analysis” model Water Resour. Ind. 7–8 pp 38–48
[7] Nugroho A, Subanar, Hartati S and Mustofa K 2014 Vector Autoregression (Var) Model for Rainfall Forecast and Isohyet Mapping in Semarang – Central Java – Indonesia Int. J. Adv. Comput. Sci. and Appl. 5 pp 44–9
[8] Fuentes M, Reich B and Lee G 2008 Spatial–temporal mesoscale modeling of rainfall intensity using gage and radar data Annals of Appl. Stat. 2 pp 1148–69
[9] World Health Organization 2020 World Health Organization Retrieved from https://www.who.int/
[10] Saikhu A, Arifin A Z and Fatichah C 2017 Rainfall Forecasting by Using Autoregressive Integrated Moving Average, Single Input and Multi Input Transfer Function Int. Conf. on Information, Communication Technol. and System (ICTS)
[11] Babu C N and Reddy B E 2014 A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data Applied Soft Computing 23 pp 27–38
[12] Remple R K and Tischler M B 2006 Aircraft and Rotorcraft System Identification (Virginia: American Institute of Aeronautics and Astronautics, Inc.)
[13] Chai T and Draxler R R 2014 Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature Geosci. Model Dev. 7 pp 1247–50