Real time prediction of four main food commodities in Indonesia and the mapping based on autoregressive integrated moving average model

R F Dianco and M Novita
Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia

Abstract. Based on the United Nations Global Pulse, food prices have a direct effect on the purchasing power of a large part of the Indonesian population. Hence, it is important to maintain stable food prices. One way to do this is by making a prediction model. Using 59 weekly rice, shallot, chicken and egg prices starting from the first week in 2018 as the cornerstone, this research uses an Autoregressive Integrated Moving Average (ARIMA) model to predict the weekly prices of these food commodities in all Indonesian provinces. This research also provides real-time prediction that can be automatically updated when any new data is inputted. Finally, this research provides a mapping visualization to make it easier for people to interpret the results. This map is equipped with a dynamic line chart to compare two provinces food trendline and shows the current price, one week and two weeks prediction. All of this is built using R 5.3.2. In this research, the error is calculated from the coefficient of variation and the result is 0.58 %, 4.1 %, 3.23 % and 2.76 % for rice, shallot, chicken, and egg weekly prices, respectively. Furthermore, this research also analyses the prediction of second-week prices of rice, shallot, chicken, and eggs in March 2019. Hopefully, this research will bring benefit to the government and farmers to make a better decision based on these predictions, so that in the future, food prices will stabilize.

Keywords: ARIMA, coefficient of error, food prices, map, prediction

1. Introduction
Based on the United Nations Global Pulse [1], food prices have a direct effect on the purchasing power of a large part of the Indonesian population. Rising food prices severely affect poor people who spend 75 % of their total spending, by making them more vulnerable to hunger and also making it harder to escape from the chain of poverty. The importance of the stability of food prices is also mentioned in Sustainable Development Goals which made the stability of food prices as the main indicator to measure the success of second goals, Zero Hunger. Unfortunately, in Indonesia, food prices have not been controlled properly yet. Based on Statistics Indonesia [2], food commodities experienced an inflation rate of 3.41 % in 2018, three times more than what happened in 2017. As seen from Indonesia main food sources, in January 2010 rice price was 6.702 rupiah/kg. Since then, it has increased to 12.211 rupiah/kg.

From the explanation above, the need for a food price prediction model is essential. By utilizing this, future uncertainty can be reduced. The government as a regulator can make a better decision in import,
export, and even distribution of commodities within a certain region. Furthermore, farmers and stockbreeders can control the production volume such that no food is being wasted because of low demand, nor is there a product void because of high demand.

Some related research includes that which was conducted by United Nations Global Pulse, Kim et al. and Janvier et al. United Nations Global Pulse in 2014 has tried to monitor the food prices in Indonesia using Twitter data to predict the inflation related to food [1]. The same attempt has also been done by Kim using Twitter data to predict beef, chicken, onion and chili prices [3], but the number of people who discuss such topics on Twitter is limited. In 2016, Janvier et al. conducted research on predicting the unit price of coffee exports in Rwanda just by using the past data by utilizing auto-regressive integrated moving average [4].

To further develop the research conducted by Janvier et al. this paper will create a food prediction model and focused on predicting weekly rice, shallot, chicken, and egg prices for every Indonesian province. This paper will also use auto-regressive integrated moving average to make that prediction. This research also provides a real-time prediction that can be automatically updated when any new data is inputted. Finally, this research provides a mapping visualization to make it easier for people to interpret the results.

2. Materials and method

2.1. Data

Ministry of trade Republic of Indonesia has published the regulation of ministry of trade Republic of Indonesia 27/M-DAG/PER/5/2017 that stated the reference prices for nine main food commodities: rice, corn, soy beans, sugar, cooking oil, shallot, beef, chicken and eggs [5]. The reference prices are made for these nine food commodities to ensure the stock, stability, and certainty of the prices. It is also stated that the ministry can order the stated owned enterprise to make some purchases and sales in accordance with the reference price for the creation of price stability. This paper will use 59 weekly rice, shallot, chicken, and egg prices starting from the first week in 2018 provided by National Strategic Food Price Information Center as the cornerstone for an Autoregressive Integrated Moving Average (ARIMA) to predict the weekly prices of these food commodities in all Indonesian provinces [6]. These four commodities are chosen because the Indonesia Bank state that these four commodities influence the value of inflation the most. The reference price for these commodities are Rp 9,500/kg, Rp 32,000/kg, Rp 32,000/kg and Rp 22,000/kg for rice, shallot, chicken and egg prices respectively.

The figure 1a show the trend of weekly chicken prices in Bengkulu (yellow) and Lampung (blue) from the first week of 2018. This figure is shown to give a general idea on how the prices change over time. While these provinces share the same border, the price of chicken in these two provinces have different trends. Sometimes the prices in Lampung will increase, while the prices in Bengkulu decrease. It would be nice if this were to happen, where the province with decreasing prices can subsidize the other one with increasing prices, and steadily approaching the reference price. The same conclusion can be concluded from figure 1b which show the trend of weekly egg prices in Bengkulu (yellow) and Lampung (blue).

2.2. Autoregressive integrated moving average

ARIMA model is focused on revealing the correlation, the degree of similarity between the given time series and the past version of itself. Janvier et al. said that ARIMA is the most popular method in analyzing time series data because of its power and model availability [4]. ARIMA has been used to predict gold prices, wind speeds, and inflation; in these cases, it was proven to give satisfactory results [7–9]. ARIMA is formulated by

\[ W_t = \mu + \varphi_1 W_{t-1} + \varphi_2 W_{t-2} + \cdots + \varphi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_p e_{t-p} - \cdots \]  

(1)
From the formula, ARIMA has three parameters that need to be adjusted, which are $p, d, q$ which respectively represent how many previous consecutive data ($W$) that explain current data; how many differencing $\Delta Y_t = Y_t - Y_{t-1}$ is done to achieve stationary; and how many previous consecutive error ($e$) that explain the current data. The error has constant variance $\sigma_e^2$. Lastly, $\mu$ is the mean of the time series, $\Phi$ and $\Theta$ respectively denote the coefficient for $W$ and $e$ for each time $t$.

ARIMA is the development of ARMA (autoregressive moving average) model, which is in turn the development of the AR (autoregressive) and MA (moving average) models. The AR model cannot utilize the correlation measured in error terms, whereas the MA model cannot utilize the correlation measured in the time series data and hence triggers the development of ARMA model (which utilizes both). However, the ARMA model requires stationary data, which is not always the case for time-series data. Therefore, to account for this, differencing is required; therefore, the ARIMA model is created, and this model can utilize the correlation that is measured in error and time series data and can handle non-stationer data.

$$-\theta_q e_{t-q}, \quad W_t = \Delta^d Y_t, \quad \Delta Y_t = Y_t - Y_{t-1}.$$  

Figure 1. (a) Chicken and (b) egg weekly prices in Bengkulu (yellow) and Lampung (blue).
2.3. ARIMA parameterization

ARIMA model is developed in three stages as explained by Box and Jenkins, which are identification, estimation and diagnostic checking. ARIMA parameter will be specified at identification step [10]. Usually, this step will result in two or more models. All coefficients in these models will be estimated in the estimation step, and one model will be selected based on either the AIC or BIC value (the lower the better). In the last step, which is diagnostic checking, the model assumption will be checked.

The first parameter in the ARIMA model that must be specified is $d$, which is obtained by differencing the data step by step until stationary data is achieved. This characteristic is checked by performing an Augmented Dickey-Fuller test which checks if the data has a unit root of not (indicator of non-stationary data).

The $p$ and $q$ parameter of ARIMA is specified by checking the extended sample autocorrelation (EACF). The idea of EACF is checking the error component of the model by firstly removing the component of the $W$ component using iterative regression [11]. After this iterative regression process, the model is formulated by:

$$ W_{t,k,j} = Y_t - \hat{\phi}_1 Y_{t-1} - \cdots - \hat{\phi}_k Y_{t-k}. $$

EACF is defined as the sample correlation of $\{W_{t,k,j}\}$. If $k$ is the value of the “right” $p$, then the time series process $\{W_{t,k,j}\}$ become ARIMA($-\cdot\cdot\cdot\cdot\cdot$), and thus the correlation value for $q + 1$ time-lag or more will become 0. Based on the parsimony principle, the value of $p$ is chosen to be smallest and the $q$ is chosen if the correlation value for $q + 1$ time-lag or more is 0. This fact is visualized by the EACF plot. Figure 2 shows the example for EACF plot. The model is selected by choosing the 0 which is furthest upper left (yellowed). Sometimes, EACF plot will give two value of $(p,q)$. To handle this, the parameter that give lowest value of AIC and BIC is selected.

To estimate the other parameters of the model ($\theta$, $\phi$, $\mu$, and the $\sigma^2_{\epsilon}$) given the parameters $p$, $d$, $q$, maximum likelihood estimation will be used. Here, denote $L$ as the likelihood function of ARIMA models of the $\theta$, $\phi$, $\mu$, and the $\sigma^2_{\epsilon}$ given the observations $Y_1,Y_2,...,Y_n$. The maximum likelihood estimators are the values of the parameters ($\theta$, $\phi$, $\mu$, and the $\sigma^2_{\epsilon}$) which make the observant data the most likely data to be observed.

ARIMA model will be diagnosed by first checking the assumptions of ARIMA which are as follows:

- **Stationary**
  A model is stationary if the mean of the series is constant for arbitrary time ($t$) and the autocovariance is constant for given time-lag. This characteristic is checked by performing Augmented Dickey-Fuller test where the null hypothesis states that the data is not stationary and the alternative hypothesis states that the data is stationary [12]. Clearly, much real-time series data cannot be reasonably modelled by stationary processes since they are not in statistical equilibrium, but are evolving over time. Hence, to achieve a stationary state, the data must be transformed. The simplest transformation that can be done is differencing. Thus, the assumption is automatically fulfilled if the $d$ (number of differencing conducted) value is correct.

- **Error is normally distributed**
  This assumption is checked by performing Shapiro Wilk on the model’s residual, where the null hypothesis states that the error of the model is normally distributed, and the alternative hypothesis states that the error is not normally distributed [13]. This assumption is important to specify the confidence interval for ARIMA coefficient which are $\Phi$ and $\Theta$, and also to ensure that the model performance is consistent to predict future data for arbitrary time.

- **The error is not correlated**
  The uncorrelated error is the indicator that the chosen model is the best one since there is no additional information can be extracted from the data or the error. To test this assumption,
Ljung Box test is performed on the model’s residual, where the null hypothesis states that the error is not correlated, and the alternative hypothesis states that at least there is some $k$ such that the error and past $k$ error is correlated [14].

After these assumptions have been checked, the model will be checked for any overfitting. To check this, the more general model will be made and will be checked; that is, a model close to (and with) the characteristics of the original model. For example, if an ARIMA (2, 0, 1) is the original model, ARIMA (3, 0, 1) and ARIMA (2, 0, 2) will be checked. For ARIMA (3, 0, 1), if the additional parameter, $\varphi_3$, is not statistically significant from zero and the value of $\varphi_1, \varphi_2$ and $\theta_1$ is not too different from the original model, then the original model is good. Accordingly, for ARIMA (2, 0, 2), if the additional parameter, $\theta_2$, is not statistically significant from zero and the value of $\varphi_1, \varphi_2$ and $\theta_1$ is not too different from the original model, then the original model is good.

2.3.1. Forecasting with ARIMA. This paper will use minimum mean square error method to forecast the future value. This method works by using the past data, $Y_1, Y_2, ..., Y_{t-1}, Y_t$ to forecast $Y_{t+1}$ where $l$ is called lead time and $t$ is called forecast origin. The formulation of such forecast is as follows:

$$\hat{Y}_t(l) = E(Y_{t+l} | Y_1, Y_2, ..., Y_t).$$

(3)

Since our data is time-series data, the future data will eventually be known. Hence, the model’s prediction must be adjusted to the data. To do such things, the following formula is used:

$$\hat{Y}_{t+1}(l) = \hat{Y}_t(l + 1) + \psi_t[Y_{t+1} - \hat{Y}_t(1)]$$

(4)

where $\psi_t[Y_{t+1} - \hat{Y}_t(1)]$ is the true error for $t+1$.

2.4. Method

From the explanation before, the ARIMA model will be built in the steps shown in figure 3. Using these steps, 136 ARIMA models will be created. R 5.3.2 program and dplyr, shiny, lazyeval, rgdal, MTS, urca, TSA, tseries, forecast, RVAideMemoire, ggplot2, and editData packages will be used to build the ARIMA model and the visualization.

| AR/MA | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-------|---|---|---|---|---|---|---|---|---|---|----|
| 0     | X | X | X | X | X | X | X | X | X | X | X  |
| 1     | X | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 2     | X | X | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 3     | X | X | X | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 4     | X | X | X | X | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 5     | X | X | X | X | X | 0 | 0 | 0 | 0 | 0 | 0  |
| 6     | X | X | X | X | X | X | 0 | 0 | 0 | 0 | 0  |
| 7     | X | X | X | X | X | X | X | 0 | 0 | 0 | 0  |

Figure 2. EACF
3. Results and discussion

In this paper, the error is calculated from the coefficient of variation of error (mean of error divided by mean of respective price) from two last periods for each commodity and the result is respectively 0.58 %, 4.1 %, 3.23 % and 2.76 % for rice, shallot, chicken and egg weekly prices.

3.1. Map visualization

This research provides a mapping visualization to make it easier for people to interpret the results. This map is equipped with a dynamic line chart to compare two provinces food trendline and shows the current price, one week and two weeks prediction. This map also equipped with a color marker that indicates the prices or the changes where the redder the color is, the higher the price is. Furthermore, this map can show the distribution line for each commodity. For the purpose of illustration, this paper will show some insight that can be obtained from the map. Only two commodities will be shown in this paper: shallots and chicken.

From figure 4, every province in Java and Sumatera island is expected to experience an increase in shallot prices from Rp 200/Kg up to Rp 1.000/Kg for Sumatra and from Rp 1.000/Kg up to Rp 5.000/Kg. Except for Bangka Belitung, DKI Jakarta, and Aceh, these increases in onion prices are good because the shallot prices slowly approach the reference price set by the government, but it must still be evaluated by the government. Furthermore, East Java can reduce the shallot commodity transported to Nusa Tenggara Barat, Central Kalimantan, and North Sulawesi since these three provinces are expected to experience a decrease in shallot prices. Looking to Sulawesi, shallot merchants in West Sulawesi can sell shallots to Central Sulawesi and West Sulawesi where the shallot price is expected to increase. By doing this, the merchants can make more profits while stabilizing prices. Lastly, good news comes from East Indonesia where the price of shallot is believed to decrease in the next week. Hopefully, this decrease can stabilize shallot prices across Indonesia.

It can be seen from figure 5 that Central Kalimantan and South Kalimantan is expected to experience an increase in chicken prices from Rp 1.000/Kg up to Rp 5.000/Kg. This increase will make the chicken price on these two provinces exceed the reference prices set by the government. It would be preferable if the farmers in the related provinces can increase their production, since these two provinces do not take any stock from any other province, or decrease the number of chickens transferred to East Kalimantan, since the chicken price in East Kalimantan is Rp 3.000 less than South Kalimantan for each kg. The government can also make a new chicken distribution line to suppress the increase. Next, from
figure 5, the chicken price in Central Sulawesi is also expected to increase from Rp 1.000/Kg up to Rp 5.000/Kg. Fortunately, this increase will not make the chicken price exceed the reference price, but the government should evaluate whether this increase is natural or not. If it is judged to be unnatural, South Sulawesi can increase the stock distributed to Central Sulawesi or Central Sulawesi can limit the stock transferred to Gorontalo. Hopefully, by doing so, the increase in prices can be mitigated.

3.2. Model update
The model from the map can be updated by inputting a new weekly price for all commodities. The user interface for inputting the data is shown in figure 6.

Figure 4. Map for shallot prices

Figure 5. Map for chicken price
After the input process is conducted, all information in the map will be updated. In this analysis, the weekly prices for rice, shallot, chicken and eggs for third and fourth week of February and first week of March 2019 has been inputted. The new error is then calculated from the coefficient of the variation of the fourth week of February and First week of March resulted in 0.59 %, 7.13 %, 3.64 % and 1.79 % for rice, shallot, chicken and eggs respectively.

4. Conclusion

In this paper, 136 ARIMA models were built to predict weekly prices of rice, shallot, chicken, and eggs for each province in Indonesia with errors of 0.58 %, 4.1 %, 3.23 % and 2.76 % respectively. These models can also be updated if new data is inputted, thereby reducing the time needed to build 136 new ARIMA models. The result of these models is visualized using a map that can show recent prices, prices for the next two weeks, and weekly price changes. The map is colored to make the results easier to interpret. Furthermore, the map can show the distribution route for each commodity and can show a dynamic line chart to compare the related commodity price in two selected provinces. Lastly, this paper has conducted some analysis and show some insight that can be mined from the map. Hopefully the output of this paper can be used by the government to make a better decision related to the price of a specific food commodity. For further improvements, the implementation of LSTM or other machine learning algorithm is recommended, since the ARIMA models rely heavily on correlation, which only quantifies the linear relationship of data, and cannot utilize non-linear relationships of data.

References

[1] UN Global Pulse 2014 Mining Indonesian Tweets To Understand Food Price Crises UN Global Pulse Jakarta available at https://www.unglobalpulse.org/document/mining-indonesian-tweets-to-understand-food-price-crisies/

[2] Statistics Indonesia 2019 Perkembangan Indeks Harga Konsumen/Inflasi Statistics Indonesia Jakarta available at https://www.bps.go.id/subject/3/inflasi.html

[3] Kim J, Cha M and Lee J G 2017 Peer. J Computer Science 3 e126
[4] Janvier M J and Mung K J 2018 *Int. J. Sci. Res.*7 375
[5] Ministry of trade Republic of Indonesia 2017 *Regulation of Ministry of Trade Republic of Indonesia* 27/M-DAG/P3R/5/2017 Jakarta available at http://jdih.kemendag.go.id/backendx/image/regulasi/25010731_Permendag_No._27_Tahun_2017.pdf
[6] Food Price Information Center 2019 *Food Price Development* available at https://hargapangan.id/tabel-harga/pasar-tradisional/daerah
[7] Guha B and Bandyopadhyay G 2016 *J. Adv. Manag. Sci.* 4 117
[8] Cadenas E, Rivera W, Amezcua C R and Heard C 2016 *Energies* 2 109
[9] Nyoni T 2018 *Dimorian Review* 6 16-40
[10] Box G E P Jenkins G M and Reinsel G C 2018 *Time Series Analysis: Forecasting And Control* (San Francisco: Holden Day)
[11] Tsay R S and Tiao G C 1984 *J. Am. Stat. Assoc.* 79 84-96
[12] Dickey D A and Fuller W A 1979 *J. Am. Stat. Assoc.* 74 427-31
[13] hapiro S S and Wilk M B 1965 *Biometrika* 52 591-611
[14] Box G E and Pierce D A 1970 *J. Am. Stat. Assoc.* 65 1509-26