JISS Journal of Indonesian Social Science

Vol. 5, No. 12, December 2024
E-ISSN: 2723 - 6692
P-ISSN: 2723 - 6595
http://jiss.publikasiindonesia.id/

Comparison of Missing Data Handling Methods and Forecasting of Broiler Egg Prices Using Autoregressive Integrated Moving Average (Case Study: Bogor Regency/City)

Shelly Selgiant Dion, Embay Rohaeti, Maya Widyastiti
Universitas Pakuan, Bogor, Indonesia
Email: shellydion512@gmail.com, embay.rohaeti@unpak.ac.id, maya.widyastiti@unpak.ac.id
Correspondence: shellydion512@gmail.com*

KEYWORDS
Missing Data; Forecasting; Linear Interpolation; Simple Moving Average; ARIMA

ABSTRACT
Fluctuations in the price of broiler eggs can have an impact on decreasing people's purchasing power, so a form of price control through forecasting is needed. The existence of missing data in broiler egg price data can interfere with the accuracy of forecasting results. This research is carried out in two stages. The first stage is handling missing data. Missing data handling is done by comparing two methods, namely the linear interpolation method and the simple moving average (SMA) method. The second stage is forecasting with the autoregressive integrated moving average (ARIMA) method. The objectives of this study are to handle missing data on the data of broiler egg prices with linear interpolation and SMA methods, evaluate the results of the comparison of missing data handling methods, forecast future broiler egg prices, and evaluate the results of forecasting. The data used is daily data on the price of broiler eggs in Bogor Regency / City in the period January 1, 2019, to December 31, 2023, as much as 1,826 data. The results of the comparison of missing data handling methods showed that the linear interpolation method is declared better with an accuracy value using MAPE of 0.005%. The results of forecasting the price of broiler eggs show that the forecasting results with the ARIMA (1,1,3) model follow the actual data pattern, with a MAPE accuracy value of 0.601%; it is stated that the forecasting performance has performed well.

Introduction

Eggs are a very important source of animal protein and a staple for some people. The price of purebred chicken eggs in the Bogor Regency / City area tends to experience significant price changes. Based on information from the National Online Market Information System (2024) in 2019, the average price of purebred chicken eggs was recorded at Rp.24,091 per kilogram. In 2020, the average price of purebred chicken eggs increased to IDR 24,599 per kilogram, and this year, the price of purebred chicken eggs reached the lowest price in the last five years at IDR 14,000 per kilogram. In 2021, the average price of purebred chicken eggs fell to IDR 22,515 per kilogram. In 2022, the average price of purebred chicken eggs experienced a significant increase of IDR 26,619 per kilogram. In 2023, the average price of broiler eggs continued to increase to Rp. 28,400 per kilogram. This year, the price of broiler eggs also reached the highest price of Rp—32,000 per kilogram.

The price of chicken eggs is one commodity that has a significant impact on the community's economy. Fluctuations in the price of chicken eggs over a long period will result in a decrease in people's purchasing power. Therefore, an effort to control prices in the future is needed. One of the efforts to control prices in the future is by forecasting. In this research, the forecasting method that will be used is the Autoregressive Integrated Moving Average (ARIMA). The ARIMA method was chosen because of its reliability in analyzing time series data and its ability to provide accurate forecasting results (Aksan & Nurfadilah, 2020; Al'afi et al., 2020; Hyndman & Athanasopoulos, 2018).

The data used in this study is chicken egg price data in Bogor Regency / City for the period January 1, 2019, to December 31, 2023. However, the chicken egg price data found problems in the form of incomplete data. Data incompleteness can reduce forecasting accuracy. This is because time series analysis is very sensitive to time (lag). The problem of missing data in this study, one of which occurred from September 2019 to November 2019. This condition needs to include data in a fairly long period. Furthermore, in the following years, data still needed to be found. The problem of missing data (incomplete data) can reduce the accuracy of forecasting results, so it needs to be handled properly (Little & Rubin, 2020). Therefore, based on these problems, an appropriate method of handling missing data is needed (Rubin, 2020).

Some methods of handling missing data on univariate time series data include linear interpolation and simple moving average (SMA) methods. The linear interpolation method estimates the value of missing data based on a linear trend between two known data points (Sumertajaya et al., 2023). Meanwhile, the SMA method uses the average value of a number of previous observation data to fill in the missing data (Putri & Wardhani, 2020; Sarifah et al., 2023).

The missing data in this research data is linear missing data. Therefore, this study compares two missing data handling methods (linear interpolation and SMA). The performance of the two missing data handling methods was evaluated on various missing data conditions in the broiler egg price data for the period January 1, 2019, to December 31, 2023. The results of the comparison of the two methods obtained a good performance in the missing data handling method. Furthermore, the best method is used for handling missing data. The results of handling missing data obtained complete data. The stage after obtaining complete data is the process of forecasting the price of broiler eggs.

Some previous studies related to handling missing data with linear interpolation and forecasting with the ARIMA method include Ismail et al. (2023) calculating missing rainfall data using the linear interpolation method. Afridar & Andriani (2022) used the ARIMA method to predict the price of shallot commodities in Tegal City. Daratullaila and Sari (2024) applied the ARIMA method to predict the number of crimes in Indonesia.

Based on the description of the problem and previous research, this research handles missing data and forecasting. The difference between this research and previous research is the process of comparing two methods of handling missing data before the forecasting process is carried out. Therefore, this research takes the title "Comparison of Missing Data Handling Methods and Forecasting of Broiler Egg Prices with Autoregressive Integrated Moving Average." The objectives of this research are to handle missing data using the linear interpolation method and the simple moving average (SMA) method. Evaluate the comparison results of missing data handling methods with the Linear Interpolation method and the Simple Moving Average (SMA) method. Forecasting the price of chicken eggs in Bogor Regency / City using the Autoregressive Integrated Moving Average Method. Evaluate the results of forecasting the price of chicken eggs in Bogor Regency / City

Research Methods

The data used in this study are data on the price of broiler eggs in the Bogor Regency / City. The data used in this study is 1826 data from January 1, 2019, to December 31, 2021. This data can be accessed on the official website of the National Livestock Online Market Information System, namely https://simponiternak. peternakan.go.id/price-region.php.

Broadly speaking, this research consists of two stages: handling missing data and forecasting chicken egg prices. The first stage compares two methods of handling missing data: the linear interpolation method and the simple moving average method. The second stage forecasts the price of chicken eggs using the Autoregressive Integrated Moving Average (ARIMA) model.

Results and Discussion

Stages of Lost Data Handling Analysis

1. Identifying Missing Data in the Whole Data

The first stage before forecasting in this study is initial data exploration; this is done by visualizing the data in the form of plots. This data exploration aims to obtain a visual overview of the data. The results of the data exploration are presented in Figure 1.

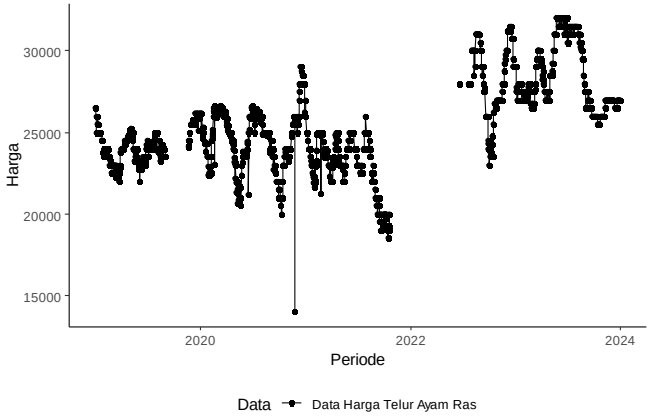

Figure 1. Broiler Egg Price Data Period January 1, 2019 to December 31, 2023

Figure 8 shows the phenomenon of missing data, which is quite varied in the data on the price of eggs in the Bogor Regency / City. The long missing data at the beginning of the first year occurred in the period from September 9, 2019, to November 18, 2019. At the end of the third year, there was sporadic missing data for a span of 9 months, namely in the period October 20, 2021, to July 31, 2019. Before going to the next stage, it is necessary to identify the percentage of missing data from the overall data. This aims to identify the magnitude of the influence of missing data on the accuracy of the forecasting results. The percentage of missing data is presented in Figure 2.

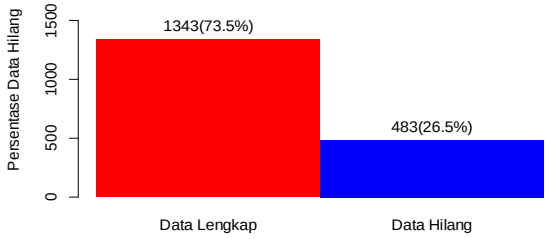

Figure 2. Percentage of Missing Data in (Overall Data)

Figure 2 shows the percentage of missing data from the overall data on the price of eggs in Bogor Regency / City. The missing data is 26.5% of the overall data. The missing data is thought to affect the accuracy of the forecasting results with the Autoregressive Integrated Moving Average model. Therefore, it is necessary to handle missing data to obtain forecasting results with good accuracy. There are several choices of missing data handling methods (missing data handling). Before handling missing data, the first step is to identify the type of missing data in the overall data.

2. Determination of Missing Data Types in the Whole Data

a. Little's MCAR Test

The missing data for five years for Sundays is 103 data; on other days, the comparison of missing data is presented in Table 1.

Table 1. Comparison of Missing Data in Overall Data

Day	Missing Data	Data Available
Sunday	103	158
Monday	56	205
Tuesday	60	201
Wednesday	64	197
Thursday	56	203
Friday	60	201
Saturday	82	179
Total	483	1343

The chi-square calculation for determining the type of missing data based on the data in Table 2 is carried out as follows:

The first step is to compare the total available data with the average data expected to be available on each day of the overall data.

$$E_{\text{tersedia}} = \frac{261 * 1343}{1826} = 187.05$$

The second step calculates the ratio of total unavailable data to the average data expected to be available on each day of the overall data.

$$E_{\text{tidak tersedia}} = \frac{261 * 483}{1826} = 73.95$$

The third step is calculating the chi square value with equation (1)

$$X^2 = \left(\frac{(158 - 187.05)^2}{187.05} + \frac{(103 - 73.95)^2}{73.95} + \cdots + \left(\frac{(178 - 187.05)^2}{187.05}\right) + \frac{(82 - 73.95)^2}{73.95}\right)$$

$X^2 = 0.99$

The next step is the calculation of degrees of freedom from the comparison of missing data with available data on the overall data.

$$df = (r - 1) \times (c - 1) = 6$$

The value of df = 6 is obtained by identifying the chi-square table in Appendix 4, resulting in a value of 12.592. The results of the above calculations are presented in Table 2.

Table 2. Little MCAR Test Results of the Whole Data

Little MCAR Test	p-value	X^2	Description
Overall data	0.319	0.99	Completely random missing data

The critical value of chi square for df = 6 at 0.05% significance level is 12.592. Since the chi square value (0.99) is less than the critical value (12.592) and the resulting p-value is more than the significant level, the decision to accept is chosen. H_0 and it can be concluded that the missing data in the overall data is a type of MCAR missing data. The MCAR missing data type confirms that the missing data loss is not influenced by the variable before or after the data loss itself, and is not influenced by the missing variable.

3. Data Sharing

The results of this comparison of missing data handling methods will then be applied to non-simulation data to estimate the missing data in non-simulation. Details of this data division can be presented in Figure 3.

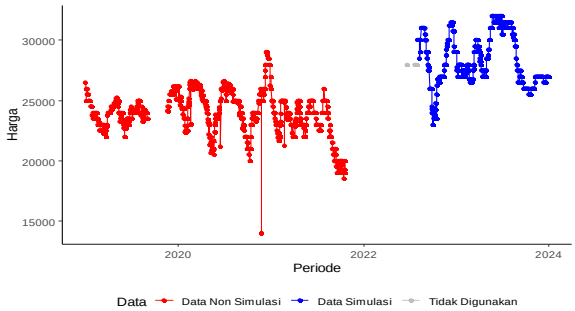

Figure 3. Distribution of Chicken Egg Price Data

Figure 3 shows the division of data in the stages of handling missing data. Based on Figure 10, the red color shows non-simulated data, namely from the period January 1, 2019, to October 20, 2021. Simulated data is marked in blue with a time period of August 1, 2022, to December 31, 2023. However, data for the time period October 21, 2021, to July 31, 2022, was not used. This is because the missing data is too long. Meanwhile, the process of handling missing data requires data before or after the missing data. This can affect the process of determining the window width for missing data, so the missing data handling method cannot perform well.

4. Identification of Missing Data Types in Non-Simulated Data

a. Type identification on non-simulated missing data

67 data points are missing during this time period for Sundays; the comparison of missing data on other days is presented in Table 3.

Table 3. Comparison of Missing Data in Overall Data

Day	Missing Data	Data Available
Sunday	67	80
Monday	20	126
Tuesday	24	122
Wednesday	26	120
Thursday	22	124
Friday	24	122
Saturday	46	101
Total	229	795

The chi-square calculation for determining the type of missing data is done as follows:

The first step is to compare the total available data with the average data expected to be available on each day of the non-simulated data.

$$E_{\text{tersedia}} = \frac{146 * 795}{1024} = 113.34$$

The second step compares the total unavailable data with the average data expected to be available on each day of the non-simulation.

$$E_{\text{tidak tersedia}} = \frac{146 * 229}{1024} = 32.65$$

The third step is calculating the chi square value with equation (1).

$$X^2 = \left(\frac{(67 - 113.34)}{113.34} + \frac{(103 - 32.65)^2}{32.65}\right) + \cdots + \left(\frac{(67 - 113.34)^2}{113.34} + \frac{(46 - 32.65)^2}{32.65}\right)$$

$X^2 = 2.33$

The next step is to calculate the degrees of freedom by comparing missing data with available data in the overall data set.

$$df = (r - 1) \times (c - 1) = 6$$

The value of df = 6 is obtained by identifying the chi-square table in Appendix 4, resulting in a value of 12.592. The results of the above calculations are presented in Table 4.

Table 4. Results of Little MCAR Test for Non-Simulation Data

Little MCAR Test	p-value	X^2	Description
Non-simulated data	0.127	2.33	Completely random missing data

Based on the test results that have been carried out, the chi-square value is smaller than the crisis value. The chi-square value obtained is 2.33, while the critical value is 12.596. Therefore, it is concluded that accept H_0. The p-value, which is greater than the significant level, supports this. This indicates that the missing data in the non-simulated data is included in the MCAR missing data type.

b. Identify the percentage of missing data in non-simulated data.

The stage after determining the type of missing data in non-simulation data is determining the percentage of missing data in non-simulation data. The percentage of missing data on non-simulation data is presented in Figure 4.

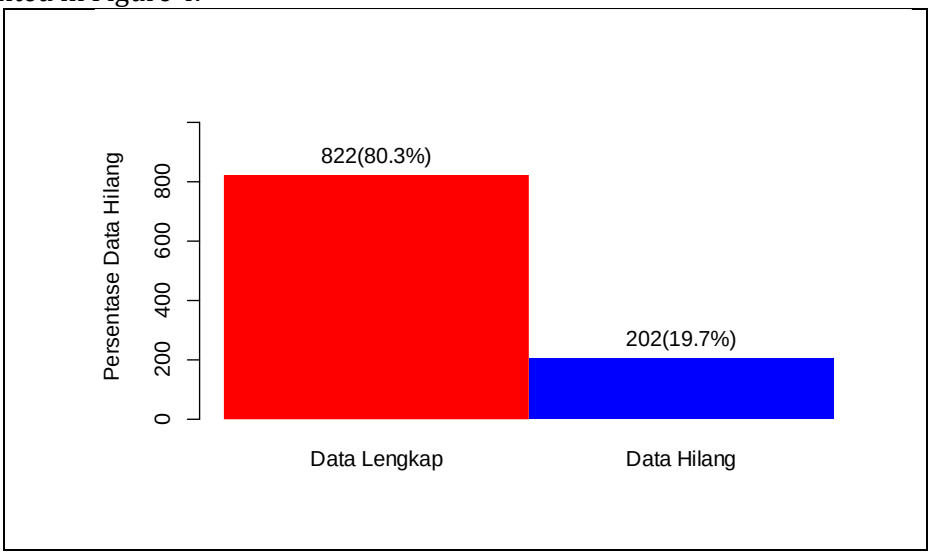

Figure 4. Percentage of Missing Data on Non-Simulated Data

Figure 4 presents the percentage of missing data in non-simulated data. The missing data is 19.7% of the total non-simulation data, which is 1024. This shows that the amount of missing data in the non-simulation data is 202 data.

The stage after obtaining the type of missing data in non-simulated data is determining the method of handling missing data. However, the selection of missing data handling methods cannot be carried out due to sporadic missing data for 9 months in the period October 21, 2021 to July 31, 2022. Therefore, the determination of the missing data handling method is carried out on the simulated data. The stages of handling missing data in simulated data begin with the data deletion stage.

5. Handling Missing Data in Simulated Data

The stage before handling missing data in simulation data is the formation of simulation data. Simulation data is obtained by generating data randomly. The simulation data used is 518 data. The simulation data is presented in the plot in Figure 12.

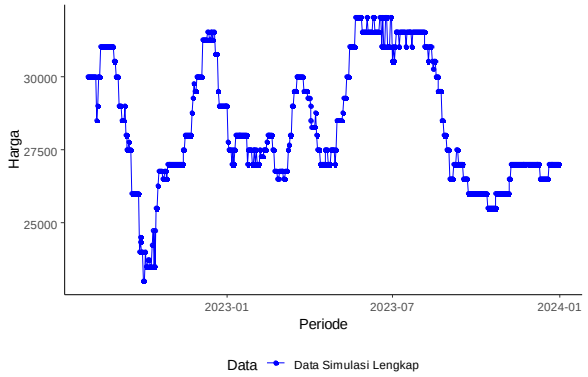

Figure 5. Complete Simulation Data

Figure 5 shows the complete simulation data for the period August 1, 2019 to December 31, 2023. The next stage of handling missing data in non-simulated data is the deletion of data in simulated data that is adjusted to the conditions of non-simulated data.

The stages of handling missing data on non-simulated data are as follows:

a. Data Deletion on Simulation Data

At this stage, data deletion is carried out on simulated data. The deletion of the type of missing data is adjusted to the type of missing data in non-simulation data and overall data. Data deletion is carried out by 20% according to the percentage of missing data in non-simulation data. The results of handling missing data on simulated data will be applied to handling missing data on non-simulated data. The results of data deletion on simulated data are shown in Figure 6.

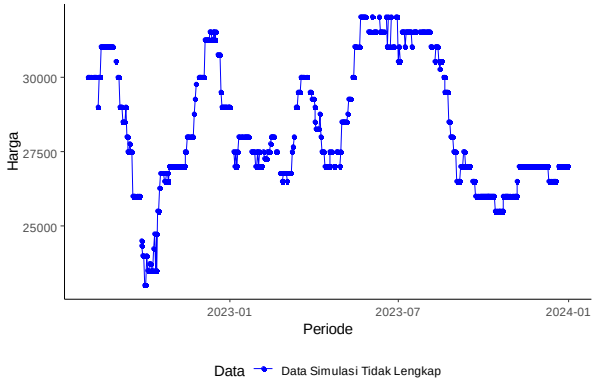

Figure 6. Incomplete Simulation Data

Figure 6 shows simulation data that has gone through the data deletion stage. The deletion of data in this simulation data aims to make the data conditions in the simulation data the same as the data conditions in the non-simulation data. The stages after the deletion of simulation data are adjusted to the conditions of non-simulation data, namely the identification of the type of missing data and the percentage of missing data in simulation data after deletion. The stages of identifying the type of missing data and the percentage of missing data in simulation data are as follows: Identification of Missing Data Types in Simulation Data: The first stage after data deletion in simulation data is the identification of missing data types. The steps for testing the type of missing data in the simulation data after deletion are the same as the steps for testing the type of missing data in the previous stages. The results of testing the type of missing data in the simulation data after deletion are presented in Table 5.

Table 5. Test Results of Little MCAR Simulation Data

Little MCAR Test	p-value	X^2	Description
Non-simulated data	0.127	2.33	Completely random missing data

Table 5 shows the results of identifying the type of missing data in the simulation data after going through the deletion stage. The p-value and chi-square value show that the simulated data after deletion has the same condition as the non-simulated data. Simulation data has contained missing data with the MCAR type. The next step is to identify the percentage of missing data in the simulation data after deletion. Identification of the Percentage of Missing Data in Simulation Data: This stage aims to determine the percentage of missing data in simulated data after deletion and whether it matches the percentage of missing data in non-simulated data or not. The results of identifying the percentage of missing data in simulated data are presented in Figure 7.

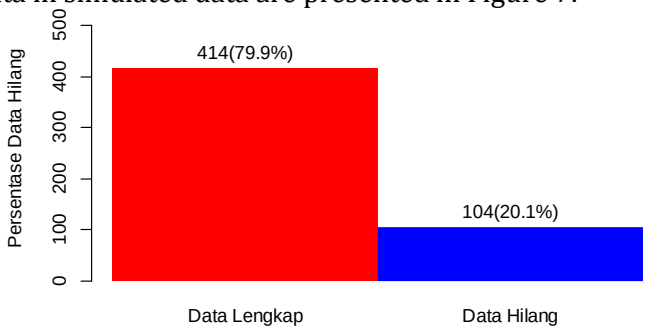

Figure 7. Percentage of Missing Data Simulation Data

Figure 7 shows that the simulation data after deletion contains 20.1% missing data. Based on the identification results for the simulation data after the deletion stage, it shows that the simulation data after deletion is in the same condition as the non-simulation data. The stage after data deletion is handling missing data. Two methods, the linear interpolation method and the SMA method, are used to handle missing data on simulation data. The first stage is handling missing data with the linear interpolation method, followed by handling missing data with the SMA method.

b. Handling Missing Data in Simulated Data with Linear Interpolation Method

At this stage, missing data handling is carried out on simulated data using the linear interpolation method. This study uses equation (4) to handle missing data using the linear interpolation method. The stages of handling missing data with the linear interpolation method are illustrated in Table 7, which presents the missing data from August 1, 2022, to August 10, 2022.

Table 6. Incomplete Chicken Egg Price Data (per Kg)

No.	Period	Initial data (Rp.)
1	01/08/2022	30.000
2	02/08/2022	30.000
3	03/08/2022	30.000
4	04/08/2022	NA
5	05/08/2022	30.000
6	06/08/2022	NA
7	07/08/2022	30.000
8	08/08/2022	30.000

9	09/08/2022	NA
10	10/08/2022	28.500

Description:
NA = missing data

Based on Table 6, the missing data is estimated using the linear interpolation method. Based on Table 4, the missing data is in the period August 3, 2022, August 4, 2022, and August 8, 2022. The steps for calculating missing data with the linear interpolation method are as follows:
The first step is to identify the missing data points. Based on Table 6, the missing data is numbers 3, 4, 8. The next step is to identify the window around the missing data point. The closest windows are numbers 2 and 5 and numbers 7 and 9. The stages of calculating missing data with the linear interpolation method:

$$y_1 = (5-3) \times \frac{(30000-30000)}{5-3} + 30000 = 30000$$

$$y_5 = (5-4) \times \frac{(30000-30000)}{5-4} + 30000 = 30000$$

Presentation of data handling results with the linear interpolation method

The results of handling missing data using the linear interpolation method are shown in Table 7.

Table 7. Results of Missing Data Handling with Linear Interpolation

No.	Period	Initial Data (Rp.)	Result of handling missing data (Rp.)
1	01/08/2022	30.000	30.000
2	02/08/2022	30.000	30.000
3	03/08/2022	30.000	30.000
4	04/08/2022	NA	30.000
5	05/08/2022	30.000	30.000
6	06/08/2022	NA	30.000
7	07/08/2022	30.000	30.000
8	08/08/2022	30.000	30.000
9	09/08/2022	NA	29.250
10	10/08/2022	28.500	28.500

Repetition of 1000 replicates

Handling missing data with the linear interpolation method is done as many as 1000 replicates. It is intended that the resulting value converges.

The results of handling missing data with the linear interpolation method on simulated data are presented in Figure 8.

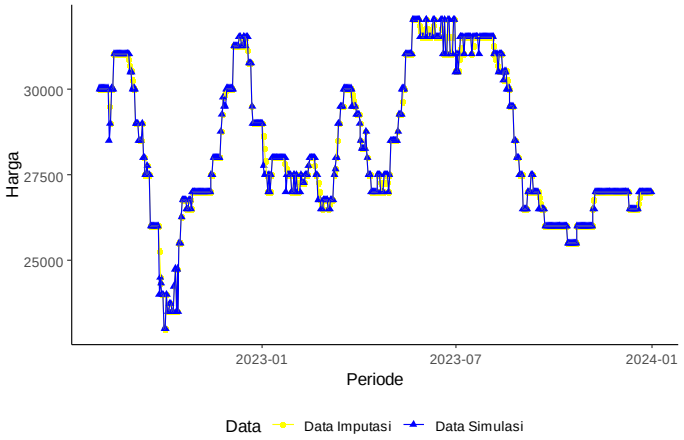

Figure 1 Handling Missing Data with Linear Interpolation

Figure 8 presents a comparison of the actual data in the simulated data with the imputed data based on the method of handling missing data with linear interpolation. Overall, the linear interpolation method has performed well. This is because visually the imputation results are close to the actual data in the simulated data. This is characterized by the side-by-side of the imputation result data with the actual data in the simulated data. Therefore, based on the plot comparison in Figure 8, an accuracy test was conducted based on MAPE.

c. Handling Missing Data in Simulated Data with Simple Moving Average (SMA) Method

At this stage, missing data handling is carried out on simulation data using the SMA method. The SMA method of handling missing data in this study uses equation (5). The stages of handling missing data with the SMA method are presented in the form of missing data illustrations for the period August 1, 2022 to August 10, 2022 as in Table 7. The steps for calculating missing data with the SMA method are as follows:

The first stage is the determination of k. In handling missing data with SMA using k = 5

Stages of handling missing data with SMA method

$$M_3 = \frac{30000 + 30000 + 30000}{3} = 30000$$

$$M_4 = \frac{30000 + 30000 + 30000 + 30000 + 30000}{5} = 29687.50$$

$$M_8 = \frac{30000 + 30000 + 30000 + 30000 + 28500}{5} = 29687.50$$

Data on the results of handling missing data with the SMA method will be presented. The results of handling missing data with SMA are presented in Table 8.

Table 8. Results of Missing Data Handling on Simulated Data with SMA Method

No.	Period	Initial data (Rp.)	Result of handling missing data (Rp.)
1	01/08/2022	30.000	30000
2	02/08/2022	30.000	30.000
3	03/08/2022	30.000	30.000
4	04/08/2022	NA	30.000
5	05/08/2022	30.000	30.000
6	06/08/2022	NA	29.688
7	07/08/2022	30.000	30.000
8	08/08/2022	30.000	30.000
9	09/08/2022	NA	29.688
10	10/08/2022	28.500	28.500

The handling of missing data with the simple moving average method was carried out for 1000 replicates.

Table 8 shows the missing data that has been handled with the SMA method. The results of handling missing data with the SMA method are presented in Figure 9.

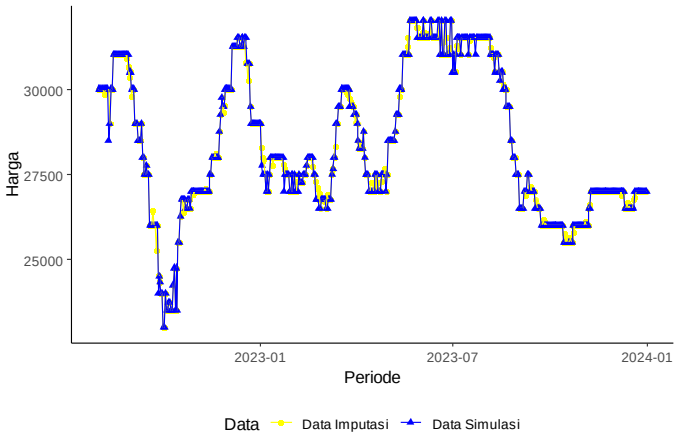

Figure 9. Handling missing data with SMA

Figure 9 compares the actual data after the simulation data with the imputed data based on the SMA missing data handling method. Overall, the SMA method has performed well. This is because visually, the imputation results are close to the actual data in the simulation data. The adjoining of the imputation characterizes this result data with the actual data in the simulation data. Therefore, based on the plot comparison in Figure 15, an accuracy test based on MAPE was conducted.

The stage after handling missing data on simulated data is to evaluate the performance of the two missing data handling methods. This stage is done by calculating the Mean Absolute Percentage Error (MAPE) value of the handling method with the linear interpolation method and the SMA method.

6. Evaluation of Lost Data Handling Methods

The results of the missing data handling method in this study are evaluated using MAPE to measure its accuracy. The first stage in evaluating the handling of missing data with the linear interpolation method and the SMA method is to calculate the percentage of error using MAPE.

The stages of this evaluation were carried out as follows:

MAPE calculation for linear interpolation method

The results of handling missing data on simulated data with the linear interpolation method are as follows:

$$\text{MAPE} = \frac{1}{518}\left|\frac{30000-30000}{30000}+\frac{30000-30000}{30000}+\cdots+\frac{27000-27000}{27000}\right|\times 100$$

$\text{MAPE} = 0.005\%$

ii. MAPE calculation for SMA method

The results of handling missing data on simulated data with the SMA method are as follows:

$$\text{MAPE} = \frac{1}{518}\left|\frac{30000-30000}{30000}+\frac{30000-30000}{30000}+\cdots+\frac{27000-27000}{27000}\right|\times 100$$

$\text{MAPE} = 0.007\%$

Comparison of accuracy results of missing data handling methods

The next step is to compare the MAPE value between the method of handling missing data with the linear interpolation method and the SMA method. The results of the calculation of the accuracy value of the comparison of the results of handling missing data using the linear interpolation method and the simple moving average method are presented in Table 9.

Table 9. Comparison of MAPE Values

No.	Lost Data Handling Methods	MAPE Value
1	Linear Interpolation	0.005%
2	Simple Moving Average	0.007%

Based on Table 9, the results of handling missing data using the linear interpolation method compared to the simple moving average method show a significant difference in the Mean Absolute

Percentage Error (MAPE) value. In handling missing data with the linear interpolation method, a MAPE value of 0.005% is obtained, while the simple moving average method produces a MAPE value of 0.007%. From these results, it can be concluded that the linear interpolation method has a better level of accuracy in handling missing data compared to the simple moving average method. The smaller MAPE value in the linear interpolation method indicates that the resulting prediction is closer to the actual value, making it reliable for further data analysis.

7. Selection of the Best Missing Data Handling Method

Based on the evaluation in the previous stage, it is known that the MAPE value for handling missing data with the linear interpolation method obtained a MAPE value of 0.005%, while for handling missing data with the simple moving average method obtained a MAPE value of 0.007%. The best missing data handling method will be selected based on the smallest MAPE value (close to zero). Therefore, from the results of the evaluation, the missing data handling method will be used: the linear interpolation method.

8. Application of Linear Interpolation Method on Non-Simulated Data

The next stage is the application of the linear interpolation method for handling missing data on non-simulated data. The stages of handling missing data on non-simulation data are the same as handling missing data on the stages of handling missing data on simulation data. The missing data is presented in Table 10.

Table 10. Incomplete Chicken Egg Price Data (per Kg)

No.	Period	Initial data (Rp.)
1	01/01/2019	NA
2	02/01/2019	26.500
3	03/01/2019	26.000
4	04/01/2019	25.000
5	05/01/2019	NA
6	06/01/2019	NA
7	07/01/2019	25.500
8	08/01/2019	25.500
9	09/01/2019	26.000
10	10/01/2019	26.500

Based on Table 10, the estimation of missing data is carried out using the linear interpolation method. Based on Table 4, the missing data is in the period January 1, 2019, January 5, 2019, and January 6, 2019. The steps for calculating missing data with the linear interpolation method are as follows:

Based on Table 10, missing data occurred in the periods January 1, 2019, January 5, 2019, and January 6, 2019. In the period January 1, 2019, because it does not have an upper window (previous data), it is concluded that the price of broiler eggs in that period is the same as the period January 2, 2019.

Stages of missing data calculation with linear interpolation method

$$y_1 = y_2 = 26500$$

$$y_5 = (5-4) \times \frac{(25500 - 25000)}{7-4} + 25500 = 25166{,}67$$

$$y_6 = (6-4) \times \frac{(25500 - 25000)}{7-4} + 25500 = 25333{,}33$$

Presentation of data handling results with the linear interpolation method

The results of handling missing data using the linear interpolation method are shown in Table 11.

Table 11. Results of Missing Data Handling on Non-Simulated Data

No.	Period	Initial data (Rp.)	Result of handling missing data (Rp.)
1	01/01/2019	NA	26.500
2	02/01/2019	26.500	26.500
3	03/01/2019	26.000	26.000
4	04/01/2019	25.000	25.000
5	05/01/2019	NA	25.166.67
6	06/01/2019	NA	25.333.33
7	07/01/2019	25.500	25.500
8	08/01/2019	25.500	25.500
9	09/01/2019	26.000	26.000
10	10/01/2019	26.500	26.500

Missing data handling with the linear interpolation method was performed for 1000 replicates.

The results of handling missing data on non-simulated data with the linear interpolation method are presented in Figure 10.

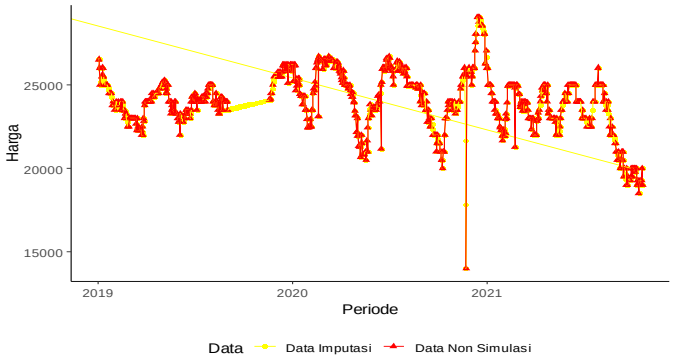

Figure 10. Plot of Missing Data Handling Results on Non-Simulated Data

Figure 10 illustrates the plot of missing data handling results on non-simulated data. Figure 17 shows the price fluctuation pattern that takes place in the period January 1, 2019 to October 20, 2021. Based on Figure 15, it is known that the missing data pattern forms a linear pattern. The stage after obtaining complete data is forecasting. In this study, forecasting the price of broiler eggs in Bogor Regency / City will be carried out using the Autoregressive Integrated Moving Average method.

Forecasting with the Autoregressive Integrated Moving Average Model

1. Full Data Exploration

Data exploration is carried out by converting data from missing data handling results with the best missing data handling method using the linear interpolation method into time series data. Data exploration aims to obtain an overview of data stationarity visually. The results of data exploration are shown in Figure 18.

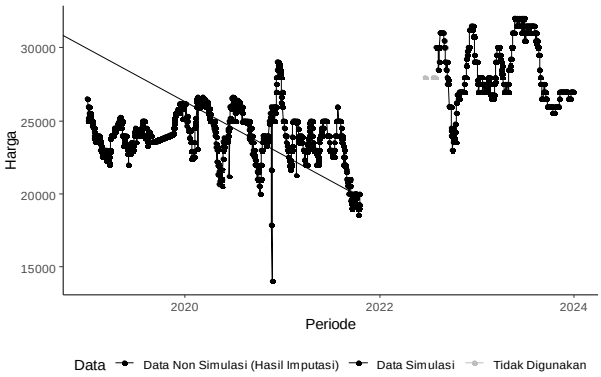

Figure 11. Time Series Data Plot

Figure 11 shows that the data on chicken egg prices fluctuated during the period January 1, 2019 to October 20, 2021. The plot in Figure 19 is more likely to have non-fixed average data, or an indication that the data is not stationary to the mean. Before forecasting, the first step is to build a model based on the train data. The model formed will then be implemented on validation data to evaluate the accuracy of the model against the forecasting results.

2. Stages of ARIMA Forecasting

a. Data Sharing

In the ARIMA forecasting stage, chicken egg price data from January 1, 2019, to December 31, 2023 will be divided into three parts: train data, validation data, and test data. The form of data division at the forecasting stage is presented in Figure 19.

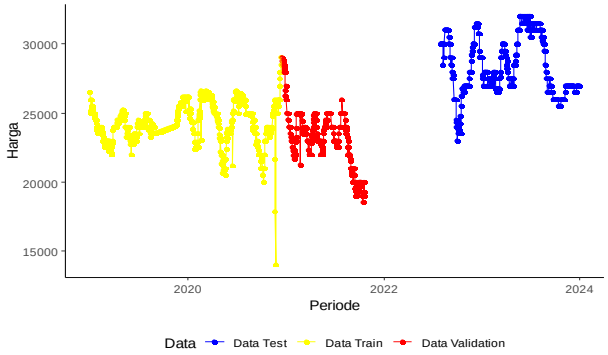

Figure 12. Data Sharing

Based on Figure 12, black color shows validation data, blue color describes train data, and purple color describes test data. The form of data division is divided for train data with a time period of January 1, 2019 to December 16, 2020 as much as 716 data. Meanwhile, the validation data starts in the period December 17, 2020 to October 20, 2021 as much as 308 data. The next stage is stationary testing of train data.

b. Stationarity Test

A data set indicated to be non-stationary in the mean requires a formal test to determine its stationarity. In this study, the test was carried out with the Augmented Dickey Fuller test (ADF Test). When the test results show that the data is not stationary, the next step is differencing. After going through the differencing process, the stationarity test is carried out again.

The results of the stationary test in this study are presented in Table 13.

Table 12. Stationarity Test

Stationarity Test	P-value	Description
At level I (0)	0.23*	Non-stationary

Difference I (1)	0.01**	Stationary

Description:

**Significant at the 5% cut-off level

Based on Table 12, the data tested at the level is not stationary because the resulting p-value is greater than the significant level limit. Data that is not stationary at the mean then needs to be differenced until the data becomes stationary to the mean. Based on Table 13, differencing on chicken egg price data is only done once. At the first differencing, the data is immediately stationary to the mean by producing a p-value smaller than the significant level.

c. Determination of ARIMA Order

The ARIMA order in this study is identified by identifying the EACF pattern. Based on the resulting EACF pattern, it can be used as a tentative model candidate. The tentative model used at this stage is through the differencing stage because previously, the data has been differenced so that the data is stationary on average data. The results of the EACF pattern are presented in Figure 20.

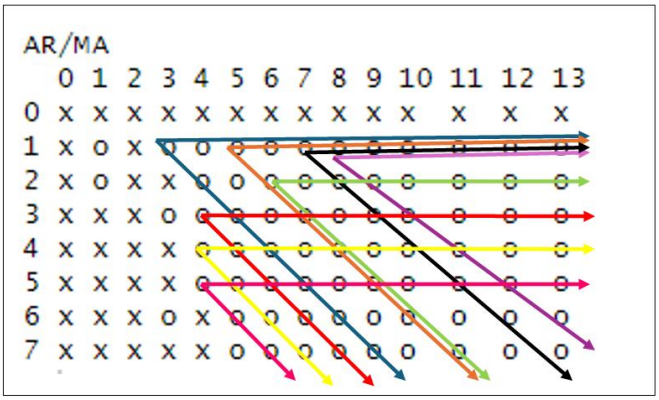

Figure 13 EACF pattern

Figure 13 shows the EACF patterns formed by the tentative models. There are 8 tentative models formed. The ARIMA models formed are as follows:

a. ARIMA (1,1,3) is shown with a blue line,
b. ARIMA (1,1,5) is shown with an orange line,
c. ARIMA (1,1,7) is shown with a black line,
d. ARIMA (1,1,8) is shown with a purple line,
e. ARIMA (2,1,6) is shown with a green line,
f. ARIMA (3,1,4) is shown with a red line,
g. ARIMA (4,1,4) is shown with a yellow line,
h. ARIMA (5,1,4) is shown with a pink line.

d. ARIMA Order Determination

After the tentative model is obtained, the next step is to estimate the model parameters. Estimation of the ARIMA model parameters is done using Maximum Likelihood Estimation (MLE). The results of the parameter estimation are presented in Table 13.

Table 13. Parameter Estimation

Model	Parameter Estimation
ARIMA (1,1,3)	$Y_t = Y_{t-1} + 0.0228(Y_{t-1} - Y_{t-2}) + \epsilon_t - 0.2809\epsilon_{t-1} - 0.1121\epsilon_{t-2} - 0.1193\epsilon_{t-3}$
ARIMA (1,1,5)	$Y_t = Y_{t-1} - 0.1198(Y_{t-1} - Y_{t-2}) + \epsilon_t + 0.1381\epsilon_{t-1} - 0.1538\epsilon_{t-2} - 0.1417\epsilon_{t-3} - 0.0246\epsilon_{t-4} + 0.0286\epsilon_{t-5}$
ARIMA (1,1,7)	$Y_t = Y_{t-1} + 0.8389(Y_{t-1} - Y_{t-2}) + \epsilon_t - 1.1005\epsilon_{t-1} + 0.0959\epsilon_{t-2} - 0.0268\epsilon_{t-3} + 0.0041\epsilon_{t-4} + 0.0352\epsilon_{t-5} + 0.0301\epsilon_{t-6} + 0.0324\epsilon_{t-7}$

ARIMA (1,1,8)	$Y_t = Y_{t-1} - 0.4537(Y_{t-1} - Y_{t-2}) + \epsilon_t + 0.1960\epsilon_{t-1} - 024038\epsilon_{t-2} - 0.1882\epsilon_{t-3} - 0.0724\epsilon_{t-4} + 00134\epsilon_{t-5} + 0.0020\epsilon_{t-6} + 0.0160 + 0.0560$
ARIMA (2,1,6)	$Y_t = Y_{t-1} - 0.0044(Y_{t-1} - Y_{t-2}) + 0.7785(Y_{t-2} - Y_{t-3}) + \epsilon_t - 0.2568\epsilon_{t-1} - 0.9058\epsilon_{t-2} + 0.0715\epsilon_{t-3} + 0.0794\epsilon_{t-4} + 0.1210\epsilon_{t-5} + 0.0329\epsilon_{t-6}$
ARIMA (3,1,4)	$Y_t = Y_{t-1} - 0.1349(Y_{t-1} - Y_{t-2}) - 0.0713(Y_{t-2} - Y_{t-3}) - 0.0609(Y_{t-3} - Y_{t-4}) + \epsilon_t - 0.1228\epsilon_{t-1} - 0.0855\epsilon_{t-2} - 0.0985\epsilon_{t-3} - 0.0519\epsilon_{t-4}$
ARIMA (4,1,4)	$Y_t = Y_{t-1} + 0.5889(Y_{t-1} - Y_{t-2}) - 0.2343(Y_{t-2} - Y_{t-3}) + 0.9169(Y_{t-3} - Y_{t-4}) - 0.4854(Y_{t-4} - Y_{t-5}) + \epsilon_t - 0.8458\epsilon_{t-1} + 0.2569\epsilon_{t-2} - 1.0211\epsilon_{t-3} + 0.7439\epsilon_{t-4}$
ARIMA (5,1,4)	$Y_t = Y_{t-1} + 0.5867(Y_{t-1} - Y_{t-2}) - 0.2338(Y_{t-2} - Y_{t-3}) + 0.8933(Y_{t-3} - Y_{t-4}) - 0.4763(Y_{t-4} - Y_{t-5}) - 0.0016(Y_{t-5} - Y_{t-6}) + \epsilon_t - 0.8423\epsilon_{t-1} + 0.2574\epsilon_{t-2} - 1.0035\epsilon_{t-3} + 0.7322\epsilon_{t-4}$

Table 14 shows the results of parameter estimation obtained by the Maximum Likelihood Estimation method. At the next stage, the best model will be selected.

e. Best Model Selection

The selection of the best model in this study is based on the smallest BIC value of the entire model formed. The comparison results for each model parameter are presented in Appendix 3. Based on Appendix 3, it shows that the best model is the ARIMA (1,1,3) model because it produces the smallest BIC value of 11287.45. The best model that will be used to produce the ARIMA (1,1,3) model can be written as follows.

$$Y_t = Y_{t-1} + 0.0228(Y_{t-1} - Y_{t-2}) + \epsilon_t - 0.2809\epsilon_{t-1} - 0.1121\epsilon_{t-2} - 0.1193\epsilon_{t-3}$$

Based on the model above, Y is the price of chicken eggs at time t.

f. Diagnostic Test

The next step is a diagnostic check of the ARIMA (1,1,3) model as the selected model. Diagnostic checks are carried out using the Ljung-Box test, which aims to check the presence of autocorrelation in the residuals. If the Ljung-Box test results show that the residuals are white noise, then the model is considered to have successfully overcome the data structure properly. The results of the diagnostic check are presented in Table 14.

Table 14 White Noise Test Results

Model	The remainder			
	Lag	p-value	Decision	Conclusion
ARIMA (1,1,3)	5	0.9274956	Accept H_0	White noise
	10	0.7754090		
	15	0.7841631		
	20	0.9352204		
	25	0.9587329		
	30	0.9914599		

Based on Table 14, the model obtained p-value data greater than 0.05. This indicates that the model is white noise. The next stage is forecasting train data using the ARIMA (1,1,3) model.

3. Forecasting with the Best ARIMA Model

Forecasting will be done with the selected ARIMA order, namely ARIMA (1,1,3). The ARIMA (1,1,3) model was selected based on previous data analysis. This forecasting process will involve applying the ARIMA (1,1,3) model to the data on the price of broiler eggs to generate future prices,

which will then be evaluated to ensure its accuracy. The forecasting results using the ARIMA (1,1,3) model are presented in Table 15.

Table 15. ARIMA (1,1,3) Forecasting Results

No.	Period	Forecasting Results (Rp.)
1	17/12/2020	28.647,93
2	18/12/2020	29.101,11
3	19/12/2020	28.831,86
4	20/12/2020	28.360,94
⋮	⋮	⋮
305	18/10/2021	19.250,29
306	19/10/2021	19.776,24
307	20/10/2021	19.296,11

Table 15 shows the forecasting results using the ARIMA (1,1,3) model, which has a price change pattern that is not too significant. This can be seen from the fluctuation pattern, which is not much different every time period. The pattern of forecasting results using the ARIMA (1,1,3) model can be seen in Figure 14.

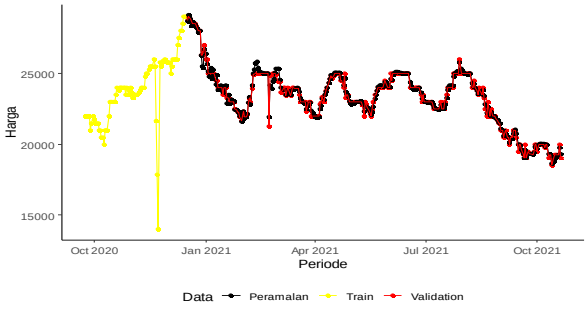

Figure 14. ARIMA (1,1,3) Forecasting Results

Figure 14 shows the comparison results of forecasting with validation data. The comparison results show that the ARIMA (1,1,3) model forecasting results closely follow the validation data pattern. This means that the forecasting accuracy is very good or that the ARIMA (1,1,3) model performs very well. After obtaining evaluation results based on comparison plots, the next stage is evaluating the performance of the ARIMA (1,1,3) model with Mean Absolute Percentage Error (MAPE).

4. Evaluation of Forecasting Results

At this stage, the performance of the ARIMA (1,1,3) model is evaluated. The accuracy measure used is MAPE. The evaluation results are presented in the form of a comparison plot of actual data with forecasting results. The evaluation results are presented in Figure 18.

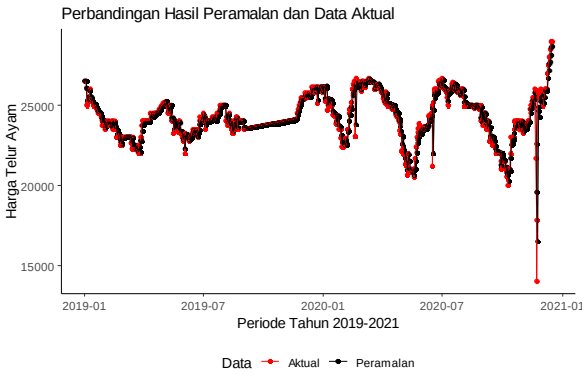

Figure 18. Comparison of Forecasting Results with Actual Data

Figure 18 shows that the pattern of the ARIMA (1,1,3) model forecasting results has a pattern that is not much different from the actual data pattern. Therefore, it can be concluded that the ARIMA (1,1,3) model is good enough to forecast the data on the price of broiler eggs. In addition, to measure the accuracy of the ARIMA (1,1,3) model in this study, MAPE is used to measure forecasting accuracy. The results of forecasting the price of broiler eggs can be presented in Table 16.

Table 16. Comparison of Actual Data and Forecasting Results

No.	Period	Price (Rp.)	Forecasting Results (Rp.)
1	17/12/2020	28.750	28.647,93
2	18/12/2020	29.000	29.101,11
3	19/12/2020	28.833,33	28.831,86
4	20/12/2020	28.666,67	28.360,94
5	21/12/2020	28.500	28.477,9
6	22/12/2020	28.500	28.620,85
7	23/12/2020	28.500	28.433,01
⋮	⋮	⋮	⋮
302	15/10/2021	19.250	18.989,30
303	16/10/2021	19.250	19.105,78
304	17/10/2021	19.250	19.195,81
305	18/10/2021	19.250	19.250,29
306	19/10/2021	20.000	19.776,24
307	20/10/2021	19.000	19.296,11

The next stage is the calculation of the MAPE value based on Table 17. The calculation of MAPE is as follows:

$$\text{MAPE} = \frac{1}{307}\left|\frac{28750 - 28647.93}{28750} + \frac{29000 - 29101.11}{29000} + \cdots + \frac{19000 - 19296.11}{19000}\right| \times 100$$

$$\text{MAPE} = 0.601\%$$

Based on the MAPE value that has been obtained, it is 0.601%. This shows that the accuracy of the ARIMA (1,1,3) forecasting model is very good. The next stage of the ARIMA (1,1,3) model is used in forecasting the final data.

5. Final Forecasting

In the final forecasting stage, forecasting is carried out on test data. The forecasting model used is the same as the forecasting model in the previous stages (forecasting on train data). Final forecasting is done to forecast the price of broiler eggs for the next 30 days.

Table 16. ARIMA (1,1,3) Forecasting Results on Test data

Period	Price (Rp)
January 1, 2024	27.005,61
January 2, 2024	27.008,37
January 3, 2024	27.010,14
January 4, 2024	27.011,72
January 5, 2024	27.013,11
⋮	⋮
January 28, 2024	27.023,26
January 29, 2024	27.023,34
January 30, 2024	27.023,41

Table 16 shows the results of forecasting the price of broiler eggs for the next 30 days. The forecasting results show an increase in the price of broiler eggs, but the increase is not too significant. The average price increase is around 0.0023%. The percentage increase is included in the category of not too large, but if this increase continues, it can have an impact on decreasing people's purchasing power. Therefore, it is necessary to control the price of broiler eggs in Bogor District/City.

Conclusion

Based on the results of the comparison of missing data handling methods and forecasting the price of broiler eggs in Bogor Regency / City, it can be concluded that missing data handling with the linear interpolation method is better than using the single moving average method. Evaluate the accuracy of missing data handling methods based on Mean Absolute Percentage Error. The Mean Absolute Percentage Error value in the linear interpolation method is 0.005%. This means that the linear interpolation method performs very well in handling missing data. The best forecasting model is the ARIMA (1,1,3) model. Forecasting the price of broiler eggs for a 30-day period shows a positive trend pattern, or the price of broiler eggs tends to increase. Evaluation of the forecasting results obtained a Mean Absolute Percentage Error value of 0.601%. This means that forecasting with the ARIMA (1,1,3) model is very good.

References

Afridar, H., & Andriani, W. (2022). Penerapan Metode Arima untuk Prediksi Harga Komoditi Bawang Merah di Kota Tegal. In *Halim Afridar IJIR* (Vol. 3, Issue 2). https://hargapangan.id/tabel-harga/pedagang-besar/daerahdengan

Aksan, I., & Nurfadilah, K. (2020). Aplikasi Metode Arima Box-Jenkins Untuk Meramalkan Penggunaan Harian Data Seluler. *Journal of Mathematics: Theory and Applications*, *2*(1), 5–10.

Al'afi, A. M., Widiart, W., Kurniasari, D., & Usman, M. (2020). Peramalan Data Time Series Seasonal Menggunakan Metode Analisis Spektral. *Jurnal Siger Matematika*, *1*(1). https://doi.org/10.23960/jsm.v1i1.2484

Asrirawan, A., Permata, S. U., & Fauzan, M. I. (2022). Pendekatan Univariate Time Series Modelling untuk Prediksi Kuartalan Pertumbuhan Ekonomi Indonesia Pasca Vaksinasi COVID-19. *Jambura Journal of Mathematics*, *4*(1), 86–103. https://doi.org/10.34312/jjom.v4i1.11717

Daratullaila, D., & Sari, R. P. (2024). Prediksi Jumlah Kejahatan di Indonesia Dengan Metode Autoregressive Integrate Moving Average (ARIMA). *Jurnal Gamma-PI*, *5*(2), 60–67. https://doi.org/10.33059/gamma-pi.v5i2.9523

Hyndman, R. J., & Athanasopoulos, G. (2018). *Peramalan: Prinsip dan Praktik* (edisi ke-2). OTexts. https://otexts.com/fpp2/

Ismail, M. R., Alfath Zain, Jamaludin, J., Dewantoro, F., & Pratiwi, D. (2023). Perhitungan Data Curah Hujan yang Hilang dengan Menggunakan Metode Interpolasi Linier. *Jurnal Teknik Sipil*, *4*(2).

Little, R. J. A., & Rubin, D. B. (2020). *Statistical Analysis with Missing Data* (3 rd Edition). Wiley.

Nailufar, E. Z., Sugianingsih, N. M. W., & Sinaga, M. O. (2023). Penerapan Metode Peramalan Arima Box-Jenkins Pada Harga Penutupan Harian Saham Alphabet Inc. *Seminar Nasional Inovasi Vokasi, 2*, 394–405. https://prosiding.pnj.ac.id/sniv/article/view/439

Nugraha, J. (2017). *Metode maximum likelihood dalam model pemilihan diskrit*. Yogyakarta: Universitas Islam Indonesia

Putri, A. N., & Wardhani, A. K. (2020). Penerapan Metode Single Moving Average Untuk Peramalan Harga Cabai Rawit Hijau. *Indonesian Journal of Technology, Informatics and Science (IJTIS)*, *2*(1), 37–40. https://doi.org/10.24176/ijtis.v2i1.5653

Ramadhan, R. H., Yusman, R., & Pranoto, G. T. (2022). Comparison of simple moving average models to identify the best model for predicting flood potential based on the normalized difference water index. *Jurnal Informatika dan Sains, 5*(2), 99–105.

Rubin, D. B. (2020). *Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. Chapman and Hall/CRC*.

Sarifah, L., Kamilah, S., & Khotijah, S. (2023). Penerapan Metode Single Moving Average Dalam Memprediksi Jumlah Penduduk Miskin Pada Perencanaan Pembangunan Daerah Kabupaten Pamekasan. *Zeta - Math Journal*, *8*(2), 47–54. https://doi.org/10.31102/zeta.2023.8.2.47-54

Sistem Informasi Pasar Online Ternak Nasional. (2024). *Harga Telur Ayam Ras Kabupaten/Kota Bogor*. https://simponiternak.peternakan. go.id/harga-daerah.php

Sumertajaya, I. M., Rohaeti, E., Wigena, A. H., & Sadik, K. (2023). Vector Autoregressive-Moving Average Imputation Algorithm for Handling Missing Data in Multivariate Time Series. *IAENG International Journal of Computer Science*, *50*(2).