JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 8, No. 3, July 2024, pp. 994-1011

Improving the Accuracy of Discrepancies in Farmers' Purchasing and Selling Index Prediction by Incorporating Weather Factors

Silvina Rosita Yulianti[1], Adhitya Ronnie Effendie[1]*, Nanang Susyanto[1]
[1]Department of Mathematics, Universitas Gadjah Mada, Indonesia
adhityaronnie@ugm.ac.id

Article History:
Received : 20-03-2024
Revised : 31-05-2024
Accepted : 31-05-2024
Online : 19-07-2024

Keywords:
Model;
ARIMA;
ARIMAX;
Forecasting;
Regression.

ABSTRACT

One measure that can be used to see the level of farmer welfare is the farmer exchange rate (NTP), which is a comparative calculation between the price index received by farmers (IJ) and the price index paid by farmers (IB), expressed as a percentage. In reality, NTP cannot explain the actual welfare situation of farmers because the ratio value has the potential to produce biased values. Another alternative that can be used to look at farmer welfare with less potential bias is to look at the difference between the sales index and the farmer purchasing index (ID). ID data forecasting can be a reference for developing and optimizing things that need to be improved in the agricultural sector. Despite the fact that a number of external factors, such as variations in the weather throughout the year, had a significant impact on the ID value, previous research used the ARIMA model to forecast without taking exogenous factors into account. Therefore, the goal of this research is to identify the optimal ARIMAX regression model for achieving accurate forecasting results with minimal error values. This research was carried out with limitations using data from the Central Statistics Agency and the Meteorological, Climatological, and Geophysical Agency in Central Java from 2008 to 2023. The first method in this research is to prepare the data, which involved collecting secondary data such as IJ and IB along with climate data such as rainfall, duration of sunlight, air pressure, wind speed, and rice prices. Next, calculate the difference between IJ and IB to determine the ID value. Then, verify the ID data's stationarity and perform AR and MA calculations. After determining the AR and MA values, construct an ARIMAX model that incorporates external factors, search for the optimal model, and utilize the optimal model to make future predictions. The results show that the accuracy of the ARIMAX model (1,1,0) has a better value than the accuracy of the ARIMA model (1,1,0), and the results obtained in this study are better than previous studies. The authors hope that the findings of this research will serve as a benchmark for the forecasting analysis of time series data in the agricultural sector, providing the local government with a foundation for policy decisions.

Crossref
https://doi.org/10.31764/jtam.v8i3.22584

——————————— ◆ ———————————

A. INTRODUCTION

Agriculture in Indonesia is a sector that has a big influence on the country's economy. In the agricultural sector, the use of data is very necessary to measure the failure or success of activities within it. In the process of processing data for further analysis that will be used in the real world, forecasting is one of the analyses most often carried out. In this case, forecasting plays an important role in optimizing the production process and increasing productivity. Apart from seasonal factors, agricultural productivity is also influenced by several external factors, such as changes in weather throughout the year (Bustos et al., 2016). The use of time series data is very necessary in this analysis. A number of studies have been conducted to explore various forecasting approaches that have been tested and applied in daily life.

Research related to forecasting is growing rapidly in response to changing times. Some research shows that it has been applied to a fairly diverse range of forecasting problems. Some of them are forecasting research on electric power systems (Amini et al., 2016; Ananthu & Neelashetty, 2021; Chodakowska et al., 2021; Javed et al., 2021; Lopez et al., 2019), natural gas (Akpinar & Yumusak, 2016; Cardoso & Cruz, 2016; Duan et al., 2022; Manowska et al., 2021), education (Bousnguar et al., 2022; Chang & Chen, 2023; Cruz et al., 2020; Qin et al., 2019; Xia & Chang, 2021), and trading strategies (Hong et al., 2022; Rostan et al., 2020; Xu et al., 2022). Apart from that, data analysis using forecasting has also been carried out in the world of health (Benvenuto et al., 2020; Capan et al., 2016; de Araújo Morais & da Silva Gomes, 2022; Juang et al., 2017; Kufel, 2020; Luo et al., 2017; Ospina et al., 2023; Roy et al., 2021; Sahai et al., 2020; Yucesan et al., 2020). In the agricultural sector, forecasting has also been widely carried out, as has been done (Zhang et al., 2021)(Jadhav et al., 2017), (Liu et al., 2018). In simple terms, in the research that has been conducted, almost all researchers use time series data obtained by taking past data. After that, create a model and predict future data.

In the world of statistics, one method that is often used in the forecasting process is the method used by George Box and Jenkins, known as the autoregressive integrated moving average (ARIMA). The ARIMA method is a powerful approach in time series data analysis that has the ability to handle complex patterns in data, such as trends, seasonality, and random fluctuations. In the context of forecasting, ARIMA's ability to adapt to variations in data makes it a very useful tool in a variety of fields, from economics and finance to social sciences and engineering. There are three main parts to ARIMA that are used in forecasting: an autoregression (AR) component, a moving average (MA) component, and an integration (I) component. The integration (I) component is used to find patterns in time series data. The ARIMA method has been proven effective in predicting future behavior from various types of data, allowing researchers to make more informed decisions based on the available information (Rosadi, 2012).

In Indonesia, one measure that is often used to measure farmer welfare is the farmer exchange rate (NTP), which is a comparison between the farmer selling index (IJ) and the farmer buying index (IB). However, this measurement has the potential to produce biased values because a large NTP does not necessarily indicate a large farmer sales index or agricultural purchasing index, and vice versa. Therefore, using a mathematical approach, a more appropriate measure to accommodate this bias is to use the difference between IJ and IB, which is called the difference index (ID). As a more accurate alternative to measuring the welfare of farmers in Indonesia, ID can provide a more precise picture of farmers' economic conditions than just using NTP (Yulianti et al., 2023).

Not many studies have analyzed ID data, including the forecasting process. In this research, ID data forecasting will be carried out using the ARIMA model, which also takes into account exogenous factors. This step is important to take because if you only pay attention to time series data from the ID or ARIMA model, it cannot accommodate the actual situation by considering that agricultural productivity is also influenced by several weather change factors and the ID value is also greatly influenced by the price of grain. This exogenous variable was chosen as one of the factors that consider the influence of ID because agricultural productivity is strongly influenced by natural factors, so the selling and buying prices of agricultural products are also very influential. The relevant government can later use the ID forecasting results to make policy decisions, such as determining insurance to protect losses of agricultural products and policies to prevent a decline or support an increase in agricultural productivity, as shown by the ID prediction value.

B. METHODS

In this study, the ARIMAX method was used, which served to predict the ID data in Central Java. This research uses open-source data taken from the Central Statistics Agency and the Meteorological, Climatological, and Geophysical Agency in Central Java from 2008 to 2023 which includes IJ and IB data, as well as exogenous variable data including rainfall data, duration of sunlight, air pressure, wind speed, and rice prices. The data used for predictions is monthly data. This research method is as follows:

1. Prepare data by aggregating data into monthly data. After data collection, the monthly IJ and IB data will be calculated for the difference to obtain ID data. Next, daily data related to climate factors such as rainfall, duration of sunlight, air pressure, wind speed, and rice prices are collected and aggregated into monthly data.
2. Check data stationarity. The prepared data is checked for stationarity to obtain optimal model formation. If the data is not stationary, differencing will be carried out until the data becomes stationary and further processing can be carried out.
3. Calculate AR and MA from the data using ACF and PACF plots.
4. Building an ARIMAX model with exogenous factors.
5. Search for the best model and forecast future data.

As for the Flowcart of study, as shown in Figure 1.

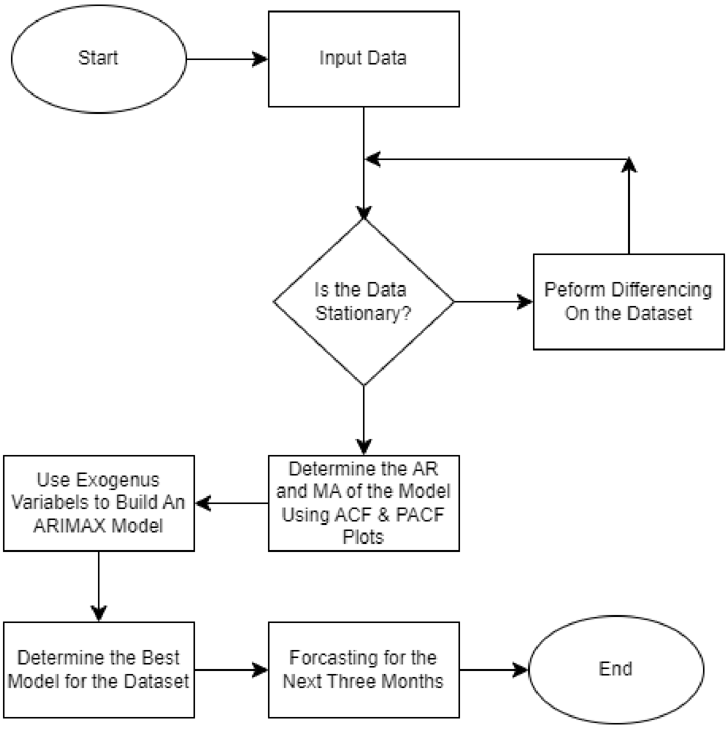

Start
Input Data
Is the Data Stationary?
Peform Differencing On the Dataset
Determine the AR and MA of the Model Using ACF & PACF Plots
Use Exogenus Variabels to Build An ARIMAX Model
Determine the Best Model for the Dataset
Forcasting for the Next Three Months
End

Figure 1. Research Methodology Using Algorithm

The ARIMA model is often used in applications involving forecasting approaches. The ARIMA model, often known as the Box-Jenkins Time Series model, is appropriate for short-term forecasting but tends to generate flat time series graphs when used for long-term forecasting. The ARIMA model consists of the following components:

1. Autoregressive Model (AR)

$$Y_t = \mathrm{c} + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots + \varphi_p Y_{t-p} + \varepsilon_t \tag{1}$$

   With $Y_t\ is$ Value of the variable at time t; c is Constant; $\varphi_1, \varphi_{2,} \dots, \varphi_p\ is$ Autoregressive coefficients; and $\varepsilon_t\ is$ Error term at time t.

2. Moving Average (MA)

$$Y_t = \mathrm{c} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon_t \tag{2}$$

   With $Y_t\ is$ Value of the variable at time t; c is Constant; $\theta_1, \theta_{2,} \dots, \theta_p$ is Moving average coefficients; and ε_t is Error term at time t

3. Autoregressive Integrated Moving Average (ARIMA)

$$Y_t = \mathrm{c} + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_p \varepsilon_{t-p} + \varepsilon_t \tag{3}$$

   With Y_t is Value of the variable at time t; c is Constant; $\varphi_1, \varphi_{2,} \dots, \varphi_p$ is Autoregressive coefficients; $\theta_1, \theta_{2,} \dots, \theta_p$ is Moving average coefficients; ε_t is Error term at time t.

The ARIMA model combined with exogenous variables (ARIMAX) is part of the dynamic regression category, including several models such as traditional multiple regression models where input factors have an immediate impact on output variables. The ARIMAX model is usually referred to as an extension or development of the ARIMA model, which includes exogenous factors in the prediction model so that the variables to be predicted do not only depend on historical data but also depend on external variables that are considered relevant. Mathematically, the ARIMAX model equation is almost the same as the ARIMA model. The difference between the two models lies in adding the $\beta_p X_{t-p}$ component to the ARIMAX model, where β_p is the coefficient of the exogenous variable X_{t-p}.

C. RESULT AND DISCUSSION

1. Data Visualization

Before starting the ARIMA and ARIMAX model building process, first focus on the dataset owned. Here is the visualization of the data.

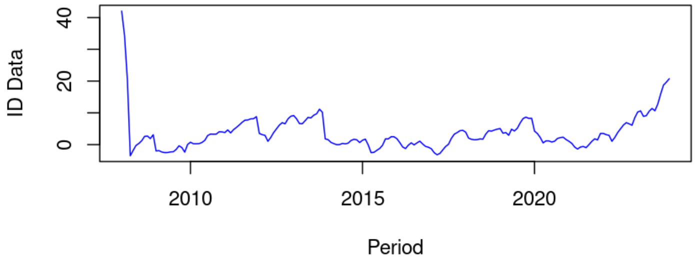


Figure 2. ID Data Visualization

Based on Figure 2, the ID data graph shows that there was a fairly large decline in ID values around 2008 but returned to a pattern that was close to stationary in the following year. This is because the IJ and IB data at the beginning of 2008 were quite high. Furthermore, based on Figure 3, Figure 4, Figure 5, Figure 6, and Figure 7, the exogenous data graphs, which include rainfall, duration of sunlight, air pressure, wind speed, and price of rice, are close to stationary from year to year.

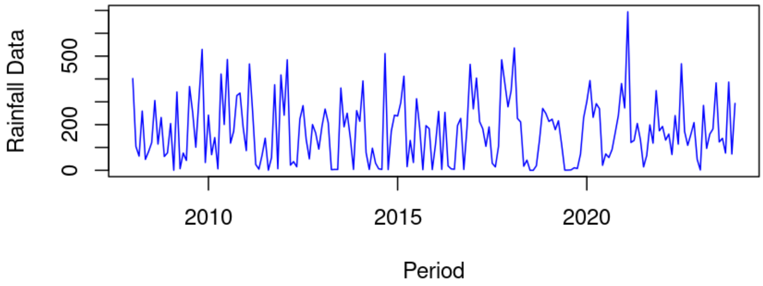


Figure 3. Rainfall Data Visualization

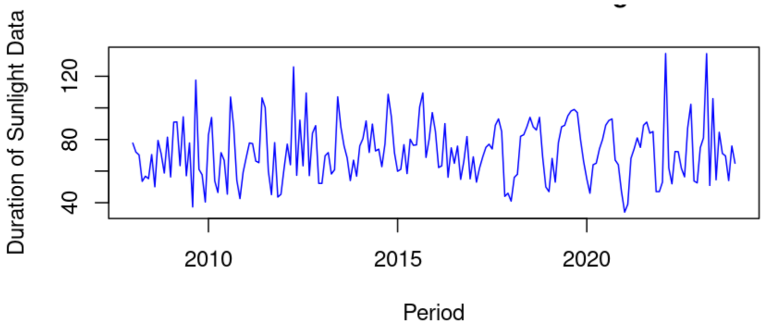


Figure 4. Duration of Sunlight Data Visualization

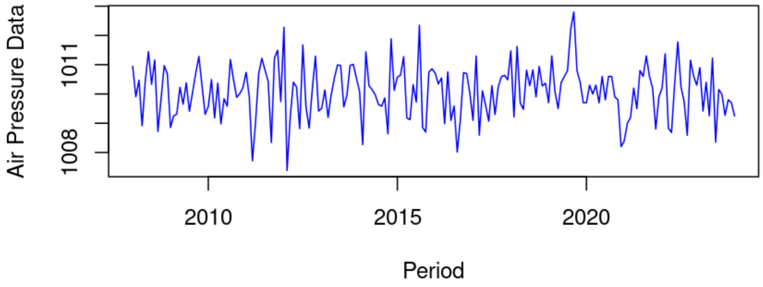


Figure 5. Air Pressure Data Visualization

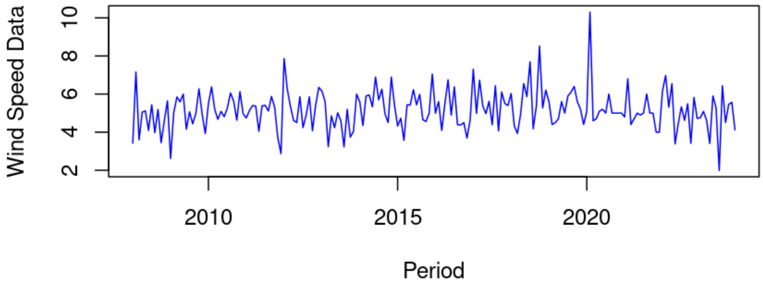


Figure 6. Wind Speed Data Visualization

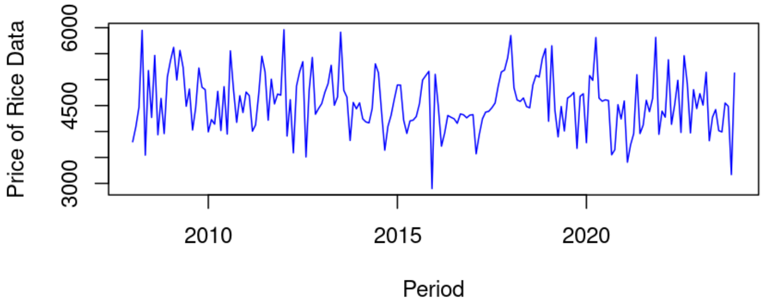


Figure 7. Price of Rice Data Visualization

The patterns formed by exogenous data based on Figure 3, Figure 4, Figure 5, Figure 6, and Figure 7 generally show high values after 2020. This can be used as a reference to show that these exogenous variables together have almost the same value character. Next, regression analysis will be carried out on the data obtained.

2. Regression Model

Since ID data is the focus of the forecasting study, it will serve as the dependent variable (Y). Moreover, potential independent variables (X) to be examined include rainfall, duration of sunlight, air pressure, wind speed, and the price of rice. Not all variables are applicable in this analysis. After conducting multiple trials, it was determined that the variables with the most significant impact on creating an ideal ARIMA model were rainfall and the price of rice. The results of the regression model conducted using R Studio program with the lm() function are as follows.

```
> reg <- lm(ID~Rainfall+PriceofRice,data=df)
> summary(reg)

Call:
lm(formula = ID ~ Rainfall + PriceofRice, data = df)

Residuals:
Min     1Q Median    3Q    Max
-7.225 -3.391 -1.444  1.944 37.755

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.9428015  3.3697676   1.764   0.0794 .
Rainfall     0.0008699  0.0029079   0.299   0.7652
PriceofRice -0.0005308  0.0007299  -0.727   0.4679
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.672 on 189 degrees of freedom
Multiple R-squared:  0.003168,             Adjusted R-squared:  -0.007381
F-statistic: 0.3003 on 2 and 189 DF,  p-value: 0.7409
```

Figure 8. Regression Code

Based on Figure 8, the following are obtained:

a. Estimation of Multiple Linear Regression Model Parameters

Table 1. Estimation

Variable	Estimation
Intercept	5,9428015
X1 (Rainfall)	0,0008699
X2 (Price of Rice)	-0,0005308

Prediction model for multiple linear regression:

$$\hat{Y} = 5{,}9428015 + 0{,}0008699\mathrm{X}_1 - 0{,}0005308\mathrm{X}_2 \tag{4}$$

Based on Table 1, it can be explained that when all explanatory variables have a value of 0, then the Y variable, or ID data, is 5.9428015. For every addition of 1 (one) to the rainfall variable, the ID data will increase by 0.0008699 on average, provided that all variables are constant. For every addition of one grain price variable, the ID data will decrease by -0.0005308 on average, provided that all variables are constant.

b. Simultaneous Significance Test

Table 2. Significance Test

Source	**Degree of Fredom (DF)**	**Sum of Square**	**Mean Square**	$\mathbf{F}_{score}$	***p-value***
Regresi	2	19.322337	9.661169	0.300301	0.7409
Residual	189	6080.435471	32.171616		
Total	191	6099.757808			

From the Table 2, it is known that the F_{score} is 0.300301 and the *p-value* is $<$ 0.7409. Results are declared significant if the $\mathrm{F}_{score} > \mathrm{F}_{table}$ or *p-value* $< \alpha$. So it can be concluded that the results of the analysis show a *p-value* of 0.7409, which is more than the significance level used, namely α = 0.05. This means that at a significance level of 0.05, all variables cannot be used to explain the significance of the variation in variable Y. The R^2 value is 0.003186, meaning that 0.31% of the diversity of variable Y can be explained by variables X1 and X2. It should be noted that this result could be caused by the variable X used being a simulation result based on some original data. The validity of this data cannot be confirmed, but it can be considered a preliminary representation. Data simulations can produce significant findings, but they require a more careful series of trials.

c. White Noise Regression Testing

```
Box-Pierce test

data:  Residue
X-squared = 112.25, df = 1, p-value < 2.2e-16
```

Figure 9. Box-Pierce Test Output

Based on Figure 9, the following are obtained: H_0 is independent residuals or residuals white noise; H_1 is the residuals are not independent of each other, or the residuals do not cause white noise. Based on the LJung-Box test, the *p-value* is 2.2 x 10-16 $< \alpha$ = 0.05, so H_0 is rejected. This means that there is sufficient evidence to state that the residuals between the lags are not independent of each other or that the residuals do not occur as white noise at the 5% level of significance.

d. Residual Stationarity of Data Regression Model.

```
Augmented Dickey-Fuller Test

data:  Residue
Dickey-Fuller = -1.2947, Lag order = 5, p-value = 0.8713
alternative hypothesis: stationary
```

Figure 10. Augmented Dickey-Fuller Test Output of Data Regression Model

Formally, the Augmented Dickey-Fuller (ADF) method can provide accurate test results to determine whether data is stationary or not. However, this ADF test only measures the level of stationarity based on the middle value. Based on Figure 10, the hypothesis is being tested as follows: H_0 is Data is not stationary; and H_1 is Data is stationary. Based on the results of the Augmented Dickey-Fuller Test (ADF Test), *p-value* = $0.8713 > \alpha = 0.05$, then H_0 is accepted. This means that there is enough evidence to say that the data is not stationary at the 0.05 significance level. To overcome this non-stationarity, differencing is necessary.

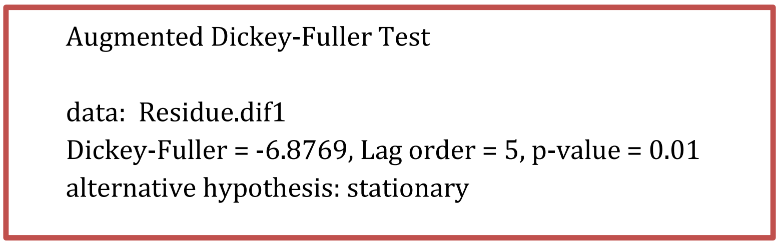
e. Differencing 1

Outputting the top 6 data points after differentiation once is presented in Table 3.

Table 3. Output The Top 6 Data Points After Differencing

2	**3**	**4**	**5**	**6**	**7**
-7.2498626	-13.9102739	-23.1164006	0.5884885	2.4051866	0.1060395

Next, data stationarity will be checked after differentiation has been carried out once, as shown in Figure 11.



```
Augmented Dickey-Fuller Test

data:  Residue.dif1
Dickey-Fuller = -6.8769, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Figure 11. Augmented Dickey-Fuller Test Output of Data After Differentiation Once

Based on Figure 11, the hypothesis is being tested as follows: H_0 is Data is not stationary; and H_1 is Data is stationary. Based on Figure 11, it was found that p-value = $0.01 < \alpha = 0.05$, so H_0 was rejected. This means that at a significance level of 0.05, the data is stationary.

f. Modeling the residuals of the regression model using ARIMA, as shown in Fiure 12 and Figure 13.

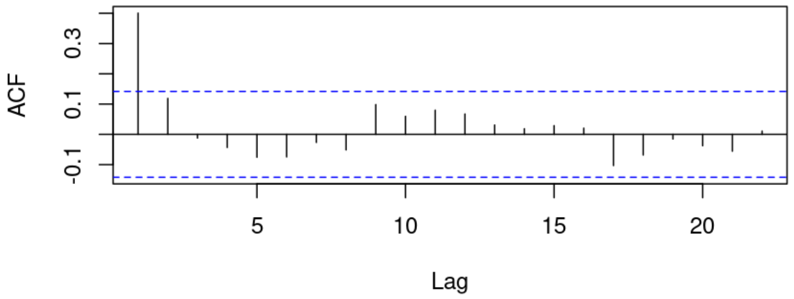


Figure 12. ACF Residual Data that has been Differentiated Once

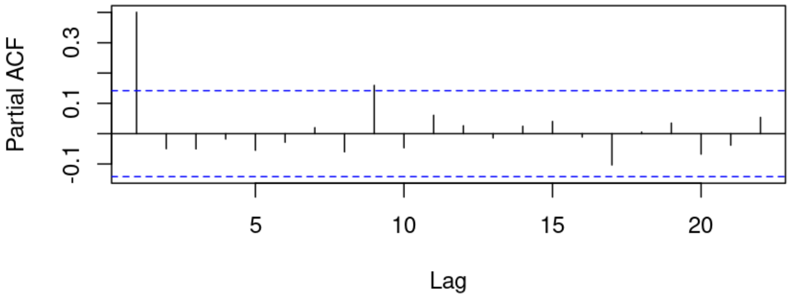


Figure 13. PACF Residual Data that has been Differentiated Once

Based on Figure 12 and Figure 13, it can be seen that the correlation value between data and lag, as in the picture above, does not decrease slowly, whereas in the ACF plot, a significant cutoff is obtained at the 1st lag, and in the PACF plot, it is found to be significant at lag 1. Based on the results of the exploration above, the models that can be formed sequentially are ARIMA (1, 1, 1), ARIMA (0, 1, 0), and ARIMA (0, 1, 1).

g. ARIMA Modeling

```
ARIMA (1,1,1)
  > modelx1 <- Arima(df$ID, xreg =
  cbind(df$Rainfall,df$PriceofRice), order = c(1,1,1), method = "ML")
  > modelx1
  Series: df$ID
  Regression with ARIMA(1,1,1) errors

  Coefficients:
           ar1     ma1   xreg1   xreg2
        0.4117  0.0486  -9e-04  -4e-04
  s.e.  0.1638  0.1753   7e-04   2e-04

  sigma^2 = 4.838:  log likelihood = -419.67
  AIC=849.35   AICc=849.67   BIC=865.61

  > lmtest::coeftest((modelx1))
  z test of coefficients:

         Estimate  Std. Error z value Pr(>|z|)
  ar1    0.41169717  0.16380491  2.5133  0.01196 *
  ma1    0.04864738  0.17527218  0.2776  0.78136
  xreg1 -0.00089227  0.00065913 -1.3537  0.17583
  xreg2 -0.00035511  0.00017465 -2.0333  0.04203 *
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 14. ARIMA (1,1,1) Code

Based on Figure 14, it is found that there are parameters in the ARIMA (1, 1, 1) model that are not significant. This can be seen from the parameter, which has a value of Pr (>|z|) > 0.05.

```
ARIMA (1,1,0)
  > modelx2 <- Arima(df$ID, xreg =cbind(df$Rainfall,df$PriceofRice), order =
  c(1,1,0), method = "ML")#terbaik
  > modelx2
  Series: df$ID
  Regression with ARIMA(1,1,0) errors

  Coefficients:
           ar1   xreg1   xreg2
        0.4525  -9e-04  -4e-04
  s.e.  0.0667   7e-04   2e-04

  sigma^2 = 4.815:  log likelihood = -419.71
  AIC=847.42   AICc=847.64   BIC=860.43
  > lmtest::coeftest((modelx2))

  z test of coefficients:

          Estimate  Std. Error z value  Pr(>|z|)
  ar1    0.45248458  0.06665691  6.7883 1.135e-11 ***
  xreg1 -0.00091186  0.00066089 -1.3797   0.16767
  xreg2 -0.00035464  0.00017589 -2.0162   0.04378 *
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 15. ARIMA(1,1,0) Code

Based on Figure 15, it is found that each parameter in the ARIMA (1, 1, 0) model is significant. This can be seen from the value of Pr (>|z|) < 0.05 for each parameter.

```
ARIMA (0,1,1)
 > modelx3 <- Arima(df$ID, xreg = cbind(df$Rainfall,df$PriceofRice), order =
 c(0,1,1), method = "ML")
 > modelx3
 Series: df$ID
 Regression with ARIMA(0,1,1) errors

 Coefficients:
          ma1   xreg1   xreg2
       0.4074  -7e-04  -4e-04
 s.e.  0.0620   7e-04   2e-04

 sigma^2 = 4.947:  log likelihood = -422.28
 AIC=852.56   AICc=852.78   BIC=865.57
 > lmtest::coeftest((modelx3))

 z test of coefficients:

          Estimate  Std. Error z value  Pr(>|z|)
 ma1    0.40739234  0.06201549  6.5692 5.059e-11 ***
 xreg1 -0.00071899  0.00065185 -1.1030   0.27003
 xreg2 -0.00037100  0.00017468 -2.1238   0.03368 *
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16. ARIMA (0,1,1) Code

Based on Figure 16, it is found that each parameter in the ARIMA (0,1,1) model is significant. This can be seen from the value of Pr(>|z|) < 0.05 for each parameter. Apart from the significance of the model parameter estimates, selecting the best model needs to be based on the smallest AIC value. There are two alternative models with all parameters that have a significant effect, namely the ARIMA (1,1,0) and ARIMA (0,1,1) models. The ARIMA (1,1,0) model has the lowest AIC value, so the ARIMA (1,1,0) model is chosen as the best model.

h. Diagnostics Test

1) Normality Test

```
Asymptotic one-sample Kolmogorov-Smirnov test

data:  Residue.arimax
D = 0.098006, p-value = 0.05002
alternative hypothesis: two-sided
```

Figure 17. Asymptotic One-sample Kolmogorov-Smirnov Test Output

Based on Figure 17, the hypothesis is being tested as follows: H_0 is Residuals follow the normal distribution; H_1 is Residuals do not follow the normal distribution. The obtained *p-value* = 0.05002 $> \alpha$ = 0.05, which means H_0 is accepted. So at a significance level of 0.05, it can be concluded that the residuals are normally distributed.

2) Mean Value Test

```
                        One Sample t-test

data:  Residue.arimax
t = -0.32112, df = 191, p-value = 0.7485
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-0.3602368  0.2593655
                         sample estimates:
                          mean of x
                         -0.05043564
```

Figure 18. One Sample T-test Output

Based on Figure 18, the hypothesis is being tested as follows: H_0 is $\mu = 0$; and H_1 is $\mu \neq 0$. The obtained *p-value* = 0.7485 $> \alpha$ = 0.05, which means H_0 is accepted. Therefore, at the significance level of 0.05, it can be concluded that the mean value of the residuals is equal to 0.

3) Autocorrelation Test

```
              Box-Ljung test

data:  Residue.arimax
X-squared = 0.0072634, df = 1, p-value = 0.9321
```

Figure 19. Box-Ljung Test Output

Based on Figure 19, the hypothesis is being tested as follows: H_0 is there is no autocorrelation. H_1 is there is autocorrelation. The obtained *p-value* = 0.9321 $> \alpha$ = 0.05, which means H_0 is accepted. H_0 is accepted, and it can be concluded that there are no symptoms of autocorrelation.

i. Comparing the ARIMA (1, 1, 0) model calculation results with original data as shown in Figure 20 and Figure 21.

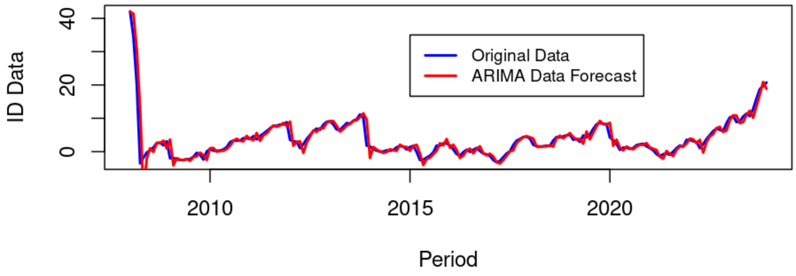


Figure 20. Original Data Vs. ARIMA (1,1,0) Data Forecast

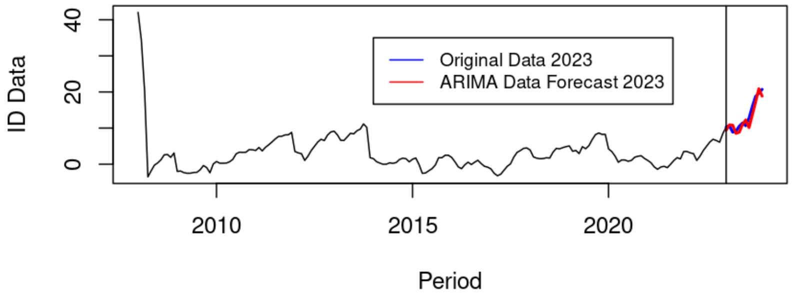


Figure 21. Original Data Vs. ARIMA (1,1,0) Data Forecast 2023

From Figure 20 and Figure 21, it can be seen that the ARIMA (1, 1, 0) model gives quite good results, as seen from the ARIMA model prediction line, which tends to run in harmony with the original data.

j. Accuracy

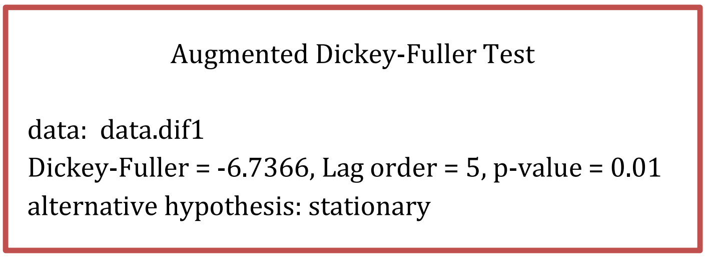
Table 4. Accuracy of ARIMAX (1,1,0)

No.	Accuracy	Measurement Forecasting
1	SSE	905.144121
2	MSE	4.714292
3	RMSE	2.171242
4	MAD	1.124557

Based on Table 4, although the ARIMAX (1, 1, 0) model shows good visual performance, evaluation based on accuracy metrics shows considerable errors. The sum of squared errors (SSE) value of 905.144121 indicates a significant level of error in the model predictions.

3. ARIMA Model without Exogenous Variables

At this stage, the data that is considered is the ID time series data. First of all, the stationarity of the ID data will be checked. Based on Augmented Dickey-Fuller Test Output of ID Data, it was found that p-value = 0.2331 $> \alpha = 0.05$. This means that at a significance level of 0.05, the data is not stationary. To overcome this non-stationarity, differencing is necessary. Next, the stationarity of the differencing ID data will be checked, as shown in Figure 22.



```
Augmented Dickey-Fuller Test

data:  data.dif1
Dickey-Fuller = -6.7366, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Figure 22. Augmented Dickey-Fuller Test Output Differencing ID Data

Based on Figure 22, the following are obtained: H_0 is Data is not stationary; and H_1 is Data is stationary. Based on Figure 22, it was found that p-value = 0.1 $< \alpha$ = 0.05, so H_0 is not accepted. This means that at a significance level of 0.05, the data is stationary. Furthermore, based on the exploration results of the ACF and PACF plots, the models that can be formed sequentially are ARIMA(1,1,1), ARIMA(1,1,0), and ARIMA(0,1,1). ARIMA (1, 1, 0) was obtained as the best model because it had the smallest AIC value. Next, the accuracy of the ARIMA model (1,1,0) is given in Table 5.

Table 5. Accuracy of ARIMA (1,1,0)

No.	Accuracy	Measurement Forecasting
1	SSE	935.742476
2	MSE	4.873659
3	RMSE	2.207636
4	MAD	1.081151

Based on Table 5, although the ARIMA (1, 1, 0) model shows good visual performance, evaluation based on accuracy metrics shows considerable errors. The sum of squared errors (SSE) value of 935.742476 indicates a significant level of error in the model predictions.

4. ARIMA VS ARIMAX

The error or accuracy value will be used to compare the two methods. **Table 6**, which contains the error values for each model, can be seen as shown in Table 6.

Table 6. ARIMA (1,1,0) VS ARIMAX (1,1,0)

Model	SSE	MSE	RMSE	MAD
ARIMA (1,1,0)	935.742476	4.873659	2.207636	1.081151
ARIMAX (1,1,0)	905.144121	4.714292	2.171242	1.124557

Next, the model is selected that has the smallest error or accuracy value. From Table 6, it can be seen that the ARIMAX (1,1,0) model as a whole has a smaller error value than the ARIMA model (1,1,0). Therefore, based on the error value, the ARIMAX (1,1,0) model has a better model than the ARIMA (1,1,0) model.

5. Forecasting

In the next stage, forecasting is carried out for the next 3 months with 10,000 simulations with ARIMAX (1,1,0) model. The step of taking the simulation more than once is carried out to obtain results with small errors. The results obtained are presented in Table 7 as follows.

Table 7. Forecasting for The Next 3 Months with 10,000 Simulations

	[,1]	**[,2]**	**[,2]**
[1,]	21.98463	22.55523	22.63905
[2,]	21.98463	22.55523	22.63905
[3,]	21.98463	22.55523	22.63905
[4,]	21.98463	22.55523	22.63905
[5,]	21.98463	22.55523	22.63905
[6,]	21.98463	22.55523	22.63905
[7,]	21.98463	22.55523	22.63905
[8,]	21.98463	22.55523	22.63905
[9,]	21.98463	22.55523	22.63905
[10,]	21.98463	22.55523	22.63905

The results based on Table 7 show that after carrying out the simulation 10,000 times, the model still produces consistent predictions. Once the ARIMA model has been configured and fitted to the data, its predictions will depend on the parameters estimated from the given time series. If, in the 10,000 simulations, there are no significant changes to the structure or basic properties of the time series, these parameters remain relatively stable.

Table 8. Original ID Data from January 2008 to December 2023 vs Forecasting ID Data from January to March 2024

2008					
Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**
42.03	34.37	20.22	-3.51	-1.83	-0.26
2023					
Jul	Aug	Sep	Oct	Nov	Dec
10.65	12.76	15.92	18.74	19.67	20.73
2024					
Jan		Feb		Mar	
21.98463		22.55523		22.63905	

Based on Table 8, this indicates that the value obtained is quite accurate based on the data analysis processes that have been carried out. The results of the prediction model analyzed have a better model compared to previous research on forecasting with the ARIMA model (Yulianti, S.R. et al., 2023). Previous research did not pay attention to exogenous factors, so external factors, in this case, weather changes, were not well accommodated, so the model built in previous research had a larger error value compared to the model built in this research.

D. CONCLUSION AND SUGGESTIONS

Data forecasting with ARIMAX (1,1,0) shows good visual performance and metric accuracy, although it shows quite large errors with SSE 905.144121, MSE 4.714292, RMSE 2.171242, and MAD 1.124557. This result is much better than the ARIMA (1,1,0) model and previous research, which was carried out without considering exogenous factors, so it had a larger error value. Even though the ARIMAX Model (1,1,0) has a much better error value than the ARIMA Model (1,1,0), the error level is still quite large, so there is still quite a lot of uncertainty in the predictions. The results of this research analysis can be a benchmark for the level of agricultural productivity in the future and can be used by the relevant government to determine agricultural insurance premiums appropriately. The author hopes that future researchers can increase accuracy by investigating other exogenous factors to further minimize the error value.

ACKNOWLEDGEMENT

I would like to thank my supervisors, Mr. Adhitya Ronnie Effendie and Mr. Nanang Susyanto, who have guided me in writing this article. I would also like to thank the Mathematics Department for giving me the opportunity to carry out research so far.

REFERENCES

Akpinar, M., & Yumusak, N. (2016). Year ahead demand forecast of city natural gas using seasonal time series methods. *Energies*, *9*(9), 1–17. https://doi.org/10.3390/en9090727

Amini, M. H., Kargarian, A., & Karabasoglu, O. (2016). ARIMA-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation. *Electric Power Systems Research*, *140*(6), 378–390. https://doi.org/10.1016/j.epsr.2016.06.003

Ananthu, D. P., & Neelashetty, K. (2021). Electrical Load Forecasting using ARIMA, Prophet and LSTM Networks. *International Journal of Electrical and Electronics Research*, *9*(4), 114–119. https://doi.org/10.37391/IJEER.090404

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in Brief*, *29*(2), 105340. https://doi.org/10.1016/j.dib.2020.105340

Bousnguar, H., Najdi, L., & Battou, A. (2022). Forecasting approaches in a higher education setting. *Education and Information Technologies*, *27*(2), 1993–2011. https://doi.org/10.1007/s10639-021-10684-z

Bustos, P., Caprettini, B., & Ponticelli, J. (2016). Agricultural productivity and structural transformation: Evidence from Brazil. *American Economic Review*, *106*(6), 1320–1365. https://doi.org/10.1257/aer.20131061

Capan, M., Hoover, S., Jackson, E. V., Paul, D., & Locke, R. (2016). Time series analysis for forecasting hospital census: Application to the neonatal intensive care unit. *Applied Clinical Informatics*, *7*(2), 275–289. https://doi.org/10.4338/ACI-2015-09-RA-0127

Cardoso, C. V., & Cruz, G. L. (2016). Forecasting Natural Gas Consumption using ARIMA Models and Artificial Neural Networks. *IEEE Latin America Transactions*, *14*(5), 2233–2238. https://doi.org/10.1109/TLA.2016.7530418

Chang, D. F., & Chen, B. Y. (2023). Exploring the Adequacy of Human Resource Investment for Quality Lower Secondary Education. *ICIC Express Letters, Part B: Applications*, *14*(2), 141–150. https://doi.org/10.24507/icicelb.14.02.141

Chodakowska, E., Nazarko, J., & Nazarko, Ł. (2021). Arima models in electrical load forecasting and their robustness to noise. *Energies*, *14*(23), 7952. https://doi.org/10.3390/en14237952

Cruz, A. P. Dela, Basallo, M. L. B., Bere, B. A., Aguilar, J. B., Kenny, C., Calvo, P., Arroyo, J. C. T., & Delima, A. J. P. (2020). Higher Education Institution (HEI) Enrollment Forecasting Using Data Mining Technique. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(2), 2060–2064. https://doi.org/10.30534/ijatcse/2020/179922020

de Araújo Morais, L. R., & da Silva Gomes, G. S. (2022). Forecasting daily Covid-19 cases in the world with a hybrid ARIMA and neural network model. *Applied Soft Computing*, *126*(6), 109315. https://doi.org/10.1016/j.asoc.2022.109315

Duan, Y., Wang, H., Wei, M., Tan, L., & Yue, T. (2022). Application of ARIMA-RTS optimal smoothing algorithm in gas well production prediction. *Petroleum*, *8*(2), 270–277. https://doi.org/10.1016/j.petlm.2021.09.001

Hong, K., Wang, X., & Xu, L. (2022). Research on price forecasting and trading strategy based on data insight. *BCP Business & Management*, *22*(1), 232–238. https://doi.org/10.54691/bcpbm.v22i.1234

Jadhav, V., Chinnappa Reddy, B. V., & Gaddi, G. M. (2017). Application of ARIMA model for forecasting agricultural prices. *Journal of Agricultural Science and Technology*, *19*(5), 981–992. https://jast.modares.ac.ir/article-23-2638-en.pdf

Javed, U., Ijaz, K., Jawad, M., Ansari, E. A., Shabbir, N., Kütt, L., & Husev, O. (2021). Exploratory data analysis based short-term electrical load forecasting: A comprehensive analysis. *Energies*, *14*(17), 1–22. https://doi.org/10.3390/en14175510

Juang, W. C., Huang, S. J., Huang, F. D., Cheng, P. W., & Wann, S. R. (2017). Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. *BMJ Open*, *7*(11), 1–7. https://doi.org/10.1136/bmjopen-2017-018628

Kufel, T. (2020). ARIMA-based forecasting of the dynamics of confirmed covid-19 cases for selected european countries. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, *15*(2), 181–204. https://doi.org/10.24136/eq.2020.009

Liu, T., Lau, A. K. H., Sandbrink, K., & Fung, J. C. H. (2018). Time Series Forecasting of Air Quality Based On Regional Numerical Modeling in Hong Kong. *Journal of Geophysical Research: Atmospheres*, *123*(8), 4175–4196. https://doi.org/10.1002/2017JD028052

Lopez, J. C., Rider, M. J., & Wu, Q. (2019). Parsimonious Short-Term Load Forecasting for Optimal Operation Planning of Electrical Distribution Systems. *IEEE Transactions on Power Systems*, *34*(2), 1427–1437. https://doi.org/10.1109/TPWRS.2018.2872388

Luo, L., Luo, L., Zhang, X., & He, X. (2017). Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models. *BMC Health Services Research*, *17*(1), 1–13. https://doi.org/10.1186/s12913-017-2407-9

Manowska, A., Rybak, A., Dylong, A., & Pielot, J. (2021). Forecasting of natural gas consumption in poland based on ARIMA-LSTM hybrid model. *Energies*, *14*(24), 1–14. https://doi.org/10.3390/en14248597

Ospina, R., Gondim, J. A. M., Leiva, V., & Castro, C. (2023). An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil. *Mathematics*, *11*(14), 1–18. https://doi.org/10.3390/math11143069

Qin, L., Shanks, K., Phillips, G. A., & Bernard, D. (2019). The Impact of Lengths of Time Series on the Accuracy of the ARIMA Forecasting. *International Research in Higher Education*, *4*(3), 58. https://doi.org/10.5430/irhe.v4n3p58

Rosadi, D. (2012). Pemanfaatan Software Open Source R dalam pemodelan ARIMA. *Seminar Nasional Dan Pendidikan Matematika 2009*, 786–795. ISBN 978-979-16353-3-2

Rostan, P., Rostan, A., & Nurunnabi, M. (2020). Options trading strategy based on ARIMA forecasting. *PSU Research Review*, *4*(2), 111–127. https://doi.org/10.1108/PRR-07-2019-0023

Roy, S., Bhunia, G. S., & Shit, P. K. (2021). Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. *Modeling Earth Systems and Environment*, *7*(2), 1385–1391. https://doi.org/10.1007/s40808-020-00890-y

Sahai, A. K., Rath, N., Sood, V., & Singh, M. P. (2020). ARIMA modelling & forecasting of COVID-19 in top five affected countries. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, *14*(5), 1419–1427. https://doi.org/10.1016/j.dsx.2020.07.042

Xia, F., & Chang, D. F. (2021). Forecasting student mobility flows in higher education: A case study in china. *ICIC Express Letters, Part B: Applications*, *12*(6), 525–532. https://doi.org/10.24507/icicelb.12.06.525

Xu, T., Wang, Z., Han, Z., Zhang, M., Liu, J., & Chang, S. (2022). A Quantitative Trading Strategy Based on A Position Management Model. *Academic Journal of Science and Technology*, *2*(1), 82–93. https://doi.org/10.54097/ajst.v2i1.901

Yucesan, M., Gul, M., & Celik, E. (2020). A multi-method patient arrival forecasting outline for hospital emergency departments. *International Journal of Healthcare Management*, *13*(S1), 283–295. https://doi.org/10.1080/20479700.2018.1531608

Yulianti, S. R., Effendie, A. R., & Susyanto, N. (2023). Revitalisasi Asuransi Pertanian: Mewujudkan Keberlanjutan dan Ketahanan Petani di Wilayah Jawa Tengah melalui Sistem Asuransi Berbasis Indeks Jual dan Beli. *Prosiding Bank Indonesia 2023*, *2*(1), 82–95. ISSN 2964-9951

Zhang, R., Guo, Z., Meng, Y., Wang, S., Li, S., Niu, R., Wang, Y., Guo, Q., & Li, Y. (2021). Comparison of arima and lstm in forecasting the incidence of hfmd combined and uncombined with exogenous meteorological variables in Ningbo, China. *International Journal of Environmental Research and Public Health*, *18*(11), 1–14. https://doi.org/10.3390/ijerph18116174