APPLICATION OF CAUSAL FORECASTING METHOD TO FORECAST SHALLOT PRODUCTION IN NORTH SUMATRA PROVINCE

Feber Wati Zebua[1], Fairus[2*], Amelia[3]

[1,2,3]Department of Mathematics, Faculty of Engineering, Universitas Samudra
Jln. Prof. Dr. Syarief Thayeb, Meurandeh, Langsa Lama, Langsa City, Aceh, 24416, Indonesia

Corresponding author's e-mail: *fairuz@unsam.ac.id

ABSTRACT

Shallot is one of the superior horticultural commodities that has a great influence on the economic value and daily needs of society. The province of North Sumatra is a strategic region for producing shallot, making it the eighth largest shallot producing province in Indonesia. The need and consumer demand for shallot along with the increasing population is a problem, as by 2021 the amount of shallot production in the household sector were only met by 11% in the North Sumatra province. The authors are therefore interested in studying the factors that influence shallot production and predicting the amount of shallot production in the future North Sumatra province. The source of data the research was carried out by the Agricultural Department of the Province of North Sumatra. The study was conducted using Causal Forecasting and ARIMA methods. The causal forecasting method used is the econometric method. The econometric method is a method for analyzing and predicting future conditions by finding and measuring several important independent variables and their influence on the variables. Dependents are observed. The ARIMA method is used to predict exogenous variables from the results of the analysis performed. Based on the analysis, it is obtained that the factors affecting the amount of shallot production are the quantity of productivity and the extent of the onion harvest. The greater the amount of productivity and the size of the harvest, the more shallot production will increase. The result of the production forecast obtained is that the lowest amount of shallot production occurs in April 2022 at 5212,763 tons and the highest amount produced on the onion occurs in March 2023 at 6502,112 tons and the average monthly amount of production is 5856,886 tons.

1. INTRODUCTION

Indonesia is one of the few developing countries with the agricultural sector being the main source of income for most of its citizens who work as farmers. Therefore, the majority of the population relies on the agricultural sector to improve their quality of life. Most of the land optimally used in Indonesia is used as agricultural land [1]. According to the Agricultural Information Center 2016, for shallot commodities, Indonesia is not among the largest producers of shallot in the world. Even to meet domestic needs, we still import large quantities of shallots [2]. Forecasting shallot is important because it can help predict the future price of the commodity. This can be useful for farmers, traders, and consumers who want to plan their production, sales, and purchases accordingly. The increased need for shallot is a potential market opportunity and can be a motivation for farmers to increase shallot production [3]. There are factors that positively influence on the production of shallots, i.e. the area of land, the amount of fertilizer usage, and the number of labor [4]. Based on the results of testing simultaneously on capital variables, the area of land, fertilizer, seedlings and labor has a significant impact on the increase in shallot production [3].

One of the regions experiencing an increase in shallot demand is the Province of North Sumatra which from 2014 to 2018 amounted to 5.197 tons or 14.60% over five years. Based on data on the amount of shallot demand, shallot production in the province of North Sumatra in 2014 - 2018, is unable to meet the quantity of consumer demand so it can cause the increase in prices [5]. At the moment, the strategy for the business of shallot is quite profitable. This is the result of consumer demand growing more intensely while also increasing the number of consumers. The average amount of onions consumed per capita per year is about 2.56 kg. The population of North Sumatra Province is approximately 14.262.147 people. Therefore, the average annual demand for shallot exceeds 36.653,7 tons. That is, the province of North Sumatra, has a deficit of about 20.550,7 tons [6].

Then the amount of production obtained in the province of North Sumatra ahead needs to be predicted. the method used to predict the level of shallot production is causal prediction. Causal prediction assumes a relationship between the independent variable and the dependent variable it affects. The Auto-Regressive Integrated Moving Average (ARIMA) method is a method that uses time series data to produce accurate short-term predictions [7].

2. RESEARCH METHODS

The type of data used is secondary data. The data in this study is based on the hypothesis of a literary study that affects the production of shallot with data sources from the Agricultural Ministry of the Province of North Sumatra. The data used is shallot production data from January 2016 to December 2021. The causal model used this time is the econometric model. The econometric model according to [8] is a quantitative analysis of actual economic phenomena based on the joint development of theory and observations, associated with the correct inference methods, as its formula is [9]:

$$Y_i = \beta_0 + \beta_1 X + u_i \tag{1}$$

Where:
Y_i = Dependent variable
β_0 = Constant
β_1 = Regression coefficient $(i = 1, 2, 3, .., k)$
X = Independent variable
u_i = Error

2.1 Correlation Analysis

Regression analysis determines the relationship between two pairing datasets, where one data group depends on the other. Correlation analysis measures how strong the relationship between variables is if the observation data consists of many variables [10]. The correlation coefficient formula is [11]:

$$r_{xy} = \frac{n(\sum XY)-(\sum X)(\sum Y),}{\sqrt{\{n \sum X2-(\sum X)2).(n \sum Y2-(\sum Y)2\}}}, \text{ where } -1 \le r_{xy} \le 1 \tag{2}$$

Where:

r = correlation coefficient,

n = number of observations,

X = first variable in the context,

Y = second variable in the context

2.2 Coefficient of Determination

A determination coefficient is used to determine how far an independent variable determines a change in the value of a bound variable [12].

$$R^2 = r^2 \ x \ 100\% \tag{3}$$

To reduce the weaknesses that occur, we use Adjusted R Square (R_{adj}^2) with the formula [13]:

$$R_{adj}^2 = R^2 - \frac{p(1-R^2)}{n-p-1} \tag{4}$$

2.3 Multicolinearity Test

The multicolinearity test is to find out if there is a lack of high correlation between independent variables in a double linear regression model. If the VIF value is not more than 10 and has a tolerance number not less than 0.10, then it pass the multicolinearity test [14]. Statistics of Testing:

$$\text{VIF} = \frac{1}{1-R_i^2} \tag{5}$$

Where:

R_i^2 = Unadjusted coefficient of determination for regressing the i-th independent variable on the remaining ones

2.4 Significance of parameters

Parameter significance testing is used to find out if there is no relationship between parameters in a regression model by calculating F (count) and t. The formula used is:

$$\text{F} = \frac{R^2/(k-1)}{1-R^2/(n-k)} \tag{6}$$

with:

F (count) value > F_table value, then H_0 is accepted.

F (count) value $\le$ F_table value, then H_1 is accepted.

Calculate the thinking value to find out whether the partial coefficient obtained is influential or not, the formula for calculating t is [15]:

$$t(\beta_i) = \frac{b_i-\beta_i}{SE(b_i)} \tag{7}$$

2.5 Assumption of Residual

a. Testing of normality

The normality test used is the Kolmogorov-Smirnov test to determine whether any particular free variable in a linear regression model is distributed normally. Normally distributed error values indicate the results of this normality test. By criteria, if the probability is greater than 0.05, the regression model distribution is normal, and if the probability is less than 0.05, the distribution of regression models is non-normal distribution [16].

b. Heteroskedasticity Test

The heteroskedasticity test uses the Glejser test to see residual variants for all homogeneous or non-homogenous observations. With a significance level of 0.05.

H_0: data is homocedastic
H_1: data is heteroscedastic

Where:
if the value $\hat{a} < a$ then accept H_0, and
if value $\hat{a} > a$ then accept H_1 [17].

c. Autocorrelation Test

The autocorrelation test aims to find out whether there is a correlation in a model between the data in the period t and the data of the period $t − 1$. [18].

$$d = \frac{\sum_{i=2}^{N}(e_i - e_{1-i})^2}{\sum_{i=1}^{N} e_i^2} \tag{8}$$

The criteria, If $d < dL$ or $4 − dL$, then there is an autocorrelation. If $dU < d < 4dU$, then there is no autocorrelation. If $dL < d < dU$ or $4 − dU < d < 4 − dL$, then no conclusion can be drawn [17].

2.6 Forecasting Error Analysis

a. Mean Absolute Percentage Error (MAPE)

MAPE is a measurement error measure that calculates the percentage of deviation between actual data (X_t) and forecast data (F_t). The value of the MAPE can be calculated with the following formula [19]:

$$\text{MAPE} = \left(\frac{1}{n}\right) \sum_{t=1}^{n} \frac{|X_t - F_t|}{X_t} x100 \tag{9}$$

Where:
X_t = actual period data t
F_t = forecasting value in period t

b. Mean Absolute Square Error (MASE)

According to Hyndman in [20], MASE is a measure of the accuracy of a prediction method that is generally applicable without having to rely on other precision measurement parameters. In research The MASE is proposed based on the MAE (Mean Absolute Error) parameter value to test the value accuracy of the naive method, where in this research the naive method produces predictions one period ahead of each data point and the error scale. The main advantage of MASE compared to MAD parameters is that MASE is more capable widely applied, because the MAD assumption is only limited to that the average is stable over time even though for fluctuating data where the data pattern goes up and down, the MAD value cannot be relied on. Otherwise MASE is suitable used for fluctuating, seasonal or trending data patterns. MASE can be used for comparing forecast methods on single time series data. The equation is:

$$\text{MASE} = \text{mean}(q_t) \tag{10}$$

$$q_t = \frac{e_t}{\frac{1}{n-t}\sum_{i=1}^{n}|y_i - y_{i-1}|} \tag{11}$$

where the numerator e_t is the forecast error for a given period, defined as the actual value (y_i) minus the forecast value (F_j) for that period: $e_t = y_i − F_i$, and the denominator is the mean absolute error of the one-step "naive forecast method" on the training set (here defined as $i = 1..t$)

2.7 Exogenous Variable Prediction

Exogenous variables act as support in performing endogenic variables and need to be predicted using the ARIMA method.

a. Autoregressive (AR)

If the initial model is:

$$(Y_t - \delta) = a_1(Y_{t-1} - \delta) + a_2(Y_{t-2} - \delta) + \cdots + a_p(Y_{t-p} - \delta) + \varepsilon_t \qquad (12)$$

b. Moving Average (MA)

The model of the Moving Average is:

$$\hat{Y}_t = \beta_0 \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \cdots + \beta_q \varepsilon_{t-q} \qquad (13)$$

c. Autoregressive Integrated Moving Average (ARIMA)

The form of the ARIMA model equation (p, d, q) is:

$$Y_t = Y_{t-1} + \delta + u_t + \varepsilon_t + \beta \varepsilon_{t-1} \qquad (14)$$

2.7.1 Differentiating Process

According to Shumway in [21], the data in the ARIMA model should be stationary, mean, and variance so that the non-stationary data is carried out an Augmented Dickey Fuller (ADF) test.

$$Y_t = \rho Y_{t-1} + \mu_t \qquad (15)$$

The variable Y_t has a unit root if ρ is 1 with the following hypothesis:

$H_0 : \rho = 1$ containing a unit of roots
$H_1 : \rho < 1$ does not contain root units

Testing is carried out by determining the statistical t-value of ADF using the formula:

$$\tau = \frac{\delta}{Se(\delta)} \qquad (16)$$

where $Se(\delta)$ is the standard error of the coefficient Y_{t-1}, if the statistical t-value ADF > the critical value ADF then H_0 is rejected which means stationary [22].

2.7.2 ARIMA Model Parameter Estimation

The measurement of the parameters can be done using the Least Square method with the equation [9]:

$$\beta = \frac{\sum_{t=2}^{n}(Z_t - \mu)(Z_{t-1} - \mu)}{\sum_{t=2}^{n}(Z_{t-1} - \mu)^2} \qquad (17)$$

To perform a significant test of the ARIMA model parameters, to test the significant parameters of the ARIMA model, the t-count test is used:

$$t_{count} = \frac{\beta}{se(\hat{\beta})}, \text{ for } se(\hat{\beta}) \neq 0 \qquad (18)$$

If t_count $|t_{count}| > t\frac{a}{2}, (n - p - 1)$ with the value $a = 0.05$, then H_0 is rejected which means the model parameter is significant [23].

2.7.3 ARIMA Model Residual Assumption Test

a. White Noise Test

Residual is white noise meaning there is no correlation between residual with mean equal to zero and constant variants performed using Ljung-Box test statistics. The hypothesis is:

$H_0: \rho_1 = \rho_2 = \cdots = \rho_k$ (residual white noise)
$H_1:$ There is a minimum of one value $\rho_k \neq 0; k = 1, 2, \dots, k$ (remaining there is no white noise)

Statistics of testing:

$$Q = (n + 2) \sum_{i=1}^{k} (n - k)^{-1} \hat{\rho}_k^2 \qquad (19)$$

b. Normality Test

The normality test is used to test whether the interfering or residual variable is distributed normally or not in the ARIMA model. Here is the normality test statistics:

$$JB = \frac{n}{6} \left(S^2 + \frac{(K-3)^2}{4} \right) \qquad (20)$$

By using a significance of 0.05, if $JB < \chi^2_{(\alpha;2)}$ or p-value $> a$ then H_0 is accepted which means the residual is distributed normally [24].

3. RESULTS AND DISCUSSION

Descriptive Statistical Analysis describes the characteristics of the data that has been obtained. What is collected is data on production quantities, planting area, number of seeds, quantity Fertilizer use, labor, price and capital of shallots. amount of shallot production in the Province North Sumatra over a period of 6 years had an average of 2042,58 ton. The highest production amount occurred in March 2021 amounting to 5.874 tons with percentage increase of 44.53% compared to the previous year. The lowest production amount occurred in April 2016 amounting to 570 tons with the percentage decrease was 15.17%. So, we get a range of amounts shallot production was 5.304 tons. The standard deviation value is 1306,250 and the variance value is 1.706.288,866. This shows data on production quantities Shallots have a large data distribution and are far from the average value.

3.1 Multicollinearity Test

Table 1 shows that the variables productivity, fertilizer, price and growing distance all have VIF values greater than 10, This means that H_1 is accepted so it can be concluded that the independent variables are not experiencing multicollinearity. There are several other variables that are removed from the analysis, according to [25], variables excluded mean there is a strong correlation so that the variables are not used.

Table 1. Multicollinearity Test Results

Model	Collinearity Statistics	
	Tolerance	VIF
(Constant)		
Productivity	0.318	3.148
Price	0.965	1.036
Labor	0.356	2.808
Harvest Area	0.325	3.076
Fertilizer use	0.000	-
Seeds	0.000	-
Capital	0.000	-
Planting area	0.000	-

3.2 Pearson Correlation Analysis

Based on the results of the normality test, the production variables (Y) and productivity variable (X_1) are not normally distributed sig (2 equals) is greater than the significant limit value of 0.05, while the price variable (X_4), labor force (X_6) and harvest area (X_8) are distributed normally. According to [26], normal undistributed data can be overcome by removing outliers and transforming data. The results of the normality test can be seen in **Table 2**.

Table 2. One-Sample Kolmogorov-Smirnov Test

		Production	Productivity	Price	Labor	Harvest Area
N		72	72	72	72	72
Normal Parameters[a,b]	Mean	2042.58	89.475	30900.56	58994.72	213.75
	Std. Deviation	1306.250	18.512	5149.692	24213.787	91.013
Most Extreme Differences	Absolute	0.210	0.185	0.147	0.144	0.129
	Positive	0.210	0.185	0.147	0.144	0.129
	Negative	-0.140	-0.090	-0.101	-0.063	-0.090
Kolmogorov-Smirnov Z		1.783	1.572	1.245	1.220	1.099
Asymp. Sig. (2-tailed)		0.003	0.014	0.090	0.102	0.179

Based on the results of the Kolmogorov Smirnov test, the Asymp. value was obtained sig (2-tailed) independent variables and dependent variables are normally distributed larger than a significant value of 0.05. The Pearson correlation test continues. The results of the normal distribution test after data transformation can be seen in **Table 3**.

Table 3. One-Sample Kolmogorov-Smirnov Test

		T_Y	T_X_1	T_X_4	T_X_6	T_X_8
N		72	72	72	72	72
Normal Parameters[a,b]	Mean	7.451	4.474	10.336	10.904	5.279
	Std. Deviation	0.570	0.198	0.164	0.406	0.414
Most Extreme Differences	Absolute	0.126	0.149	0.069	0.065	0.076
	Positive	0.126	0.149	0.069	0.065	0.076
	Negative	-0.077	-0.085	-0.061	-0.053	-0.058
Kolmogorov-Smirnov Z		1.067	1.264	0.582	0.554	0.644
Asymp. Sig. (2-tailed)		0.205	0.082	0.887	0.919	0.801

It appears that the variable correlation coefficient depends on each independent variable with a positive value which means the oriented variable relationship direction and a very strong relationship shift rate. The correlation coefficient of a dependent variable with a negative value means that the direction of the relationship between the dependent and the independent variables is non-directed and the level of relationship is very weak.

Table 4. Pearson Correlation Test Results

Correlation Probability	$Y*$	X_1*	X_4*	X_6*	X_8*
$Y*$	1	0.852461	-0.075470	0.794296	0.968396
X_1*	0.852461	1	-0.006454	-0.037219	0.695302
X_4*	-0.075470	-0.006454	1	0.006858	-0.103444
X_6*	0.725641	0.689595	0.014109	1	0.668060
X_8*	0.968396	0.695302	-0.103444	0.668060	1

3.3 Regression Analysis and Selection of the Best Causal Model

The next model will be determined based on the variable that has the largest p-value and the smallest beta coefficient. The results of the regression analysis of the selected model can be seen in **Table 5**.

Table 5. The Results of The Selected Model Regression Analysis

Variable	Coefficient	Std. Error	*t*-Statistic	*F*-statistic	Prob.
C	-2.302259	0.014844	-155.1014	390961.7	0.000
X_1*	0.999671	0.004536	220.4038		0.000
X_8*	1.000286	0.002164	462.2305		0.000

Parameter estimates are then performed simultaneously (*t* test) and partially (*F* test) to test the accuracy of the model. Based on **Table 5**, the known value of F (count) as 390961.7 and on the distribution table F obtained F table as 3.12. then it can be concluded that the value of f counts > table F meaning simultaneously independent variables have a significant influence on dependent variables. Further on the t test, the t value counted on the variable X_1* can be determined using the **Equation (5)** which is:

$$t \text{ (count)} = \frac{0.999671}{0.004536} = 220.4038$$

Then the t test for the variable X_8* is:

$$t \text{ (count)} = \frac{1.000286}{0.002164} = 462.2305$$

Based on the result of the calculated t value, it is obtained that the t count > t table and p-value > 0.05 which means variables X_1* and X_8* have a significant influence on the variable Y.

Based on the regression analysis, the following models are obtained:

$$Y* = -2.302259 + 0.999671\,X_1* + 1.000286\,X_8* + u_i$$

3.4 Residual Assumption Test

3.4.1 Normality Test

Based on the **Table 6** obtained results of the Kolmogorov Smirnov test selected models have asymp. values sig (2-tailed) by 0.174 is larger than a significant 0.05, which means the model has been distributed normally.

Table 6. Results of the One-Sample Kolmogorov-Smirnov Test

		Unstandardized Residual
N		71
Normal Parameters[a,b]	Mean	0.0000451
	Std. Deviation	0.00066239
Most Extreme Differences	Absolute	0.131
	Positive	0.129
	Negative	-0.131
Kolmogorov-Smirnov Z		1.104
Asymp. Sig. (2-tailed)		0.174

3.4.2 Heteroscedasticity Test

The probability value of the heteroscedasticity outcome is obtained on the variable X_1* of 0.3085 and the X_8* variable of 0.9217, which is greater than $\alpha = 0.05$. The results of the test can be seen in **Table 7**.

Table 7. Heteroscedasticity Test Results

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	0.016803	0.008537	1.968179	0.0531
X_1*	-0.002673	0.002605	-1.026024	0.3085
X_8*	-0.000123	0.001242	-0.098699	0.9217

3.4.3 Autocorrelation Test

Durbin Watson's value is 1.706735. If $n = 71$ and $k = 2$, then the value of dU is 1.6733 and dL is 1.5577. Using the criterion $dU < d < 4 - dU$ is $1.6733 < 2.005285 < 2.3267$ which means H_1 rebound. It can then be concluded that the selected model does not occur autocorrelation.

3.5 Exogenous Variable Prediction with ARIMA

Before using the chosen model to predict the amount of shallot production, the exogenous variable prediction is first made, namely, the prediction of the productivity variable (X_1) and harvest area variable.
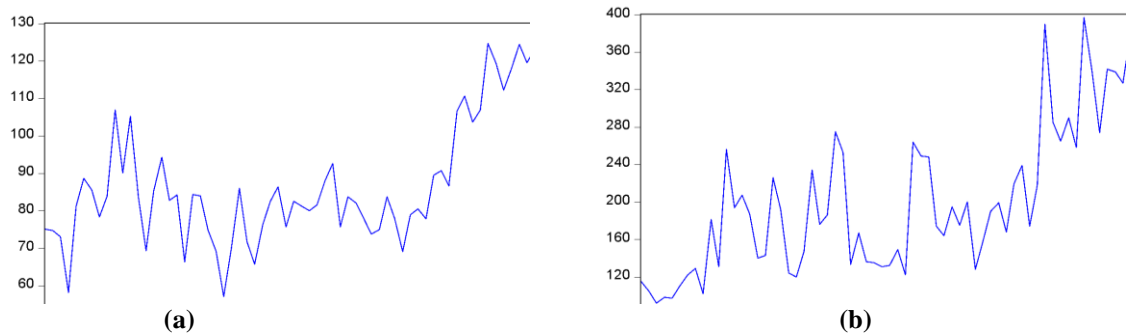
Figure 1. Exogenous Variable Data Plots
(a) Number of Productivity, (b) Harvest Area

Figure 1 (a) dan **Figure** (b) shows that the amount of productivity and size of onion harvest in North Sumatera Province increased from January 2016 to December 2021. This shows a fluctuating trend where there are indications of non-stable data.

3.5.1 Stationary Testing and Identification of ARIMA Models

In **Table 8**, we obtain data on the quantity of productivity and the extent of non-stationary harvest at the level that has a t-statistical value of ADF greater than the critical value and a p-value > 0.05, so that a root unit test is carried out on the 1st difference.

Table 8. Results Of Stationary Test Quantity

Variable	Test	t-Statistic	Probability
shallot productivity (X_1)	Augmented Dickey-Fuller test statistic	0.401573	0.9816
	Test critical values: 1% level	-3.533204	
	5% level	-2.906210	
	10% level	-2.590628	
large amount of shallot (X_8)	Augmented Dickey-Fuller test statistic	-1.487578	0.5340
	Test critical values: 1% level	3.528515	
	5% level	-2.904198	
	10% level	-2.589562	

In **Table 9**, we obtain data of the quantity of productivity and the area of non-stationary harvest at the level that has a statistical t-value of ADF smaller than the critical value and p-values < 0.05. This means that the amount of productivity and the size of the harvest has been stationary.

Table 9. Total Productivity (X_1) at 1st Difference

Variable	Test	t-Statistic	Probability
shallot productivity (X_1)	Augmented Dickey-Fuller test statistic	-8.641156	0.000
	Test critical values: 1% level	-3.530030	
	5% level	-2.904848	
	10% level	-2.589907	
large amount of shallot (X_8)	Augmented Dickey-Fuller test statistic	-12.45900	
	Test critical values: 1% level	-3.528515	0.0001
	5% level	-2.904198	
	10% level	-2.589562	

Next, determine p and q for ARIMA model assumptions based on the plot ACF and PACF on 1st Difference.

Figure 2 (a)

Lag	AC	PAC	Q-Stat	Prob
1	0.796	0.796	46.905	0.000
2	0.718	0.231	85.661	0.000
3	0.666	0.122	119.45	0.000
4	0.641	0.125	151.22	0.000
5	0.633	0.127	182.67	0.000
6	0.512	-0.247	203.56	0.000
7	0.428	-0.118	218.37	0.000
8	0.355	-0.083	228.74	0.000
9	0.340	0.079	238.42	0.000
10	0.313	0.029	246.74	0.000
11	0.237	-0.042	251.58	0.000
12	0.130	-0.184	253.06	0.000
13	0.095	0.038	253.87	0.000
14	0.075	-0.013	254.38	0.000
15	0.080	0.103	254.97	0.000
16	0.023	-0.054	255.02	0.000
17	-0.027	0.028	255.08	0.000
18	-0.024	0.045	255.14	0.000
19	-0.018	0.025	255.17	0.000
20	0.016	0.044	255.20	0.000
21	-0.032	-0.081	255.30	0.000
22	-0.058	-0.035	255.66	0.000
23	-0.061	-0.008	256.06	0.000
24	-0.047	-0.022	256.31	0.000
25	-0.047	-0.062	256.56	0.000
26	-0.084	-0.045	257.37	0.000
27	-0.105	0.004	258.67	0.000
28	-0.103	0.017	259.96	0.000
29	-0.114	-0.090	261.56	0.000
30	-0.114	0.004	263.20	0.000
31	-0.149	-0.046	266.07	0.000
32	-0.181	-0.010	270.44	0.000

Figure 2 (b)

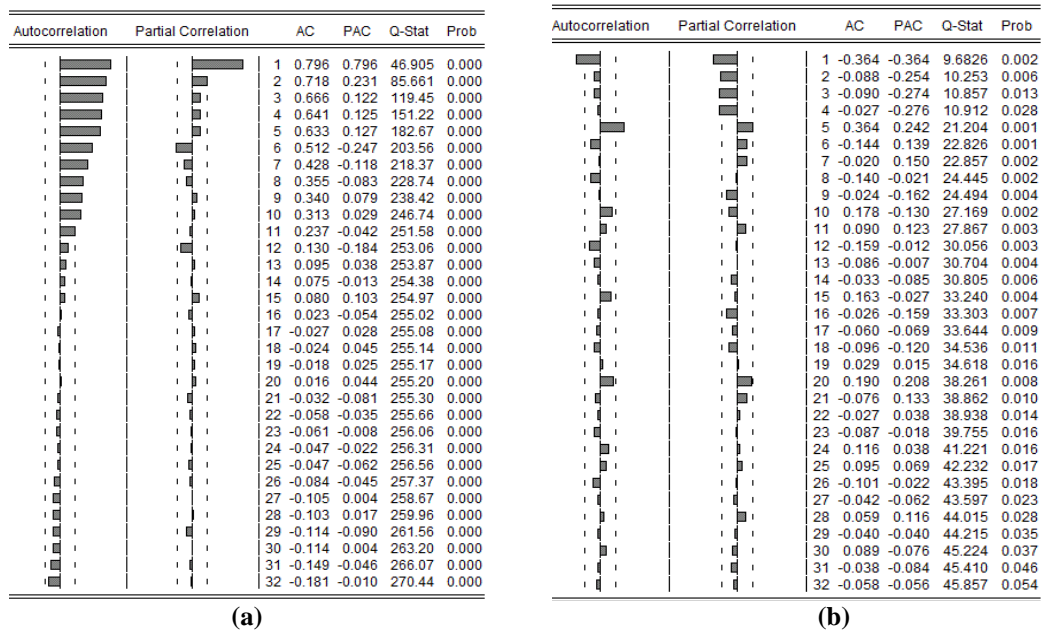| Lag	AC	PAC	Q-Stat	Prob
1	-0.364	-0.364	9.6826	0.002
2	-0.088	-0.254	10.253	0.006
3	-0.090	-0.274	10.857	0.013
4	-0.027	-0.276	10.912	0.028
5	0.364	0.242	21.204	0.001
6	-0.144	0.139	22.826	0.001
7	-0.020	0.150	22.857	0.002
8	-0.140	-0.021	24.445	0.002
9	-0.024	-0.162	24.494	0.004
10	0.178	-0.130	27.169	0.002
11	0.090	0.123	27.867	0.003
12	-0.159	-0.012	30.056	0.003
13	-0.086	-0.007	30.704	0.004
14	-0.033	-0.085	30.805	0.006
15	0.163	-0.027	33.240	0.004
16	-0.026	-0.159	33.303	0.007
17	-0.060	-0.069	33.644	0.009
18	-0.096	-0.120	34.536	0.011
19	0.029	0.015	34.618	0.016
20	0.190	0.208	38.261	0.008
21	-0.076	0.133	38.862	0.010
22	-0.027	0.038	38.938	0.014
23	-0.087	-0.018	39.755	0.016
24	0.116	0.038	41.221	0.016
25	0.095	0.069	42.232	0.017
26	-0.101	-0.022	43.395	0.018
27	-0.042	-0.062	43.597	0.023
28	0.059	0.116	44.015	0.028
29	-0.040	-0.040	44.215	0.035
30	0.089	-0.076	45.224	0.037
31	-0.038	-0.084	45.410	0.046
32	-0.058	-0.056	45.857	0.054

(a)   (b)

Figure 2. Correlogram Of Productivity Variables (X_1)
(a) Lot ACF And PACF Before the First Difference, (b) Plot ACF And PACF After the First Difference

In **Figure 2** (a) the ACF plot decreases exponentially, meaning X_1 is not stationary against the average so it is necessary to make 1st Difference. **Figure 2** (b) shows lag 1 on the ACF plot interrupted, then the temporary model estimate for the MA order is (1). On the plot PACF there is a lag 1 until lag 4 is interrupted, then the estimate of the temporary model for the AR order is (4). So obtained the initial model the amount of shallot productivity while ARIMA (4,1,1).
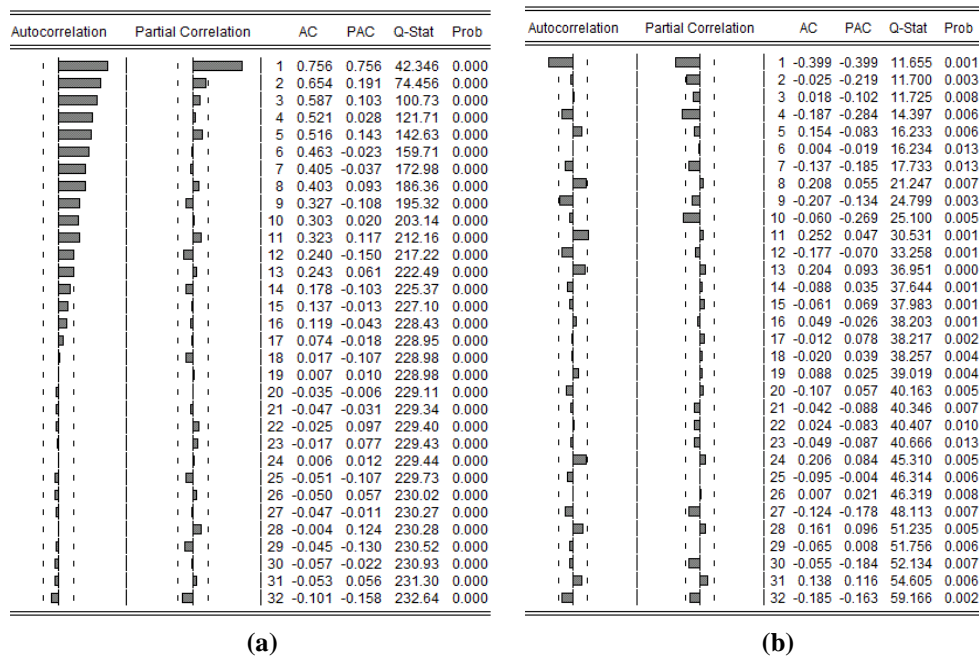
Figure 3 (a)

Lag	AC	PAC	Q-Stat	Prob
1	0.756	0.756	42.346	0.000
2	0.654	0.191	74.456	0.000
3	0.587	0.103	100.73	0.000
4	0.521	0.028	121.71	0.000
5	0.516	0.143	142.63	0.000
6	0.463	-0.023	159.71	0.000
7	0.405	-0.037	172.98	0.000
8	0.403	0.093	186.36	0.000
9	0.327	-0.108	195.32	0.000
10	0.303	0.020	203.14	0.000
11	0.323	0.117	212.16	0.000
12	0.240	-0.150	217.22	0.000
13	0.243	0.061	222.49	0.000
14	0.178	-0.103	225.37	0.000
15	0.137	-0.013	227.10	0.000
16	0.119	-0.043	228.43	0.000
17	0.074	-0.018	228.95	0.000
18	0.017	-0.107	228.98	0.000
19	0.007	0.010	228.98	0.000
20	-0.035	-0.006	229.11	0.000
21	-0.047	-0.031	229.34	0.000
22	-0.025	0.097	229.40	0.000
23	-0.017	0.077	229.43	0.000
24	0.006	0.012	229.44	0.000
25	-0.051	-0.107	229.73	0.000
26	-0.050	0.057	230.02	0.000
27	-0.047	-0.011	230.27	0.000
28	-0.004	0.124	230.28	0.000
29	-0.045	-0.130	230.52	0.000
30	-0.057	-0.022	230.93	0.000
31	-0.053	0.056	231.30	0.000
32	-0.101	-0.158	232.64	0.000

Figure 3 (b)

Lag	AC	PAC	Q-Stat	Prob
1	-0.399	-0.399	11.655	0.001
2	-0.025	-0.219	11.700	0.003
3	0.018	-0.102	11.725	0.008
4	-0.187	-0.284	14.397	0.006
5	0.154	-0.083	16.233	0.006
6	0.004	-0.019	16.234	0.013
7	-0.137	-0.185	17.733	0.013
8	0.208	0.055	21.247	0.007
9	-0.207	-0.134	24.799	0.003
10	-0.060	-0.269	25.100	0.005
11	0.252	0.047	30.531	0.001
12	-0.177	-0.070	33.258	0.001
13	0.204	0.093	36.951	0.000
14	-0.088	0.035	37.644	0.001
15	-0.061	0.069	37.983	0.001
16	0.049	-0.026	38.203	0.001
17	-0.012	0.078	38.217	0.002
18	-0.020	0.039	38.257	0.004
19	0.088	0.025	39.019	0.004
20	-0.107	0.057	40.163	0.005
21	-0.042	-0.088	40.346	0.007
22	0.024	-0.083	40.407	0.010
23	-0.049	-0.087	40.666	0.013
24	0.206	0.084	45.310	0.005
25	-0.095	-0.004	46.314	0.006
26	0.007	0.021	46.319	0.008
27	-0.124	-0.178	48.113	0.007
28	0.161	0.096	51.235	0.005
29	-0.065	0.008	51.756	0.006
30	-0.055	-0.184	52.134	0.007
31	0.138	0.116	54.605	0.006
32	-0.185	-0.163	59.166	0.002

(a)   (b)

Figure 3. Correlogram Of Harvest Area Variables (X_8)
(a) Lot ACF And PACF Before the First Difference, (b) Plot ACF And PACF After the First Difference

In **Figure 3** the ACF plot shows a lag 1 that comes out of a significant line (cuts off), then the estimated temporary model for the MA order is (1). On the PACF plot there are lag 1 and lag 2 that are out of a significant line (cuts off), then the estimated temporary model for the AR order is (2). So based on the plot ACF and PACF obtained a model of the large amount of shallot harvest while ARIMA (2,1,1).

Subsequently, estimate and test the significance of the ARIMA assumption model. **Table 13** shows the significant and non-significant model parameters on the ARIMA model. for the insignificant mode

parameters can be removed from the model to obtain a significant model, so that the model is continued on the test of residual diagnostic assumptions.

Table 10. Estimate and Test Results of Significance of Shallot Productivity Model

Model Assumptions	Parameters	Estimate	T-Value	P-Value	Test Results
ARIMA (1,1,1)	ϕ_1	0.073424	0.371205	0.7117	Not Significant
	θ_1	-0.615463	-4.052543	0.0001	Significant
ARIMA (2,1,1)	ϕ_1	-0.024141	-0.073772	0.9414	Not Significant
	ϕ_2	-0.143515	-0.756004	0.4524	Not Significant
	θ_1	-0.513234	-1.824184	0.0727	Not Significant
ARIMA (3,1,1)	ϕ_1	-0.285135	-0.840327	0.4039	Not Significant
	ϕ_2	-0.277835	-1.331423	0.1878	Not Significant
	ϕ_3	-0.240690	-1.446721	0.1529	Not Significant
	θ_1	-0.274258	-0.705814	0.4829	Not Significant
ARIMA (4,1,0)	ϕ_1	-0.601254	-4.311728	0.0001	Significant
	ϕ_2	-0.482113	-3.162075	0.0024	Significant
	ϕ_3	-0.418507	-3.097602	0.0029	Significant
	ϕ_4	-0.274041	-2.057845	0.0437	Significant
ARIMA (4,1,1)	ϕ_1	-1.020427	-3.948459	0.0002	Significant
	ϕ_2	-0.708352	-3.437513	0.0010	Significant
	ϕ_3	-0.587690	-3.382306	0.0012	Significant
	ϕ_4	-0.409083	-3.424056	0.0011	Significant
	θ_1	0.465279	1.669539	0.1000	Not Significant

Table 11 shows that there are non-significant model parameters. Non-significant model parameters can be removed from the model in order to be continued on the further analysis of the following test of residual diagnostic assumptions and model evaluation as one of the conditions to obtain the best model.

Table 11. Estimation and Test Results of Significance Model Volume of Shallot Harvest

Model Assumptions	Parameters	Estimates	T-Value	P-Value	Test Results
ARIMA (0,1,1)	θ_1	-0.688978	-7.417566	0.000	Significant
ARIMA (1,1,0)	ϕ_1	-0.406005	-3.242882	0.0014	Significant
ARIMA (1,1,1)	ϕ_1	0.217203	0.931977	0.3547	Not Significant
	θ_1	-0.809928	-7.237078	0.000	Significant
ARIMA (2,1,0)	ϕ_1	-0.488995	-2.902526	0.0050	Significant
	ϕ_1	-0.216001	-1.310113	0.1947	Not Significant
ARIMA (2,1,1)	ϕ_1	0.232647	0.947929	0.3467	Not Significant
	ϕ_1	0.062074	0.383506	0.7026	Not Significant
	θ_1	-0.835090	-5.994457	0.000	Significant

The first residual diagnostic assumption test is the white noise test. **Table 12** shows that the ARIMA models that meet the white noise assumption are ARIMA (4,1,0) and ARIMA (4,1,1). It can be seen on statistical values Ljung-Box based on p-value for lag 12 to 48, having p-value > 0.05 or $dvid$ based on $Q < X^2{}_{a;k-p}$ which means H_0 is accepted (residual white Noise).

Table 12. Residual White Noise Assumption for Productivity

Model	Lag k	Df	Q	χ^2	P-Value	Conclusion
ARIMA	12	9	23.983	16.918	0.008	No white noise
(1,1,1)	24	21	39.002	32.670	0.014	No white noise
	36	33	49.190	47.399	0.044	No white noise
	48	45	70.227	61.656	0.012	No white noise
ARIMA	12	7	11.668	14.067	0.167	*White Noise*
(4,1,0)	24	19	23.849	30.143	0.249	*White Noise*
	36	31	36.184	44.985	0.279	*White Noise*
	48	43	51.911	59.303	0.193	*White Noise*

Model	Lag k	Df	Q	χ^2	P-Value	Conclusion
ARIMA	12	6	9.280	12.591	0.233	*White Noise*
(4,1,1)	24	18	19.337	28.869	0.435	*White Noise*
	36	30	29.738	43.772	0.531	*White Noise*
	48	42	44.119	58.124	0.424	*White Noise*

Further testing of the assumption of residual white noise for the amount of crop area can be seen in **Table 13**. **Table 13** shows Ljung-Box statistics for lag 12 to 48, having a p-value > 0.05 or a view based on $< X^2{}_{a;k-p}$ where H_0 is accepted which means all ARIMA models until amount of shallot crop area meets the assumption of white noise.

Table 13. **Residual White Noise Assumption for Large Harvest Volumes**

Model	Lag k	Df	Q	χ^2	P-Value	Conclusion
ARIMA	12	10	13.055	18.307	0.290	*White Noise*
(0,1,1)	24	22	25.441	33.924	0.328	*White Noise*
	36	34	35.657	48.602	0.437	*White Noise*
	48	46	48.209	62.829	0.428	*White Noise*
ARIMA	12	10	19.315	18.307	0.056	*White Noise*
(1,1,0)	24	22	31.899	33.924	0.102	*White Noise*
	36	34	45.394	48.602	0.112	*White Noise*
	48	46	60.188	62.829	0.094	*White Noise*
ARIMA	12	9	11.644	16.918	0.310	*White Noise*
(1,1,1)	24	21	22.266	32.670	0.444	*White Noise*
	36	33	33.288	47.399	0.502	*White Noise*
	48	45	46.044	61.656	0.470	*White Noise*
ARIMA	12	9	15.831	16.918	0.105	*White Noise*
(2,1,0)	24	21	30.599	32.670	0.105	*White Noise*
	36	33	42.968	47.399	0.139	*White Noise*
	48	45	56.986	61.656	0.129	*White Noise*
ARIMA	12	8	11.497	15.507	0.243	*White Noise*
(2,1,1)	24	20	21.957	31.410	0.402	*White Noise*
	36	32	33.059	46.194	0.464	*White Noise*
	48	44	46.095	60.480	0.427	*White Noise*

In **Table 14**, the productivity variable p-value of 0.973518 is greater than the significance value of 0.05, meaning that the ARIMA (4,1,0) and ARIMA models (4,1,1) are normally distributed. For wider variables the harvest has a p-value value less than a significance value of 0.05, which means that all ARIMA models are not normally distributed. Based on the Central Limit Theorem, if the number of samples is greater than 30, then the distribution of the sample will be considered normal regardless of the probability value so that the analysis can continue **[27]**.

Table 14. **Results of the ARIMA Model Normality Test**

Variable	Model	Jarque-Bera	Df	P-Value
Productivity	ARIMA (4,1,0)	0.053678	2	0.973518
	ARIMA (4,1,1)	0.258807	2	0.878619
Large Harvest	ARIMA (0,1,1)	11.16750	2	0.003
Volumes	ARIMA (1,1,0)	7.7909	2	0.020
	ARIMA (1,1,1)	10.95605	2	0.004
	ARIMA (2,1,0)	7.790335	2	0.020
	ARIMA (2,1,1)	11.44708	2	0.003

3.5.2 Model Evaluation for Choosing the Best Model

The ARIMA model (4,1,1) has a smaller error value of 8.210% and the MASE of 1.355, so the ARIMA (4,1,1) model is well suited to predict the amount of shallot productivity. On the crop width variable, the smallest MAPE and MASE values are held by the ARIMA model (1,1,0) which is 12.381% and MASE is 0.706. So the ARIMA model (1,1,0) is the model used to predict the size of the upcoming harvest, with the parameter $\delta = 0.791808, a_1 = -1.020427, a_2 = -0,708352, a_3 = -0.587690, a_4 = -0.409083, \theta_1 = 0,465279$ is:

$$Y_t = (1-a_1 - a_2 - a_3 - a_4)\delta + (1 + a_1)Y_{t-1} + (a_2-a_1)Y_{t-2} + (a_3 - a_2)Y_{t-3} + (a_4 - a_3)Y_{t-4} - a_4 Y_{t-5} + \theta_1 \varepsilon_{t-1}$$

Next for the model equation ARIMA (1,1,0), where the parameter $\delta = 4.210130$, $a_1 = -0.406005$ so that the model equation ARIMA (1,1,0) is as follows:

$$Y_t = (1-a_1)\,\delta + (1 + a_1)Y_{t-1} - a_1 Y_{t-2}$$

Based on the process of the ARIMA method that has been carried out, the results of the forecast of the amount of production of shallot contained in the **Table 15**.

Table 15. Results of The Forecast of The Amount of Shallot Production in North Sumatra Province in January 2022 to March 2023

Year	Month	Results of Prediction (tons)
	January	5374.766
	February	5212.763
	March	5435.831
	April	5456.535
	May	5611.474
2022	June	5657.693
	July	5744.93
	August	5849.864
	September	5928.791
	October	6041.308
	November	6114.253
	December	6214.117
	January	6308.039
2023	February	6400.82
	March	6502.112

Table 15 shows that the production of shallot has an average production volume of 5856,886 tons. Where the lowest amount of shallot production occurred in February 2022 at 5212.763 tons and the highest production of shallot was in March 2023 at 6502.112 tons.

4. CONCLUSION

Based on the results of the research, it can be concluded that:

1. The variables affecting the amount of shallot production in the province of North Sumatra are the variables X_1 (sum of productivity) and the variable X_8 (large amount of crop). If the quantity of productivity and the size of the harvest increases, then the output of the amount of shallot production will also increase.

2. The output forecast of shallot production in the province of North Sumatra has increased. Where the lowest amount of shallot production occurred in April 2022 was 5212.763 tons and the highest volume of shallot production occurs in March 2023 was 6502.112 tons.

REFERENCES

[1] D. Deperiky, Santosa, R. A. Hadiguna, and Nofialdi, "Sinergi Supply Chain Yang Efektif : Literature Review Agroindustri Bawang Merah Di Sumatera Barat," *J. Teknol. Ind. Pertan.*, vol. 29, no. 2, pp. 124–131, 2019, doi: 10.24961/j.tek.ind.pert.2019.29.2.124.

[2] M. S. I. S. Hindarti and L. R. Maula, *Agribisnis Bawang Merah*, Accessed: Yogyakarta: Deepublish, 2020. doi: Available: https://www.google.co.id/books/edition/Agribisnis_Bawang_Merah/Jc0XEAAAQBAJ?hl=id&gbpv=1.

[3] M. Taufiq, R. Rahmanta, and S. F. Ayu, "Permintaan Dan Penawaran Bawang Merah Di Provinsi Sumatra Utara," *J. Agrica*, vol. 14, no. 1, pp. 104–115, 2021, doi: 10.31289/agrica.v14i1.4759.

[4] H. Hasri, J. Zakaria, and A. Arifin, "Faktor-Faktor Yang Mempengaruhi Produksi Bawang Merah Di Kecamatan Banggae Timur Kabupaten Majene," *Parad. J. Ilmu Ekon.*, vol. 3, no. 4, pp. 64–72, 2020, doi: 10.33096/paradoks.v3i4.599.

[5] M. Rijal, F. Jakfar, and W. Widyawati, "Analisis Faktor-Faktor Yang Mempengaruhi Produksi Usahatani Bawang Merah Di Desa Lam Manyang Kecamatan Peukan Bada," *J. Ilm. Mhs. Pertan.*, vol. 1, no. 1, pp. 488–497, 2016, doi: 10.17969/jimfp.v1i1.1321.

[6] D. Novita, M. Asaad, and T. Rinanda, "Potensi Dan Peluang Pengembangan Sentra Produksi Bawang Merah Provinsi

Sumatera Utara," *J. Agrica*, vol. 12, no. 2, p. 92, 2019, doi: 10.31289/agrica.v12i2.2870.

[7]   Haslina, Hasmah, K. W. F. Fitriani, M. A. Asbar, and Asrirawan, "PENERAPAN METODE ARIMA (AUTOREGRESSIVE INTEGRATED MOVING AVER- AGE) BOX JENKINS UNTUK MEMPREDIKSI PERTAMBAHAN JUMLAH PENDUDUK TANSMIGRAN (JAWA DAN BALI) DI KECAMATAN SUKAMAJU, KA-BUPATEN LUWU UTARA PROPINSI SULAWESI SELATAN," vol. 09, no. 1, pp. 55–67, 2018.

[8]   E. K. Sugiyanto *et al.*, "Konsep dan Praktik Ekonometrika Menggunakan Eviews," *Academia Publication*. p. 336, 2022.

[9]   N. Wahyuningsih, B. U. Daryono, and R. A. D. Zatadini, "Pendekatan Model Ekonometrika Untuk Memprediksi Indeks Saham Syariah Indonesia," *Univ. Airlangga*, vol. 294, 2017.

[10]   Z. Zuhri, "R ANCANG BANGUN APLIKASI REGRESI LINIER DAN KORELASI (Study Kasus : Luas Lahan Padi dan Hasil Produksi Padi Di Sumatera Utara)," *Serunai J. Ilm. Ilmu Pendidik.*, vol. 7, no. 1, pp. 116–122, 2021, doi: 10.37755/sjip.v7i1.428.

[11]   C. V. Bertan, A. K. T. Dundu, and R. J. M. Mandagi, "Pengaruh Pendayagunaan Sumber Daya Manusia (Tenaga kerja) terhadap Hasil Pekerjaan (Studi Kasus Perumahan Taman Mapanget Raya (Tamara)," *J. Sipil Statik*, vol. 4, no. 1, pp. 13–20, 2016.

[12]   B. Indrawan and R. Kaniawati Dewi, "Pengaruh Net Interest Margin (NIM) Terhadap Return on Asset (ROA) Pada PT Bank Pembangunan Daerah Jawa Barat Dan Banten Tbk Periode 2013-2017," *J. E-Bis*, vol. 4, no. 1, pp. 78–87, 2020, doi: 10.37339/e-bis.v4i1.239.

[13]   D. T. CAHYANINGRUM, "MODEL EKONOMETRIKA UNTUK ANALISIS KAPASITAS PENERBANGAN RUTE JEMBER – SURABAYA DI BANDARA NOTOHODINEGORO," Institut Teknologi Sepuluh Nopember, 2016.

[14]   Setiawati, "Analisis Pengaruh Kebijakan Deviden terhadap Nilai Perusahaan pada Perusahaan Farmasi di BEI," *J. Inov. Penelit.*, vol. 1, no. 8, pp. 1581–1590, 2021, [Online]. Available: https://stp-mataram.e-journal.id/JIP/article/view/308/261

[15]   Sugiharto and S. N. Anggraeny, "Mengukur Kualitas Akuntan Publik Menurut Pengguna Jasa Akuntan Publik," *J. Monex*, vol. 7, no. 1, pp. 1–9, 2018.

[16]   I. Fauzi, "Pengaruh Just in Time Dan Total Quality Management Terhadap Efisiensi Biaya Produksi Pada Restoran Joko Solo Medan," 2021.

[17]   I. Nurdin, Sugiman, and Sunarmi, "Penerapan Kombinasi Metode Ridge Regression (RR) dan Metode Generalized Least Square (GLS) untuk Mengatasi Masalah Multikolinearitas dan Autokorelasi," *J. Mipa*, vol. 41, no. 1, pp. 58–68, 2018.

[18]   M. Tiyani and Anneke Iswani Achmad, "Uji Stasioneritas Dickey Fuller Generalized Least Square pada Data Inflasi Indonesia dari Bulan Januari 2006 Hingga Bulan Mei 2020," *Pros. Stat.*, vol. 2, no. 2, pp. 27–33, 2020.

[19]   P. P. Krisma, Alviani Azhari, Muhammad Widagdo, "Perbandingan Metode Double Exponential Smoothing Dan Triple Exponential Smoothing Dalam Parameter Tingkat Error Mean Absolute Percentage Error (MAPE) dan Means Absolute Deviation (MAD) Alviani Krisma Putut Pamilih Widagdo Kata kunci-forecasting, Double Ex," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, pp. 81–87, 2019.

[20]   K. Auliasari, M. Kertaningtyas, and M. Kriswantono, "Penerapan Metode Peramalan untuk Identifikasi Permintaan Konsumen," *INFORMAL Informatics J.*, vol. 4, no. 3, p. 121, 2020, doi: 10.19184/isj.v4i3.14615.

[21]   K. V. Laura, "Peramalan banyaknya penabung di Credit Union Sumber Kasih Teraju dengan metode Box-Jenkins," Universitas Sanata Dharma, Yogyakarta, 2019.

[22]   S. Aktivani, "Uji Stasioneritas Data Inflasi Kota Padang," *Statistika*, vol. 20, no. 2, pp. 83–90, 2020.

[23]   M. H. Pamungkas, "Estimasi Parameter Model ARIMA Menggunakan Kalman Filter untuk Peramalan Permintaan Darah (Studi Kasus: UTD PMI Surabaya)," *Repos. ITS*, pp. 1–6, 2016, [Online]. Available: http://repository.its.ac.id/901/

[24]   Mamuroh, Sudarno, and H. Yasin, "IDENTIFIKASI BREAKPOINT DAN PEMODELAN AUTOREGRESSIVE STRUCTURAL CHANGE PADA DATA RUNTUN WAKTU (Studi Kasus Indeks Harga Konsumen Umum Kota Semarang Tahun 1994 – 2010)," vol. 30, no. 17, pp. 91–100, 2014.

[25]   Sulaminingsih, "Analisis Faktor-Faktor Yang Mempengaruhi Impor Beras DiKabupaten Berau," *J. Agroekoteknologi Trop. Lembab*, vol. 2, no. 2, pp. 61–70, 2020, doi: 10.35941/JATL.

[26]   H. Yusuf, "Uji Beda Preferensi Belanja Impor Barang Kiriman Atas Perubahan Batas Pembebasan Bea Masuk," *J. Perspekt. Bea Dan Cukai*, vol. 4, no. 1, pp. 20–35, 2020, doi: 10.31092/jpbc.v4i1.748.

[27]   R. Yuliyanti and E. Arliani, "Peramalan Jumlah Penduduk Menggunakan Model ARIMA," *Kaji. dan Terap. Mat.*, vol. 8, no. 2, pp. 114–128, 2022.