Forecasting poverty in East Java using vector autoregressive method and vector error correction model

A'yunin Sofro[1*], *Safira Diah Nur* Aidha[2], and *Khusnia Nurul* Khikmah[1,2]

[1]Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, East Java, Indonesia 60231
[2]Statistics Department, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, West Java, Indonesia 16680

Abstract. People experiencing poverty are people who are unable to fulfil their basic needs. A region with a dense population is prone to problems overcoming poverty. In this instance, the gross regional domestic product, the human development index, and the open unemployment rate are the variables impacting poverty. Therefore, more study is required to address this issue of poverty. The vector autoregressive and error correction models are two possible approaches. The East Java Central Bureau of Statistics provided the data, which included gross regional domestic product, human development index, open unemployment rate, and percentage of poverty. Forecasting the number of poverty people is obtained using estimates from data that can affect forecasting results. In this article, the best forecasting results are obtained with an RMSE value of 21.51062 using the vector error correction model, namely with a percentage of poverty value of 7.2619.

1 Introduction

The inability to fulfill basic things is called poverty, where the poverty rate in Indonesia in 2020 reached 10.19% [1]. This poverty directly affects the economy, one of which is the growth of household consumption expenditure on gross regional domestic product, where economic growth is a benchmark indicator of the success of development implementation [2,3]. In addition to gross regional domestic product, other influencing factors include the human development index and the open unemployment rate [4]. The increase in poverty in Indonesia can be predicted by using previous data on the factors that cause an increase in poverty. In this case, the author uses the factors of gross regional domestic product, human development index, and open unemployment rate. This forecasting can be used to prevent or overcome so that the influencing factors can be overcome first so that they do not experience a decline.

* Corresponding author: ayuninsofro@unesa.ac.id

Forecasting is an activity of guessing or estimating a future situation based on past and present circumstances that are needed as a decision support system to determine when an event will occur so that appropriate action can be taken [5]. This research uses data with four variables, namely the consumption rate of gross regional domestic product, human development index, open unemployment rate, and percentage of poverty. This variable was chosen because it refers to an existing article, where it is mentioned that the regression analysis of the variables of gross regional domestic product consumption rate, human development index, and open unemployment rate shows that the three variables influence the poverty rate.

Regression analysis with multivariate data can be predicted using the vector autoregressive method. This method was chosen because it moves from regression analysis, where in prediction, there are two methods, namely the ARIMA method for univariate data and the vector autoregressive method for multivariate data. Vector autoregressive is used because the behavior of a variable is expected to be affected by other variables, but the influence does not have an immediate impact but requires an interval or can be called a time lag. the advantages of the vector autoregressive model, include a simple model form where all variables in this model can be considered as endogenous variables. The vector autoregressive model uses ordinary least square (OLS) parameter estimation and better forecasting accuracy between other more complex simultaneous equation models [6, 7].

Other research [8, 9] shows that the vector error correction model has good forecasting results. The Vector Error Correction Model is a derivative of the Vector Autoregressive model where the vector error correction model must be stationary in the first difference. By using the Vector Error Correction Model, data can be analyzed for short-term and long-term relationships between variables. Referring to previous research conducted by [10, 11], Vector Autoregressive and Vector Error Correction Models provide good forecasting results. Both methods will be used in this study.

2 Methods

2.1 Vector autoregressive (VAR)

The vector autoregressive method is a combination of several autoregressive models whose variables will affect each other. The vector autoregressive model is based on the statistical properties of the data. In vector autoregressive, each endogenous variable in the system is considered as the lag value of all endogenous variables in the system [12]. Thus, the univariate autoregressive model is generalized to a vector autoregressive model consisting of multivariate time series variables. The vector autoregressive model is one of the dynamic linear models (MLD) that is widely used in the application of predicting economic variables both in the long term and in the medium to long term. In addition, this model can also be used to determine the causal relationship. [13] wrote the vector autoregressive model equation with k variables and order p or VAR(p) as follows:

$$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2.} + \cdots + \Phi_p Y_{t-p} + a_t \quad (1)$$

Where $Y_t = (Y_{1,t}, Y_{2,t}, \dots, Y_{k,t})$ denotes the vector Y_t of size $(k \times 1)$ of the time series variable, Φ_i denotes the matrix of size $(k \times k)$, $\Phi_0 = (\Phi_{10}, \Phi_{20}, \dots, \Phi_{k0})$ is a vector of dimension k and $a_t = \left(a_{1,t}, a_{2,t}, \dots, a_{k,t}\right)^{\mathrm{T}}$ is the error of size $(k \times 1)$ which is assumed to be multivariate normal with $E(\mu_t) = 0, E(\mu_t \mu_t^{\mathrm{T}}) = \sum \mu$ and $E\, E(\mu_t \mu_s^{\mathrm{T}}) = 0$ for $s \neq t$. The covariance matrix $\sum \mu$ must be positive definite.

The vector autoregressive model in equation (1) will be stable if:

$$\det\left(l_k - A_1 z - A_2 Z^2 - \cdots - A_p Z^p\right) \neq 0, |z| \leq 1 \qquad (2)$$

Equation (2) is the reverse characteristic polynomial of the VAR(p) model which has no roots and is on the unit circle.

2.2 Vector error correction models

Place the figure as close as possible after the point where it is first referenced in the text. If there is a large number of figures and tables, it might be necessary to place some before their text citation. The vector error correction model is a vector autoregressive model with cointegration constraints. Since there is a cointegration relationship in the vector error correction model, when there is an extensive range of short-term dynamic fluctuations, the vector error correction model expression can be restricted to the long-term behavior of the endogenous variables and converge to the cointegration relationship [9]. Assuming $y_t = (y_{1t}, y_{2t}, \dots, y_{kt})$ as a k-dimensional stochastic time series, $t = 1,2, \dots, T$ and $y_t \sim I(1)$ where $y_{it} \sim I(1)$, $i = 1,2, \dots, k$ is affected by the d-dimensional exogenous time series $x_t = (x_{1t}, x_{2t}, \dots, x_{dt})'$; then the vector error correction model can be formed as follows:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + B x_t + \mu_t \ , t = 1,2, \dots, T \qquad (3)$$

If y_t is not influenced by a d-dimensional exogenous time series, then the vector autoregressive model from equation (1) can be written as follows:

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \mu_t \ , t = 1,2, \dots, T \qquad (4)$$

With the cointegrating transformation of equation (2), we can obtain

$$\Delta y_t = \prod y_{t-1} + \sum_{i=1}^{p=1} \Gamma_i \Delta y_{t-1} + \mu_t \qquad (5)$$

where

$$\prod = \sum_{i=1}^{p} A_i - I \qquad (6)$$

and

$$\Gamma_i = \sum_{i=1}^{p} A_j \qquad (7)$$

If yt has a cointegrating relationship, then $\prod y_{t-1} \sim I(0)$ and equation can be written as

$$\Delta y_t = \alpha\beta' y_{t-1} + \sum_{i=1}^{p=1} \Gamma_i \Delta y_{t-1} + \mu_t \quad (8)$$

Where $\beta' y_{t-1} = ecm_{t-1}$ is the error correction term, which describes the long-run equilibrium relationship between variables, so it can be written with the following formula:

$$\Delta y_t = \alpha ecm_{t-1} + \sum_{i=1}^{p=1} \Gamma_i \Delta y_{t-1} + \mu_t \quad (9)$$

Equation (3) is a vector error correction model where each equation is an error correction model.

2.3 Root Mean Square Error (RMSE)

Root mean square error (RMSE) is a calculation method that is often used to obtain the minor error value between two different methods or equation models [14]. The formula is as follows:

$$\text{RMSE} = \sqrt{\overline{(f-o)^2}} \quad (10)$$

Where, f is the forecast and o is the observed value.

2.4 Data Sources

In this study, the data used are data on gross regional domestic product, human development index, open unemployment rate, and percentage of poverty in all provinces in Indonesia. Where gross regional domestic product, human development index, and open unemployment rate are independent variables, and the percentage of poverty is a dependent variable. Each variable is taken from 2004 to 2020 for East Java Province. The data is taken from the Central Statistics Agency (BPS). The variables used in this study are divided into 4, namely the gross regional domestic product variable ($Y_{1,t}$), the human development index variable ($Y_{2,t}$), the open unemployment rate variable ($Y_{3,t}$), and the percentage of poverty variable ($Y_{4,t}$). The dataset used is secondary data taken from the Central Bureau of Statistics.

3 Result and discussion

This study begins with data stationarity testing. Testing is done using the Augmented Dickey-Fuller method where the hypothesis used is $H_0 = |statistic\ t| < |critical\ value\ t|$ and $H_1 = |statistic\ t| > |critical\ value\ t|$. The test criteria will reject H_0 if the test statistic value of ADF has a value smaller than the critical region value or

$p - value < \alpha$. The test results show that the data is not yet stationary, so it is necessary to do differencing. The first difference results in the data and the following Table 1.

Table 1. ADF first difference test results.

	Gross regional domestic product	Human development index	Open unemployment rate	Percentage of poverty
Statistic t	-7.668	-4.519	-2.878	-4.936
Critical Value t	-1.95	-1.95	-1.95	-1.95

From Table 1, it can be seen that all data has been stationary because of the results of $|statistic\ t| > |critical\ value\ of\ t|$. So, the data becomes stationary in the first difference. Then, the requirements for the vector error correction model are also met because the data has been stationary in the first difference. Furthermore, the differencing results are modeled using vector autoregressive. The first step in this modeling is to determine the optimal lag based on the optimal AIC value, which is 1.260790 at lag-1 and -0.90666 at lag-2 so that the optimal lag of the data used is two or it can be said to be a VAR(2) model. Furthermore, parameter estimation is carried out, and parameter estimates for the VAR(2) model with four variables can be obtained in Table 2.

Table 2. Parameter estimation results of vector autoregressive.

	Y			
	1	2	3	4
Gross regional domestic product.l1	0.1328	−0.1552	0.0364	−0.2625
Human development index.l1	0.0105	0.0680	−0.0419	−0.0365
Open unemployment rate.l1	−0.0006	0.7176	0.5498	0.4432
Percentage of poverty.l1	−0.3816	−3.1447	0.1361	0.9063
Gross regional domestic product.l2	−0.1484	−0.9799	0.0981	0.3362
Human development index.l2	−0.1716	−2.1576	−0.0542	0.031
Open unemployment rate.l2	0.3015	2.3788	−0.0166	−0.5474
Percentage of poverty.l2	0.0114	−0.9414	−0.1232	0.1349

Const.	22.138	278.844	8.0933	−0.9967

The results of parameter estimation are tested with Granger Causality, which aims to determine the relationship between related variables. If the variables are not interconnected, then the vector autoregressive model cannot be formed, and the results are obtained in Table 3.

Table 3. Granger causality test results.

	$p-value$
Gross regional domestic product	4.99×10^{-10}
Human development index	9.83×10^{-1}
Open unemployment rate	8.04×10^{-4}
Percentage of poverty	1.37×10^{-2}

Based on the test results obtained, gross regional domestic product affects the human development index, open unemployment rate, and percentage of poverty. human development index does not affect gross regional domestic product, the open unemployment rate, and percentage of poverty; open unemployment rate affects gross regional domestic product, human development index, and percentage of poverty. percentage of poverty does not affect gross regional domestic product, human development index, and open unemployment rate. So the model that is formed is:

$$\mathrm{Y}_{1.\mathrm{t}} = 0.1328\text{Gross regional domestic product.l1} + 0.0105\text{Human development index.l1} - 0.0006\text{Open unemployment rate.l1} - 0.3816\text{Percentage of poverty.l1} - 0.1484\text{Gross regional domestic product.l2} - 0.1716\text{Human development index.l2} + 0.3015\text{Open unemployment rate.l2} + 0.0114\text{Percentage of poverty.l2} + 22.138$$

$$\mathrm{Y}_{2.\mathrm{t}} = -0.1552\text{Gross regional domestic product.l1} + 0.0680\text{Human development index.l1} + 0.7176\text{Open unemployment rate.l1} - 3.1447\text{Percentage of poverty.l1} - 0.9799\text{Gross regional domestic product.l2} - 2.1576\text{Human development index.l2} + 2.1576\text{Open unemployment rate.l2} - 0.9414\text{Percentage of poverty.l2} + 278.844$$

$$\mathrm{Y}_{3.\mathrm{t}} = 0.0364\text{Gross regional domestic product.l1} - 0.0419\text{Human development index.l1} + 0.5498\text{Open unemployment rate.l1} + 0.1361\text{Percentage of poverty.l1} + 0.0981\text{Gross regional domestic product.l2} - 0.0542\text{Human development index.l2} - 0.0166\text{Open unemployment rate.l2} - 0.1232\text{Percentage of poverty.l2} + 8.0933$$

$$Y_{4.t} = -0.2625\text{Gross regional domestic product.l1} - 0.0365\text{Human development index.l1} + 0.4432\text{Open unemployment rate.l1} + 0.9063\text{Percentage of poverty.l1} + 0.3362\text{Gross regional domestic product.l2} + 0.031\text{Human development index.l2} - 0.5474\text{Open unemployment rate.l2} + 0.1349\text{Percentage of poverty.l2} - 0.9967$$

The second model to be modeled is the vector error correction model, starting with the determination of the Optimal Lag. The optimal lag obtained the optimal AIC value of -0.90666 at lag-2 so that the optimal lag of the data used is 2. Furthermore, stability testing is carried out. The vector error correction model is said to be stable if the modulus value is at a radius < 1. If the most significant modulus value is less than one and is located at the optimal point, then the vector autoregressive model is stable, and the following results are obtained in Table 4.

Table 4. Stability test results.

	Modulus		Modulus
[1]	1.4314739	[5]	0.4453617
[2]	1.4314739	[6]	0.4453617
[3]	1.0153875	[7]	0.4031710
[4]	0.5575469	[8]	0.1289675

Based on Table 4, modulus values 1 to 3 show more than 1, but after that, the modulus gradually decreases. It is concluded that the vector error correction model used is relatively stable. After the stability test is carried out, cointegration testing will be carried out using the Johansen cointegration test. The test results are in Table 5.

Table 5. Johansen cointegration test results.

	Statistic trace	Critical value
$r \leq 3$	4.9043	9.24
$r \leq 2$	16.7792	19.96
$r \leq 1$	57.1662	34.91
$r = 0$	114.8108	53.12

Based on the results in Table 5, at $r = 0$ and $r \leq 1$, the significance value of the trace is more significant than its value (with a significance level of 5%). Then, it can be interpreted that the data has a long-run equilibrium. Then, the model can be estimated parameters. The vector error correction model parameter estimates obtained are as follows.

Table 6. Vector error correction model parameter estimation results.

	Gross regional domestic product	Human development index	Open unemployment rate	Percentage of poverty
Gross regional domestic product.l1	0.177	−0.39	−0.095	−0.168
Human development index.l1	−0.003	0.095	0.0013	−0.045
Open unemployment rate.l1	−0.034	0.514	−0.102	0.353
Percentage of poverty.l1	−0.39	−0.33	0.305	0.095
Gross regional domestic product.l2	−0.125	−0.14	0.048	0.16
Human development index.l2	−0.069	−1.91	−0.031	0.017
Open unemployment rate.l2	0.141	0.176	−0.078	−0.026
Percentage of poverty.l2	−0.07	0.313	0.106	−0.197
Const.	−0.155	1.259	0.222	−0.325

After obtaining parameter estimates, the vector error correction model can be formed as follows:

$$\begin{bmatrix} \text{Gross regional domestic product}_t \\ \text{Human development index}_t \\ \text{Open unemployment rate}_t \\ \text{Percentage of poverty}_t \end{bmatrix} = \begin{bmatrix} -0.155 \\ 1.259 \\ -0.108 \\ 0.155 \end{bmatrix} + \begin{bmatrix} 0.177 & -0.003 & -0.034 & -0.39 & -0.125 & -0.069 & 0.141 & -0.07 \\ -0.39 & 0.095 & 0.514 & -0.33 & -0.14 & -1.91 & 0.176 & 0.313 \\ -0.095 & 0.0013 & -0.102 & 0.305 & 0.048 & -0.031 & -0.078 & 0.106 \\ -0.168 & -0.045 & 0.353 & 0.095 & 0.16 & 0.017 & -0.026 & -0.197 \end{bmatrix} \begin{bmatrix} \text{Gross regional domestic product.l1} \\ \text{Human development index.l1} \\ \text{Open unemployment rate.l1} \\ \text{Percentage of poverty.l1} \\ \text{Gross regional domestic product.l2} \\ \text{Human development index.l2} \\ \text{Open unemployment rate.l2} \\ \text{Percentage of poverty.l2} \end{bmatrix}$$

The model obtained based on the analysis carried out is then carried out forecasting. Vector autoregressive and vector error correction model for forecasting, forecasting is

carried out gross regional domestic product, human development index, open unemployment rate, and percentage of poor people in East Java for the next six years. Then, the forecasting results with the VAR(2) model are obtained in Table 7.

Table 7. Forecasting results using vector autoregressive and vector error correction model parameter estimation results.

Year	Gross regional domestic product		Human development index		Open unemployment rate		Percentage of poverty	
	VAR	VECM	VAR	VECM	VAR	VECM	VAR	VECM
2021	7.840	6.409	93.229	90.842	−0.095	4.871	−0.168	11.550
2022	8.587	6.572	96.615	92.954	0.001	4.700	−0.045	10.272
2023	5.055	5.486	48.608	50.238	−0.102	3.692	0.353	10.300
2024	3.692	4.965	41.072	44.835	−0.078	4.241	−0.026	11.542
2025	11.215	6.986	142.001	129.258	0.106	6.435	−0.197	10.620
2026	14.987	8.027	168.217	146.104	0.222	5.711	−0.325	7.2619

Based on Table 7, the forecasting results show the results of forecasting for the next five years, where each method has different results. Then, we will check the RMSE value to find out the best model. By using equation (25), the RMSE value for the vector autoregressive method is 26.91299, while the RMSE value for the vector error correction model method is 21.51062. It can be concluded from this study that the most appropriate forecasting method is to use the vector error correction model with an optimal lag of 2.

This study provides the results of modeling and forecasting poverty data in East Java, Indonesia, based on the vector autoregressive method and vector error correction model. Based on the forecasting results, this study found that in 2021, the level of gross regional domestic product and the percentage of poor people decreased, while the community development index and the open unemployment rate increased. Gross regional domestic product and open unemployment rate significantly affect the poverty rate in East Java. However, when viewed from long-term forecasting, namely until 2026, the community development index decreased in 2022 but gradually improved in subsequent years. This results in the percentage of poor people also reducing. This study also shows that the percentage of poor people in East Java is influenced by gross regional domestic product, community growth index, and open unemployment rate.

In addition, we found that joint modeling, based on the analysis results, provides more complicated information. The data analysis based on the statistical model also shows that

the vector error correction model shows more accurate forecasting results than the vector autoregressive model. In addition, the forecasting results can suggest considerations for policymakers in determining steps and attitudes toward the percentage of poor people in East Java.

4 Conclusion

Two different methods can be used, namely vector autoregressive and vector error correction models. Where both methods have standard multivariate variables, two different forecasting models can be obtained. From the different forecasting results, the best method can be selected. In this study, the best model was obtained using the VAR(2) model with an RMSE of 1.1738. The smallest RMSE value indicates that the model has a small error value so that the results obtained are expected to be close to future data. In future research, forecasting using the vector autoregressive and vector error correction model methods can be done using more variables and a more comprehensive range of coverage. This is done to determine the best forecasting and the effect of each variable on the results of forecasting other variables.

References

[1] R. Izzati Al, Situasi Kemiskinan Selama Pandemi (The SMERU Research Institute, 2021)

[2] N. Nursini, Development Studies Research **7**, 1 (2020)

[3] E.-J. De Bruijn, G. Antonides, Theory Decis. **92**, 1 (2022)

[4] D. Dahliah, A. N. Nur, Golden Ratio of Social Science and Education **1**, 2 (2021)

[5] J. Dupuy et al., Ann For Sci, **77**, 2 (2020)

[6] A. Ibrahim, R. Kashef, M. Li, E. Valencia, E. Huang, Journal of Risk and Financial Management **13**, 9 (2020)

[7] S. Guefano, J. G. Tamba, T. E. W. Azong, L. Monkam, Energy **214** (2021)

[8] M. Ding, H. Zhou, H. Xie, M. Wu, Y. Nakanishi, R. Yokoyama, Neurocomputing **365** (2019)

[9] P. Giudici, P. Pagnottoni, Appl Stoch Models Bus Ind. **36**, 1 (2020)

[10] E. N. Ni'mah, URECOL **12** (2017)

[11] A. P. Desvina, Jurnal Sains Teknologi dan Insdustri **13**, 2 (2016)

[12] W. B. Nicholson, I. Wilms, J. Bien, D. S. Matteson, The Journal of Machine Learning Research **21**, 1 (2020)

[13] H. Lütkepohl, New introduction to multiple time series analysis (Springer Science & Business Media, 2005)

[14] M. Ćalasan, S. H. E. A. Aleem, A. F. Zobaa, Energy Convers Manag, **210** (2020)