BAREKENG: Jurnal Ilmu Matematika dan Terapan
March 2022 Volume 16 Issue 1 Page 099–112
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss1pp099-112

PRINCIPAL COMPONENT ANALYSIS-VECTOR AUTOREGRESSIVE INTEGRATED (PCA-VARI) MODEL USING DATA MINING APPROACH TO CLIMATE DATA IN THE WEST JAVA REGION

Devi Munandar[1], Budi Nurani Ruchjana[2*], Atje Setiawan Abdullah[3]

[1,2]Mathematics Department, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran
[3]Computer Science Department, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran
Jln. Raya Bandung Sumedang, km 21, Jatinangor, Sumedang, 45363, Indonesia

Corresponding author e-mail: [2] budi.nurani@unpad.ac.id*

Abstract. *Over a long time, atmospheric changes have been caused by natural phenomena. This study uses the Principal Component Analysis (PCA) model combined with Vector Autoregressive Integrated (VARI) called the PCA-VARI model through the data mining approach. PCA reduces ten variables of climate data into two principal components during ten years (2001-2020) of climate data from NASA Prediction Of Worldwide Energy Resources. VARI is a non-stationary multivariate time series to model two or more variables that influence each other using a differencing process. The Knowledge Discovery in Database (KDD) method was conducted for empirical analysis. Pre-processing is an analysis of raw climate data. The data mining process determines the proportion of each component of PCA and is selected as variables in the VARI process. The post processing is by visualizing and interpreting the PCA-VARI model. Variables of solar radiation and precipitation are strongly correlated with each measurement location data. A forecast of the interaction of variables between locations is shown in the results of Impulse Response Function (IRF) visualization, where the climate of the West Java region, especially the Lembang and Bogor areas, has strong response climate locations, which influence each other.*

Keywords: *climate, data mining, forecasting, IRF, KDD, PCA-VARI, variable reduction*

Article info:

Submitted: 9th November 2021 *Accepted: 3rd February 2022*

How to cite this article:
D. Munandar, B. N. Ruchjana and A. S. Abdullah, "PRINCIPAL COMPONENT ANALYSIS-VECTOR AUTOREGRESSIVE INTEGRATED (PCA-VARI) MODEL USING DATA MINING APPROACH TO CLIMATE DATA IN THE WEST JAVA REGION", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 1, pp. 099-112, Mar, 2022.

1. INTRODUCTION

Climate phenomena are long-term (minutes to months) changes in the atmosphere. The difference between weather and climate is a matter of time. Weather is the condition of the atmosphere over a short period, and climate is how the atmosphere changes over a relatively long period. Climate change, in long-term averages of daily weather such as temperature, humidity, precipitation, cloudiness, brightness, visibility, wind, and atmospheric pressure as at high and low pressure. Climatic variables can be observed based on the order of location and time as the realization of a stochastic process with a large data size and many measurement variables. A Principal Component Analysis (PCA) process is needed to reduce variables through linear combinations or orthogonal transformations like application in marine climate [1]. The relationship between climate variables and time index can be analyzed using univariate and multivariate time series models. Climatic variables such as temperature, humidity, air pressure, solar radiation, and wind speed vary due to the nature of the monthly climate data, which is uneven and has quite large dimensions. Using Knowledge Discovery in Database (KDD), the data mining approach includes three steps of preprocessing, data mining, and postprocessing can be used on climate phenomena with some of these variables. Multivariate time series models such as temperature variables observed at several observation stations with a fairly high correlation value can use the Vector Autoregressive (VAR) or Vector Autoregressive Integrated (VARI) models [2]. To model the impact of climate variability (rainfall, maximum temperature, and relative humidity) on the number of malaria sufferers with the estimated variance decomposition showing varying degrees of dependence of the number of malaria sufferers on climate variables that are large enough from the variability at a maximum temperature [3] [4].

Meanwhile, the use of PCA is integrated with VAR in Econometrics with an estimate of the impact on humans from climate change, affecting the economy, environment, social community, and analysis of variables that influence the dominant factors in development. The PCA analysis integrated with the VAR model was carried out on the research on exhaust gas emissions produced by Iran, including the top ten CO2 emitters globally, using data from 1992 to 2015 on each subsystem. The interaction between subsystems was investigated through the short and long term with the Vector Autoregressive (VAR) model [5].

The description above, the purpose of this study is to integrate the PCA model with the Vector Autoregressive Integrated (VARI) using the data mining approach [6] on climate data in the West Java region of Indonesia to produce a multivariate time series model based on variable reduction. Implementing the PCA-VARI model further simplifies forecasting and interpreting the model by combining two different models. PCA reduces the variables in higher dimensions to fewer dimensions with a linear combination of the initial data and obtains results based on the proportions of the principal components. The principal component is obtained of the cumulative percentage of the total variance, getting new, more straightforward data for VARI modeling [7][8]. Previous studies' results explain that this study uses PCA and VARI approaches to assess the relationship between variables in climate data in the West Java region.

2. RESEARCH METHODS

2.1 Principal Component Analysis

Before determining the main components based on the correlation matrix and how the matrix is formed, first, the KMO test is a statistical method [9] that shows the proportion of variance in the variables that the underlying factors may cause

$$KMO = \sum_{i \neq j} \tau_{ij}^2 \Big/ \sum \sum_{i \neq j} \tau_{ij}^2 + \sum \sum_{i \neq j} \alpha_{ij}^2 , \; i = 1,2,...,n; \; j = 1,2,..,n \tag{1}$$

τ_{ij}^2 is the simple correlation coefficient between variable-*i* and variable-*j*, α_{ij}^2 is partial correlation coefficient between variable-*i* and variable-*j*. PCA explains the covariance variance matrix structure on a group of variables into a linear combination of these variables so that it becomes less [10]. If there are p components, then by performing PCA, there will be k components, where $p > k$. Thus, k components can explain and replace the initial variables, which amount to p [11]. We have a vector $X' = \{X_1, X_2, X_2, ...X_n\}$ that has a covariance matrix Σ with p eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_p \geq 0$. Furthermore, there is a linear combination, with the transformation that occurs as much as p equations as follows:

$$\begin{aligned}\eta_1 &= \mathbf{a}_1'\mathrm{X} = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ \eta_2 &= \mathbf{a}_2'\mathrm{X} = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \\ &\cdot \\ &\cdot \\ &\cdot \\ \eta_p &= \mathbf{a}_p'\mathrm{X} = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p\end{aligned} \tag{2}$$

If there are p variables, then p can be made linear combinations such that each variable is uncorrelated or linearly independent by taking only k variables, so that obtained $\mathrm{Var}(\eta_i) = \mathbf{a}_i'\Sigma\mathbf{a}_i$ and $\mathrm{Cov}(\eta_i, \eta_k) = \mathbf{a}_i\Sigma\mathbf{a}_k$ with *i,k =1,2,...p*. To get the i^{th} principal component, a chosen linear combination maximizes $\mathrm{Var}(\mathrm{a}_i'\mathrm{X})$ under the condition $\mathrm{a}_i'\mathrm{a}_i = 1$ and all covariates for $k < i = 0$ or $\mathrm{Cov}(\mathrm{a}_i'\mathrm{X}, \mathrm{a}_k'\mathrm{X}) = 0$ for $k < i$. Linear algebra [12] can be defined if $\mathrm{X}' = \{X_1, X_2, X_2, ...X_n\}$ there is a covariance matrix Σ with pairs of eigenvalues and eigenvectors $(\lambda_1, \mathbf{a}_1), (\lambda_2, \mathbf{a}_2), ...(\lambda_p, \mathbf{a}_p)$ where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_p \geq 0$ as well as $\eta_1 = \mathbf{a}_1'\mathrm{X}, \eta_2 = \mathbf{a}_2'\mathrm{X}, ..., \eta_p = \mathbf{a}_p'\mathrm{X}$ is the first principal component, then to obtain another principal component $\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \lambda_1 + \lambda_2 \geq \lambda + \geq ... + \lambda_p = \sum_{i=1}^{p} \mathrm{Var}(\eta_i)$ the proportion of the total variance for the k^{th} principal component is $\lambda_k / \lambda_1 + \lambda_2 + ... + \lambda_p$.

2.2 Vector Autoregressive Integrated (VARI)

The VARI model following the Box-Jenkins procedure is used to analyze and forecast stationary time series data. This method consists of three steps; the first is to identify the model by checking the stationary data. If the data is not stationary, then a differencing process is carried out to achieve stationary data. Then proceed with looking at the data plot using the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). Second, estimate the model parameters. Third, checking diagnostics is useful to see whether the model meets or not. The following process, which is the main goal of Box-Jenkin, is to use the model for forecasting. A sufficient condition for a model to be used for forecasting is that the residuals must be white noise and normal. If you do not meet the criteria for these two conditions, then re-identify [13].

Model Identification and Specification

In VARI modeling, the assumption that must be met is that the stationarity of time series data is the basic method before data analysis is carried out. Stationary time series data has no trend, seasonal pattern, and constant variance over time, which means that the stationary time series does not depend on the unit root. To overcome the stationarity of the data on average, the Augmented Dickey-Fuller test can cover high-order autoregressive including ΔZ_{t-p} in the model with $\gamma = 0$

$$\Delta Z_t = Z_t - Z_{t-1} = \alpha + \beta t + \gamma Z_{t-1} + \delta_1 \Delta Z_{t-1} + \delta_2 \Delta Z_{t-2} + ... + \delta_p \Delta Z_{t-p} \tag{3}$$

Z_t is time series data, so it can be determined through linear regression of ΔZ_t to *t* and Z_{t-1} by testing whether $\gamma \neq 0$. If $\gamma = 0$ then there is a unit root. If there is no unit root, $-1 < 1 + \gamma < 1$ then the process can be said to be stationary [14]. Statistical hypothesis testing is done by *H*0: ϕ = 1, there is a unit root or non-stationary data in the mean and *H*1: ϕ < 1, there is no unit root or stationary data in the mean. Then the error rate (α), is a critical value, and the test statistic

$$\varsigma_{test} = \frac{\hat{\phi} - 1}{\mathrm{SE}(\hat{\phi})} \tag{4}$$

The significance level rejects the null hypothesis with a *p-value* with a significance level of not less than 0.05 [15].

We can apply the Box-Cox transformation test if the data is not stationary in the variance [16]. The Box-Cox transformation is the exponential lambda (λ) , which varies from -5 to 5. All values are considered and the optimal value for the selected data. While the optimal value is the one that produces the best approximation of the normal distribution curve. The transformation of y has the formula

$$y(\lambda)=\begin{cases}\dfrac{y^{\lambda}-1}{\lambda}, \text{ jika } \lambda\neq 0;\\ \ln y, \text{ jika } \lambda=0.\end{cases} \quad (5)$$

Before the VARI modeling was built, the level of accuracy was evaluated by calculating the lag value, generally by obtaining the *p-value*. In this study, to assess the feasibility of the climate forecast model, the application was chosen for the calculation of the Aikake Information Criterion (AIC) for several *m* independent variables where the AIC value is generally defined using the following mathematical equation

$$AIC(m)=\ln\left|\hat{\varsigma}(m)\right|+\frac{2k^2m}{T} \quad (6)$$

where $\hat{\varsigma}(m)=\dfrac{\sum_{t=1}^{T}\hat{\varepsilon}_t(\hat{\varepsilon}_t)}{T}$

$\hat{\varsigma}(m)$ the residual covariance matrix $\hat{\varepsilon}_t$ is the error value, k the number of parameters in the model, m number of observations at a time t [17].

Parameter Estimation

The VARI model Maximum Likelihood (MLE) function uses the combined probability density function to obtain parameter estimates. Assumptions in the regression model are related to random error values in a multivariate normal distribution. For a VARI model with two variables, the error component must meet the assumptions $e(t)\overset{iid}{\sim}\boldsymbol{N}(\mathbf{0},\sigma^2\boldsymbol{I}_2)$. Equality $\dot{Z}_{t_{(N\times1)}}=\Phi_{(N\times N)}\dot{\boldsymbol{Z}}_{(t-1)_{(N\times1)}}+e_{(t)_{(N\times1)}}$ or in linear equality $\boldsymbol{Y}=\mathbf{X\Phi}+\mathbf{E}$ as the initial VAR model equation, by making changes to the matrix Y,X,Φ, and E estimation in parameter solution using MLE:

$$\boldsymbol{y}-\boldsymbol{x\phi}=\boldsymbol{e}\sim N(\mathbf{0},\sigma^2\mathbf{I})$$
$$(\boldsymbol{y}-\boldsymbol{x\phi})\sim N(\mathbf{0},\sigma^2\mathbf{I}) \quad (7)$$

the probability density function for the equation (7), which becomes a function *likelihood*: [18]

$$f(\mathbf{e})=\frac{1}{\sigma\sqrt{2\pi}}\,exp\left(-\frac{\mathbf{e}^2}{2\sigma^2}\right)$$
$$L(\boldsymbol{\phi},\sigma_e^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_e^2}}exp\left(-\frac{\mathbf{e}_{\mathrm{i}}^2}{2\sigma_e^2}\right)$$
$$L(\boldsymbol{\phi},\sigma_e^2)=\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}exp\left[-\left(\frac{1}{2\sigma^2}\right)(\boldsymbol{y}-\boldsymbol{x\phi})^T(\boldsymbol{y}-\boldsymbol{x\phi})\right]$$
$$\ln\left(L(\boldsymbol{\phi},\sigma_e^2)\right)=-\frac{n}{2}ln(2\pi\sigma_e^2)-\left(\frac{1}{2\sigma_e^2}\right)(\boldsymbol{y}-\boldsymbol{x\phi})^T(\boldsymbol{y}-\boldsymbol{x\phi}) \quad (8)$$

Next, derive the equation (8) to $\boldsymbol{\phi}$:

$$\frac{\partial\left(ln\left(L(\boldsymbol{\phi},\sigma_e^2)\right)\right)}{\partial(\boldsymbol{\phi})}=-\frac{1}{\sigma_e^2}(\boldsymbol{y}-\boldsymbol{x\phi})^T(-\boldsymbol{x}) \quad (9)$$

by maximizing the function *likelihood* pada equation (9), then :

$$-\frac{1}{\sigma_e^2}(\boldsymbol{y}-\boldsymbol{x\phi})^T(-\boldsymbol{x})=0$$
$$-\boldsymbol{y}^T\boldsymbol{x}+\boldsymbol{\phi}^T\boldsymbol{x}^T\boldsymbol{x}=0$$
$$\boldsymbol{\phi}^T=\boldsymbol{y}^T\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{x})^{-1} \quad (10)$$

then remove transpose on $\boldsymbol{\phi}$, equation (10) become :

$$\boldsymbol{\phi}=(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}$$

so an estimator of $\boldsymbol{\phi}$ become:

$$\hat{\boldsymbol{\phi}}=(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}. \quad (11)$$

PCA-VARI Model

The PCA process obtained in equation (2) is a linear combination from the Vector Autoregressive Integrated (VARI) process. VARI model develops the Autoregressive Integrated (ARI) model, a form of regression results against itself and other variables in the previous period on non-stationary data. Suppose the VAR(*p*) model is not stationary. In that case, the differencing process is carried out k times so that the data

is stationary, and the VARI(p,k) model is obtained with a formulation with n variables. If p=1 and k=1, then the VARI(1,1) model can be obtained by the equation in matrix form as follows:

$$\begin{bmatrix} \dot{Z}_{1,t} \\ \dot{Z}_{2,t} \\ \vdots \\ \dot{Z}_{N,t} \end{bmatrix} - \begin{bmatrix} \dot{Z}_{1,t-1} \\ \dot{Z}_{2,t-1} \\ \vdots \\ \dot{Z}_{N,t-1} \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1N} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{N1} & \phi_{N2} & \cdots & \phi_{NN} \end{bmatrix} \left(\begin{bmatrix} \dot{Z}_{1,t-1} \\ \dot{Z}_{2,t-1} \\ \vdots \\ \dot{Z}_{N,t-1} \end{bmatrix} - \begin{bmatrix} \dot{Z}_{1,t-2} \\ \dot{Z}_{2,t-2} \\ \vdots \\ \dot{Z}_{N,t-2} \end{bmatrix} \right) + \begin{bmatrix} e_{1,t} \\ e_{2,t} \\ \vdots \\ e_{N,t} \end{bmatrix}$$

or equal to

$$\dot{Z}_t - \dot{Z}_{t-1} = \mathbf{\Phi}_1(\dot{Z}_{t-1} - \dot{Z}_{t-2}) + \mathbf{e}(t) \tag{12}$$

It is assumed that the error value is normally distributed, i.e. $e_t \overset{iid}{\sim} N(0,\sigma^2)$, and $\dot{Y}_t = \dot{Z}_{1,t} - \dot{Z}_{1,t-1}$, then equation (12) satisfies with $t = 2,3,...,T$ and VARI equation

$$\dot{\mathbf{Y}}_{t_{((N\times(T-1))\times 1)}} = \dot{\mathbf{Y}}_{t-1_{((N\times(T-1))\times(N\times N))}} \mathbf{\Phi}_{((N\times N)\times 1)} + \mathbf{e}_{t_{((N\times(T-1))\times 1)}} \tag{13}$$

Diagnostic Checking

The feasibility test of the model is very well done through residual series analysis, which is white noise resulting parameter estimates assumed to be randomly uncorrelated with zero mean and constant variance.

The Ljung-Box test is a diagnostic checking that uses all residual ACF samples as a unit for testing the null hypothesis through 2 steps, namely; Hypothesis $H_0 : \rho_1 = ... = \rho_k = 0$, H_1: there is at least one value $\rho_k \neq 0$ $k = 1,2,3,...,n$ (there is a correlation between residuals, and k is the order of time); statistic test

$$Q = n(n+2)\sum_{k=1}^{m}(n-k)^{-1}\hat{\rho}_k^2 \tag{14}$$

Model statistical test Q modified to determine statistical test H_0 following distribution criteria $\chi^2(K-m)$ with $m = p+q$ and $p-value > \alpha_{0.05}$ shows that the residuals in the multivariate model meet the white noise assumption.

Granger causality test is a statistical hypothesis test to determine whether a time series is useful for analyzing the causal relationship between the observed variables [19]. If we assume that X and Y are two stationary series, to determine whether X Granger causes Y, we have to look at autoregressive and determine the lag length. Granger causality can be defined as follows

$$y_t = \alpha_0 + \sum_{i=1}^{p}\alpha_i y_{t-i} + \sum_{j=1}^{q}\beta_i x_{t-j} + a_t \tag{15}$$

$$y_t = \delta_0 + \sum_{i=1}^{p}\delta_i y_{t-i} + e_t \text{ (null hypothesis model)} \tag{16}$$

while the F-statistic uses the following equation

$$F = \frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/(n-k)} \tag{17}$$

$ESS_R = \sum_{i-1}^{T}\hat{a}_t^2$, $ESS_{UR} = \sum_{i-1}^{T}\hat{e}_t^2$, n = observation data, k = estimation parameters, q = lag length. Decline decision H_0 if $F-test > p-value > \alpha_{0.05}$.

Impulse Response Function (IRF)

The impulse response function (IRF) is the reaction of any dynamic system in response to some external change. The impulse-response function is used to test the impact of shock or response a variable on itself and other variables of the short-term and long-term VAR models. The VAR model can be written in vector (∞) as

$$\Gamma_t = \mu + \varepsilon_t + \Psi_1\varepsilon_{t-1} + \Psi_2\varepsilon_{t-2} + \Psi_3\varepsilon_{t-3} + ...$$

The Ψ_s matrix is interpreted as

$$\frac{\partial \Gamma_{t+s}}{\partial \varepsilon_t'} = \Psi_s \tag{18}$$

The row *i*, column *j* is the identification element Ψ_s that set an increase of one unit of innovation to *j*th variable at date *t* (ε_{jt}) for the value of the *i*th variable at time $t+s(\Gamma_{i,t+s})$, while retaining all other innovations at constant dates. The plot of the *i*th row, *j*th column of the elements Ψ_s as a function of *s* is called the impulse response function.

2.3 PCA-VARI using Knowledge Discovery in Databases (KDD)

Data mining collects important information from large data using statistical mathematical methods to utilize artificial intelligence technology. Alternative data mining is also known as Knowledge discovery (mining) in databases (KDD) [20] for PCA-VARI process in West Java climate shown in Figure 1. Data mining has several functions. The primary function is descriptive, which is to understand more about the observed data to understand the behavior of the data used to find out the characteristics of the data in question to find specific hidden patterns and predictive functions [21][22]. Namely, how a process will later find patterns of data. This study KDD process on climate data has several steps: preprocessing by cleaning, aggregating, and integrating data. The data mining step is to carry out the PCA-VARI integration process. The data generated from the PCA process is the input for the VARI process. The post processing step is that knowledge presentation performs the process of visualizing and interpreting the data resulting from the mining process.

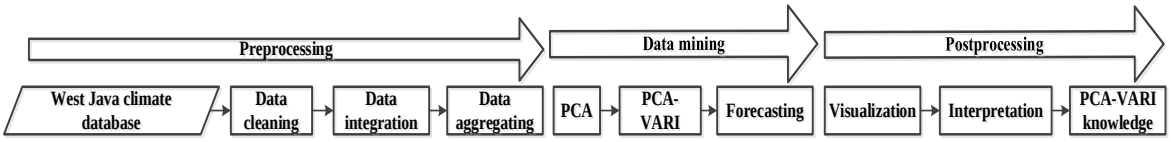

Figure 1. PCA-VARI using Knowledge Discovery in Databases

3. RESULTS AND DISCUSSION

3.1 Preprocessing

The data mining approach is used to complete the modeling with PCA-VARI. The data used is sourced from climatic data in West Java, Indonesia, consisting of the Lembang (latitude = -6.8117, longitude = 107.6175), Bogor (latitude = -6.5944, longitude = 106.7892), and Tasikmalaya (latitude = -7.3274, longitude = 108.2207) from 2001-2020 with daily time intervals. The data is obtained from secondary data from POWER NASA[1].

Table 1. Variables for Lembang, Bogor, and Tasikmalaya Locations

Measurement Station	Lembang (LM)	Bogor (BG)	Tasikmalaya (TS)
Variable	LM_1 : UV Index	BG_1 : UV Index	TS_1 : UV Index
	LM_2 : Temperature (2 meters)	BG_2 : Temperature (2 meters)	TS_2 : Temperature (2 meters)
	LM_3 : Dewpoint	BG_3 : Dewpoint	TS_3 : Dewpoint
	LM_4 : Solar Radiation	BG_4 : Solar Radiation	TS_4 : Solar Radiation
	LM_5 : Humidity	BG_5 : Humidity	TS_5 : Humidity
	LM_6 : Precipitation	BG_6 : Precipitation	TS_6 : Precipitation
	LM_7 : Air Pressure	BG_7 : Air Pressure	TS_7 : Air Pressure
	LM_8 : Wind Speed (2 meter)	BG_8 : Wind Speed (2 meter)	TS_8 : Wind Speed (2 meter)
	LM_9 : Root Soil Wetness	BG_9 : Root Soil Wetness	TS_9 : Root Soil Wetness
	LM_{10} : Surface Soil Wetness	BG_{10} : Surface Soil Wetness	TS_{10} : Surface Soil Wetness

[1] https://power.larc.nasa.gov/

To build the PCA-VARI model, three climate value locations were needed from 2001-2020 to compare the results of each location. Table 1 shows some variables that are the same and some that are different for each location [23]. It should be noted that the selection of variables for various indicators of each location is based on two measurements, namely theoretical basis and data availability [24]. As shown in Table 1, the climate indicators of mean UV Index, temperature, dewpoint, solar radiation, humidity, precipitation, air pressure, wind speed represent each location [25]. It can be noted that the temperature and wind speed is a measurement at the height of 2 meters. The data is obtained daily, then clean up some missing values with the interpolation technique. The data is then integrated into csv file format to facilitate grouping in one location. Finally, aggregating data is obtained to get monthly data that will be used to build the model.

3.2 Data Mining Processing using PCA-VARI

Principal Component Location Index for PCA-VARI

The initial data processing uses PCA to see the KMO value by testing the feasibility of the three climate data locations. The results obtained from this test are 0.6102, 0.5405, 0.7018 for the Lembang, Bogor, and Tasikmalaya locations, respectively. Furthermore, PCA modeling can be processed to meet the test criteria > 0.5.

Table 2. Estimation of Eigenvalues for Lembang, Bogor and Tasikmalaya Locations

Number	LM		BG		TS	
	Proportion	Eigenvalue	Proportion	Eigenvalue	Proportion	Eigenvalue
1	0.82381	74.9496	0.67352	47.0451	0.83252	82.2342
2	0.11610	10.5620	0.24064	1.68031	0.13281	13.1200
3	0.05704	5.18991	0.08236	5.75260	0.02988	2.95159
4	0.00199	0.18130	0.00191	0.13351	0.00366	0.36152
5	0.00073	0.06673	0.00120	0.08406	0.00071	0.07038
6	0.00028	0.02512	0.00033	0.02336	0.00026	0.02598
7	0.00006	0.00504	0.00007	0.00486	0.00007	0.00660
8	0.00003	0.00266	0.00004	0.00283	0.00005	0.00480
9	0.00000	0.00030	0.00001	0.00099	0.00000	0.00016
10	0.00000	0.00010	0.00000	0.00003	0.00000	0.00000

Table 2 shows that after forming the covariance at each LM, BG, and TS location data, the eigenvalues are obtained by its matrix. PC1 at the LM location gave a proportion of 82.38% and PC2 with a proportion of 11.61%. In contrast, the BG location for PC1 gives a proportion of 67.35% and PC2 with a proportion of 24.06%. The location of TS for PC1 has a proportion of 83.25% and PC2 with a proportion of 13.28%. Then the proportion of variance of all locations in the first principal component (PC1) exceeds 60%, while the proportion of variance in the second principal component (PC2) is between 10%-25%. PC1 and PC2 are sufficient to explain the overall data from the three climate measurement locations with two main components.

Table 3. Estimation of Eigenvector for The Principal Component of LM, BG, and TS Locations

Variable	LM		Variable	BG		Variable	TS	
	PC1	PC2		PC1	PC2		PC1	PC2
$\mathbf{LM_1}$	-0.00075	0.01037	$\mathbf{BG_1}$	-0.00085	0.00708	$\mathbf{TS_1}$	-0.00788	0.00592
$\mathbf{LM_2}$	-0.06091	0.11157	$\mathbf{BG_2}$	0.02185	0.13530	$\mathbf{TS_2}$	-0.09820	-0.04645
$\mathbf{LM_3}$	-0.10616	-0.02373	$\mathbf{BG_3}$	0.12298	-0.01765	$\mathbf{TS_3}$	-0.10358	-0.03498
$\mathbf{LM_4}$	-0.82420	0.52107	$\mathbf{BG_4}$	0.70716	0.66436	$\mathbf{TS_4}$	-0.78299	-0.60387
$\mathbf{LM_5}$	-0.21289	-0.59167	$\mathbf{BG_5}$	0.44426	-0.65164	$\mathbf{TS_5}$	-0.03228	0.04096
$\mathbf{LM_6}$	-0.51015	-0.60293	$\mathbf{BG_6}$	0.53527	-0.33661	$\mathbf{TS_6}$	-0.60398	0.79380
$\mathbf{LM_7}$	0.00422	0.00469	$\mathbf{BG_7}$	-0.00516	0.00550	$\mathbf{TS_7}$	0.00782	0.00452
$\mathbf{LM_8}$	0.00371	-0.03016	$\mathbf{BG_8}$	0.01548	-0.03877	$\mathbf{TS_8}$	0.02263	-0.00727
$\mathbf{LM_9}$	-0.00769	-0.02233	$\mathbf{BG_9}$	0.00874	-0.01602	$\mathbf{TS_9}$	-0.00679	0.00638
$\mathbf{LM_{10}}$	-0.00664	-0.01838	$\mathbf{BG_{10}}$	0.00948	-0.01530	$\mathbf{TS_{10}}$	-0.00551	0.00492

The PCA analysis for each climate measurement location in the West Java region is shown in Table 3. At the LM location, considering the calculated PC1, it can be concluded that the highest correlation for component with the first principal component (PC1) is solar radiation (LM_4) and the second component is precipitation (LM_6). Equation (19) shows a linear combination between the first principal components and other variables as follow:

$$\begin{aligned} PC1_{LM} = &-0.00075LM_1 - 0.06091LM_2 - 0.10616LM_3 - 0.82420LM_4 - 0.21289LM_5 \\ &- 0.51015LM_6 + 0.00422LM_7 + 0.00371LM_8 - 0.00769LM_9 - 0.00664LM_{10} \end{aligned} \quad (19)$$

Calculating the coefficients for the ten components of the BG location, the highest correlation of the first principal components is solar radiation (BG_4) followed by precipitation (BG_6). Equation (20) shows a linear combination and other variables

$$\begin{aligned} PC1_{BG} = &-0.00085BG_1 + 0.02185BG_2 + 0.12298BG_3 + 0.70716BG_4 + 0.44426BG_5 \\ &+ 0.53527BG_6 - 0.00516BG_7 + 0.01548BG_8 + 0.00874BG_9 + 0.00948BG_{10} \end{aligned} \quad (20)$$

Furthermore, the estimated coefficients of the first principal components are shown at the TS location. This location concludes that the highest correlation in the first principal component (PC1) is solar radiation (TS_4), followed by precipitation (TS_6). Equation (21) shows a linear combination

$$\begin{aligned} PC1_{TS} = &-0.00788TS_1 - 0.09820TS_2 - 0.10358TS_3 - 0.78299TS_4 - 0.03288TS_5 \\ &- 0.60398TS_6 + 0.00782TS_7 + 0.02263TS_8 - 0.00679TS_9 - 0.00551TS_{10} \end{aligned} \quad (21)$$

PCA-VARI for Three Locations

Augmented Dickey Fuller (ADF) and Box-Cox tests in equations (3) and (5) are used in this paper to test the locationary behaviour in the mean and variance of the $PC1_{LM}$, $PC1_{BG}$ and $PC1_{TS}$ locations respectively as $Z_{1,t}, Z_{2,t}, Z_{3,t}$. This test is applied to each indicator in 2 modes: i) non-stationary data at each differencing location; ii) differencing data for each location after processing the PCA results. It should be noted that the critical value of the ADF and Box-Cox tests varies according to the data assumptions, and all conditions are reported in Table 4. Thus, the validation test obtains an approved estimate. The results of the ADF $\alpha < 0.05$ and Box-Cox ≈ 1 test show that the time series climate has shown changes. Non-stationary to stationary through a one-time differencing process when the value meets the test criteria limits

Table 4. Result of Augmented Dickey-Fuller Unit Root and Box-Cox Tests

$Z_{1,t}$			$Z_{2,t}$			$Z_{3,t}$		
ADF	***p-value***	**Box-Cox**	**ADF**	***p-value***	**Box-Cox**	**ADF**	***p-value***	**Box-Cox**
-9.8104	0.0100	0.9995	-10.388	0.0100	0.8794	-9.8424	0.0109	0.8869

The optimum lag is obtained through the stationary variable in the PCA process by taking the first to fifth order difference. Then repeat the test of these variables at each location successively. Finally, the minimum fifth-order difference of the variables becomes the input of the VARI model after differencing-1. It should be noted that the fifth-order lag was chosen as the optimal VARI model lag using the Akaike Information Criterion (AIC). Due to the close difference between HQ(n) and SC(n) calculating the lag values, the three calculations have a minimum value at lag-5.

Table 5. Order-Lag Selection Criteria on The VARI Model

	VARI(1)	**VARI(2)**	**VARI(3)**	**VARI(4)**	**VARI(5)**
AIC(n)	6.975300	6.911906	6.715734	6.507818	6.435011*
HQ(n)	7.029224	7.019753	6.877505	6.723513	6.704629
SC(n)	7.109010	7.179325	7.116863	7.042656	7.103559

Applying MLE method, the parameters of selected PCA-VARI(5) model are estimated. The obtained are shown in Table 6. From this table, the second column, third column, and fourth column represent the estimated parameters of equation (12). Asterisks define the significance coefficient of the estimated model. As shown in Table 5 and Table 6, the order lag of the PCA-VARI model to be used is the fifth-order lag time and estimated parameters as we use it with predictions at the significance level of the model $\alpha < 0.05$.

Base on Table 6, $\dot{Z}_{1,t}$ represents the PCA-VARI model input of the first principal component of climate measurement at the Lembang location. The interpretation of the model is if there is a change in the values of solar radiation and precipitation (representation of two strong correlations) that occurred in the previous months at Lembang climate, causing an increase in values variable by 0.54044 times this month. Then if there is a change in the value of the previous four month at Bogor climate, it will cause an increase of 0.52312 times this month. If there is a change in the value previous three months at Tasikmalaya climate, it will cause an increase of 0.46896 times this month. $\dot{Z}_{2,t}$ represents the input process of the first principal component of the Bogor location. The interpretation of the model is that if there is a change in the values of solar radiation and precipitation (a strong correlation) that occurred in the previous 2 and 5 months at Lembang, it causes a decrease in the value of variables by 0.55685 and 0.46282 times this month. If there is a change in the value previous four month at Tasikmalaya, it will cause a decrease of 0.37717 times this month. $\dot{Z}_{3,t}$ represents the input process of the first principal component of the Tasikmalaya location. The interpretation model changes the solar radiation and precipitation values that occurred in the previous 3 and 4 months at Bogor climate, causing an increase in the value of variables by 0.57864, and 0.59738 times this month.

Table 6. The Estimated Coefficients of PCA-VARI(5) Model for Equation (13)

Variable	$\dot{Z}_{1,t}$	$\dot{Z}_{2,t}$	$\dot{Z}_{3,t}$
$\dot{Z}_{1,t-1}$	$\hat{\phi}^1_{1,1}$ = 0.54044 **	$\hat{\phi}^1_{2,1}$ = -1.04559	$\hat{\phi}^1_{3,1}$ = 1.14312
$\dot{Z}_{1,t-2}$	$\hat{\phi}^1_{1,2}$ = -0.03887	$\hat{\phi}^1_{2,2}$ = -0.55685 **	$\hat{\phi}^1_{3,2}$ = 0.35741
$\dot{Z}_{1,t-3}$	$\hat{\phi}^1_{1,3}$ = -0.38754	$\hat{\phi}^1_{2,3}$ = -0.34986	$\hat{\phi}^1_{3,3}$ = -0.15099
$\dot{Z}_{1,t-4}$	$\hat{\phi}^1_{1,4}$ = -0.42624	$\hat{\phi}^1_{2,4}$ = -0.18991	$\hat{\phi}^1_{3,4}$ = -0.09879
$\dot{Z}_{1,t-5}$	$\hat{\phi}^1_{1,5}$ = 0.22979	$\hat{\phi}^1_{2,5}$ = -0.46282 **	$\hat{\phi}^1_{3,5}$ = 0.42433
$\dot{Z}_{2,t-1}$	$\hat{\phi}^2_{1,1}$ = 0.40708 *	$\hat{\phi}^2_{2,1}$ = -1.14727	$\hat{\phi}^2_{3,1}$ = 0.44998 *
$\dot{Z}_{2,t-2}$	$\hat{\phi}^2_{1,2}$ = 0.29083	$\hat{\phi}^2_{2,2}$ = -0.87751	$\hat{\phi}^2_{3,2}$ = 0.38414
$\dot{Z}_{2,t-3}$	$\hat{\phi}^2_{1,3}$ = 0.42744 *	$\hat{\phi}^2_{2,3}$ = -0.84748	$\hat{\phi}^2_{3,3}$ = 0.57864 **
$\dot{Z}_{2,t-4}$	$\hat{\phi}^2_{1,4}$ = 0.52312 **	$\hat{\phi}^2_{2,4}$ = -0.75270	$\hat{\phi}^2_{3,4}$ = 0.59738 **
$\dot{Z}_{2,t-5}$	$\hat{\phi}^2_{1,5}$ = 0.55123 ***	$\hat{\phi}^2_{2,5}$ = -0.65733	$\hat{\phi}^2_{3,5}$ = 0.55015 ***
$\dot{Z}_{3,t-1}$	$\hat{\phi}^3_{1,1}$ = -0.16296	$\hat{\phi}^3_{2,1}$ = 0.14897	$\hat{\phi}^3_{3,1}$ = -0.70702
$\dot{Z}_{3,t-2}$	$\hat{\phi}^3_{1,2}$ = 0.28221	$\hat{\phi}^3_{2,2}$ = -0.11535	$\hat{\phi}^3_{3,2}$ = -0.02981
$\dot{Z}_{3,t-3}$	$\hat{\phi}^3_{1,3}$ = 0.46896 **	$\hat{\phi}^3_{2,3}$ = -0.17841	$\hat{\phi}^3_{3,3}$ = 0.32512
$\dot{Z}_{3,t-4}$	$\hat{\phi}^3_{1,4}$ = 0.67272 ***	$\hat{\phi}^3_{2,4}$ = -0.37717 **	$\hat{\phi}^3_{3,4}$ = 0.39469 *
$\dot{Z}_{3,t-5}$	$\hat{\phi}^3_{1,5}$ = 0.14986	$\hat{\phi}^3_{2,5}$ = -0.04504	$\hat{\phi}^3_{3,5}$ = -0.04076

Note: The asterisks ***,**, and * represent the statistical significance at 0.01, 0.05 and 0.1, respectively

The interpretation of this model shows that the principal component variables influence each other at each climate location. Indicating the VARI process, the performance obtained at the three locations shows a strong correlation between the two variables that affect each location, namely solar radiation and precipitation, representing the principal component. Different sample locations for data analysis at the three locations show that this modeling in the West Java region tends to be influenced by those variables that affect each other and have a strong relationship with predictions at the significance level of the model $\alpha < 0.05$.

Table 7. Diagnostic Test The Coefficient of Estimation of PCA-VARI(5) Model

Diagnostic Test	Variable	F-Test	*p-value*
Granger	$Z_{1,t}$	4.2841	0.00000829
	$Z_{2,t}$	1.4027	0.1746
	$Z_{3,t}$	2.4569	0.006922
Ljung-Box	$Z_{1,t}$		0.9800
	$Z_{2,t}$		0.9789
	$Z_{3,t}$		0.9956

Verifying the model obtained to meet the significance of multivariate modeling is done by performing a diagnostic test to see the frequency distribution of the residuals with the Normality and white noise test with the Ljung-Box method. In Table 7, the diagnostic test results for the PCA-VARI(5) model are presented.

In Table 7, it can be seen that in the Granger causality test, the $Z_{1,t}$ Granger causality for $Z_{2,t}$ and $Z_{3,t}$, the significant test with *p-value* = 0.00000829 (<0.05) and a null hypothesis is rejected and can be concluded that it meets the model diagnostic test. Then $Z_{2,t}$ not Granger for $Z_{1,t}$ and $Z_{3,t}$ caused *p-value* = 0.1746. Granger causality $Z_{3,t}$ for $Z_{1,t}$ and $Z_{2,t}$ the significant test with p-value = 0.006922 (<0.05) and a null hypothesis is rejected and can be concluded that it meets the model diagnostic test. Then the white noise test with Ljung-Box gives *p-values* close to 1. Therefore, the obtained PCA-VARI(5) model fulfils the multivariate diagnostic test.

3.3 Postprocessing

To analyze the effect of climate measurement variables between locations (Lembang, Bogor, Tasikmalaya) by taking the first principal component at each location of the climate measurement location from the PCA process, then modeling it with VARI. $Z_{1,t}, Z_{2,t}, Z_{3,t}$ represent of each principal component of Lembang (PCLM1), Bogor (PCBG1), and Tasikmalaya (PCTS1) according to (19), (20), and (21), respectively. The PCA-VARI(5) model is obtained as in Table 6 with the significance level of the model $\alpha < 0.05$. The relationship between variables and location is obtained by interpretation using IRF for each location that forecasts for the next ten months and affects each other.

Figure 2. (PCBG1 to PCBG1) shows that IRF, if the climate value index at the Bogor location (PCBG1) changes to a standard deviation, it will have an increasing effect 5.1231 on its area in the first month. In other words, solar radiation and precipitation in the current month will decrease until the second month and increase 0.2494 to the third month and decrease to the sixth month.

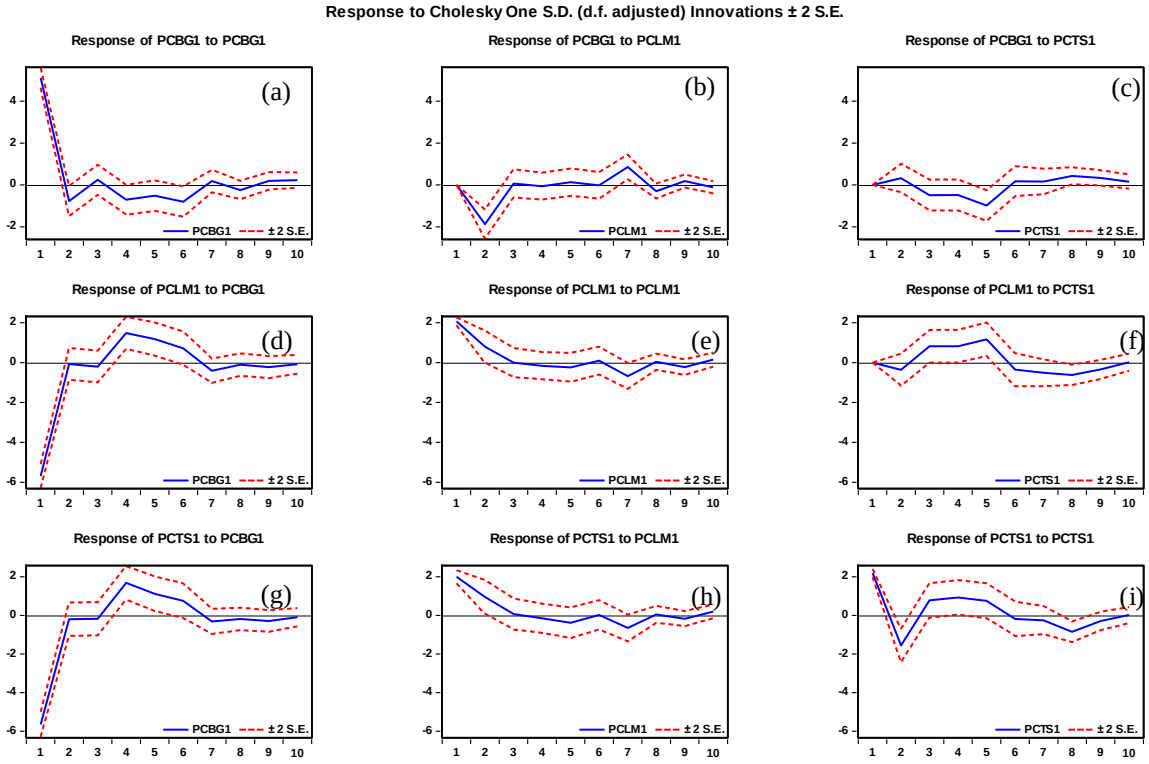

Figure 2. Impulse Response Function (IRF) from the Cholesky Decomposition for PCA-VARI(5) Model

In Figure 2, we use the scale in the X-axis indicates the lag in month and forecasting values and we have 9 plots of IRF. Figure 2(a) shows a negative shock to the standard deviation is introduced to the PCLM1, there is no effect PCBG1 on PCLM1 like Figure 2(b) in the first month but a decrease of 1.8701 in the second month, there is an increase of 0.07183 to the third month and zero effect until the sixth month. This means that solar radiation and precipitation are the response of the Bogor location to Lembang. On the other hand, Figure 2(d) has a similar effect as Figure 2(g). Starting with a negative response in the first month, then showing an increase until the sixth month, and decreasing again to the tenth month.

For Figure 2(f), it shows a not so significant effect on the standard deviation of the first month but decrease 0.3588 in the second month, the increase fluctuation of response the PCTS1 until the sixth month and moves close to below zero point to the tenth month. This indicates the response of Figure 2(f) has a small standard deviation. The response PCTS1 to PCLM1 in Figure 2(h) beginning in the first month with an increase of 1.9996 and approached zero effect in the third month and continued to decrease in the fifth and seventh months, then increased by 0.2021 at the tenth month.

In Figure 2(e) which is Lembang response to itself, it shows similar to the response like Figure 2(h) beginning in the first month of 2.0735 approaching zero effect in the third month. The fluctuating effect occurred until the ninth month with a value below zero which means that the influence of solar radiation and precipitation did not increase in Lembang climate between these months and only increased by 0.1491 in the tenth month.

Finally, based on the results of the impulse-response function, the climate of Bogor strongly are responded by Lembang and Tasikmalaya climate. Then the response of the Lembang climate variable is significant affected by Tasikmalaya climate and Bogor climate. Furthermore, Tasikmalaya response affects Lembang and also Bogor. Therefore, Bogor climate response to Lembang and otherwise has a similar response and influences each other, but the shock of Bogor and Lembang to Tasikmalaya is of minimum significance, although Lembang increase response is more volatile in the third to sixth months. This is due to the relatively high Solar Radiation and Precipitation in the Bogor and Lembang areas, Through PCA-VARI modeling, we hope that it can help climate forecasting in the West Java region, especially in the agricultural sector to see the flexibility of the planting season as a recommendation for the next few months.

4. CONCLUSIONS

Empirical analysis of climate data in the West Java region has been very important lately. This study includes an analysis of climate parameters such as UV Index, temperature, dewpoint, solar radiation, humidity, precipitation, air pressure, wind speed, root soil wetness, surface soil wetness located in the three areas of Lembang, Bogor, and Tasikmalaya from January 2001 to December 2020 from POWER NASA Agroclimatology. We were conducted using an integrated PCA model with VARI. The result of the PCA process, which reduces the variables for each location, is the input for VARI modeling. At the Lembang climate location, solar radiation and precipitation variables are strongly correlated and influence each other. Climate measurements at the Bogor location show similar variables as Lembang and influence each other. While the climate variables at Tasikmalaya location have a strong correlation and influence, each other but minimum responded by Lembang and Bogor. The three locations represented the analysis of the PCA-VARI model in the West Java region obtain the modeling predictions involving variables correlations of solar radiation and precipitation were strongly correlated and influenced each other for the next ten months. While the influence of inter-location variables, using IRF shows that the climate variables of the Lembang and Bogor regions influence each other, but the minimum response to Tasikmalaya has a standard deviation close to zero. It can be concluded that the variables of solar radiation and precipitation have impacted the climate of West Java, which influence dominant each other in the Lembang and Bogor regions.

ACKNOWLEDGEMENT

The author thanks the Rector of Universitas Padjadjaran, who has provided financial support for dissemination lecturer and student research through the Academic Leadership Grant with contract number: 1959/UN6.3.1/PT.00/2021. The author also thanks to the reviewers, Prof. Dr. Ir. Eddy Hermawan, M.Sc., and Dr. Diah Chaerani, M.Si., who have provided input to improve this paper. Also, the author gratefully thanks the Head of the National Research and Innovation Agency who has supported the funding for Magister Program by Research 2020.

REFERENCES

[1] R. Kishimoto, T. Shimura, N. Mori, and H. Mase, "Statistical modeling of global mean wave height considering principal component analysis of sea level pressures and its application to future wave height projection," *Hydrol. Res. Lett.*, vol. 11, no. 1, pp. 51–57, 2017, doi: 10.3178/hrl.11.51.

[2] B. J. Washington and L. Seymour, "An adapted vector autoregressive expectation maximization imputation algorithm for climate data networks," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 12, no. 6, 2020, doi: 10.1002/wics.1494.

[3] S. Ankamah, K. S. Nokoe, and W. A. Iddrisu, "Modelling Trends of Climatic Variability and Malaria in Ghana Using Vector Autoregression," *Malar. Res. Treat.*, vol. 2018, 2018, doi: 10.1155/2018/6124321.

[4] F. Pretis, "Econometric modelling of climate systems: The equivalence of energy balance models and cointegrated vector autoregressions," *J. Econom.*, vol. 214, no. 1, pp. 256–273, 2020, doi: 10.1016/j.jeconom.2019.05.013.

[5] S. Mamipour, M. Yahoo, and S. Jalalvandi, "An empirical analysis of the relationship between the environment, economy, and society: Results of a PCA-VAR model for Iran," *Ecol. Indic.*, vol. 102, pp. 760–769, 2019, doi: 10.1016/j.ecolind.2019.03.039.

[6] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed. Massachusetts: Elsevier Inc., 2012.

[7] J. Niu *et al.*, "A comparative study on application of data mining technique in human shape clustering: Principal component analysis vs. factor analysis," in *Proceedings of the 2010 5th IEEE Conference on Industrial Electronics and Applications, ICIEA 2010*, 2010, pp. 2014–2018, doi: 10.1109/ICIEA.2010.5515577.

[8] W. L. Cerón, J. Molina-Carpio, I. Ayes Rivera, R. V Andreoli, M. T. Kayano, and T. Canchala, "A principal component analysis approach to assess CHIRPS precipitation dataset for the study of climate variability of the La Plata Basin, Southern South America," *Nat. Hazards*, vol. 103, no. 1, pp. 767–783, 2020, doi: 10.1007/s11069-020-04011-x.

[9] Snedecor, G. W., and W. G. Cochran, *Statistical Methods*. Iowa State University Press, 1989.

[10] T. Singh, A. Ghosh, and N. Khandelwal, "Dimensional reduction and feature selection: Principal component analysis for data mining," *Radiology*, vol. 285, no. 3, p. 1055, 2017, doi: 10.1148/radiol.2017171604.

[11] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Six editio. New Jersey: Pearson Prentice Hall, 2007.

[12] H. Anton and C. Rorrers, *Elementary Linear Algebra*, 11th ed. Wiley, 2014.

[13] G. E. P. Box and G. M. Jenkins, *Time Series Analysis Forecasting and Control*. Holden-Day. Inc, 1976.

[14] D. A. Dickey and W. A. Fuller, "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *J. Am. Stat. Assoc.*, vol. 74, no. 366, pp. 427–431, 1979, doi: 10.1080/01621459.1979.10482531.

[15] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, 2nd ed. New York: Springer-Verlag, 2002.

[16] G. E. P. Box and D. Cox, "An analysis of transformations," *J. R. Stat. Soc.*, vol. B.26, no. 2, pp. 211–252, 1964.

[17] A. Hoyyi, Tarno, D. A. I Maruddani, and R. Rahmawati, "Vector autoregressive model approach for forecasting outflow cash in Central Java," 2018, vol. 1025, no. 1, doi: 10.1088/1742-6596/1025/1/012105.

[18] Y. Nalita, R. Rahani, E. R. Tirayo, T. Toharudin, and B. N. Ruchjana, "Ordinary least square and maximum likelihood estimation of VAR(1) model's parameters and it's application on covid-19 in China 2020," 2021, doi: 10.1088/1742-6596/1722/1/011002.
[19] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969, doi: 10.2307/1912791.
[20] E. Chandra and P. Ajitha, "PCA for heterogeneous data sets in a distributed data mining," 2011, doi: 10.1145/1980422.1980451.
[21] Y. Yu and D. Wang, "Similarity Study of Hydrological Time Series Based on Data Mining," *Adv. Intell. Syst. Comput.*, vol. 1303, pp. 1049–1055, 2021, doi: 10.1007/978-981-33-4572-0_150.
[22] X. Du and F. Zhu, "A novel principal components analysis (PCA) method for energy absorbing structural design enhanced by data mining," *Adv. Eng. Softw.*, vol. 127, pp. 17–27, 2019, doi: 10.1016/j.advengsoft.2018.10.005.
[23] J. C. L. Chan and J.-E. Shi, "Application of projection-pursuit principal component analysis method to climate studies," *Int. J. Climatol.*, vol. 17, no. 1, pp. 103–113, 1997.
[24] S. M. Shaharudin, N. Ahmad, N. H. Zainuddin, and N. S. Mohamed, "Identification of rainfall patterns on hydrological simulation using robust principal component analysis," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 11, no. 3, pp. 1162–1167, 2018, doi: 10.11591/ijeecs.v11.i3.pp1162-1167.
[25] M. A. Shahin, M. A. Ali, and A. B. M. S. Ali, *Vector Autoregression (VAR) modeling and forecasting of temperature, humidity, and cloud coverage*, vol. 9789401786. 2014.