BAREKENG: Journal of Mathematics and Its Applications
June 2024 Volume 18 Issue 2 Page 1273–1282
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss2pp1273-1282

CANONICAL CORRELATION ANALYSIS OF ECONOMIC GROWTH AND UNEMPLOYMENT RATE

Joko Purwadi [1*], Bagus Gumelar[2], Tri Widiantoro[3], Zhilvia Noviana Ningsih[4]

[1]Mathematics Study Program, Faculty of Applied Science and Technology Universitas Ahmad Dahlan
Jln. Ringroad Selatan, Yogyakarta, 55191, Indonesia

[2]Management Study Program, Faculty of Economics and Business, Universitas Ahmad Dahlan
Jln. Kapas No 9, Yogyakarta, 55166, Indonesia

[3]BPS-Statistics Indonesia
Jln. T. Bedussamad no 201, Aceh Tenggara, 24651, Indonesia

[4]Department of Statistics, Faculty of Science and Data Analitics, Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo,Surabaya, 60111, Indonesia

*Corresponding author's e-mail: * **joko@math.uad.ac.id***

ABSTRACT

Article History:

Received: 13th January 2024
Revised: 24th February 2024
Accepted: 2nd May 2024
Published: 1st June 2024

Keywords:

Canonical Correlation;
Economic growth;
Multivariate Analysis;
Unemployment Rate.

The paper discusses the relationship between economic growth and the unemployment rate in Indonesia in each province in 2021. Both variables are considered as the dependent variables, and there are 5 independent variables used in this research, such as human index development, wage minimum region, poor citizens percentage, investment, and farmer rate value in each province. The method used to analyze is canonical correlation analysis, which is one of the dependent methods that are used for multivariate analysis. This method was used to determine which variable had the most significant relationship between dependent and independent variables. The data was taken from the Center of Statistics Bureau Indonesia in 2021. The result shows that among independent variables, the human index development had the strongest relation at 79%, while the correlation between the dependent and independent variables, the unemployment rate gives the strongest influence at 68%.

How to cite this article:

J. Purwadi, B. Gumelar, T. Widiantoro and Z. N. Ningsih., "CANONICAL CORRELATION ANALYSIS OF ECONOMIC GROWTH AND UNEMPLOYMENT RATE," *BAREKENG: J. Math. & App.*, vol. 18, iss. 2, pp. 1273-1282, June, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*
Research Article · Open Access

1. INTRODUCTION

There are three main components that affect economic growth; the first is capital accumulation, obtained from savings and investments set aside from current income to increase production and income in the future. The second is population growth and workforce, traditionally population growth is considered a positive factor in encouraging economic growth, and the third is technological progress, which is a new way and improvement of operating, where there are three main groups of technological progress, namely neutral, labor saving, and capital saving [1].

Economic growth and unemployment rate are essential problems, especially during the post-pandemic COVID-19. Every province in Indonesia tries its best to restore it to make it more stable, controllable, and even better. The low per capita investment level is caused by the low per capita domestic demand as well and this happens because of the high level of poverty and so on, thus forming a circle of poverty as a cause-and-effect relationship [2]. Research on economic growth investment was done in 2022 by Cili; the result shows that inflation give a significant impact [3].

Canonical correlation analysis (CCA) is one of the interdependence techniques in multivariate statistics, which deals with analysis for the discovery and quantification of associations between two sets of variables [4] [5] [6] [7]. CCA aims to maximize the association (measured by correlation) between the low-dimensional projections of the two data sets. Research has been done by [8]. They use CCA to analyze the relationship between egg production traits and body weight, egg weight, and age at sexual maturity in layers, and the result gives a nice interpretation. In the year 2021, the overview of the CCA is used in multi-view classification, and the idea is to map data from different views onto a common space with maximum correlation [9].

In other research in CCA under a mild condition, which tends to hold for high-dimensional data, CCA in the multilabel case can be formulated as a least-squares problem [10]. CCA model with extension is considered, with three or more sets of variables [11]. The method is also extended to handle non-linear relations via kernel trick (this increases the complexity to quadratic complexity). The scalability is demonstrated on a large-scale cross-lingual information retrieval task [12]. The stochastic algorithm converges to the stationarity equations for the determination of the canonical variables and the canonical correlations in 1998 using the neural network algorithm [13]. General method on CCA using kernel function already done, they study about to learn a semantic representation of web images and their associated text [14]. The sparse CCA to an important genome-wide association study problem, mapping was done in 2012 by Chen et al., the empirical results in their research show that the proposed optimization algorithm is more efficient than existing state-of-the-art methods [15].

In determining the economic growth and the unemployment rate, there are several factors to be considered, such as human index development, wage minimum region, poor citizens percentage, investment, and farmer rate value in each province. The purpose of this research is to determine the correlation between each factor and how significant those factors are using the CCA methods.

2. RESEARCH METHODS

2.1 Canonical Correlation Analysis

Canonical correlation analysis focuses on the correlations between linear combinations of sets of the dependent variables with a linear combination of the set of variable independent. The idea of this analysis is to determine the pair of this linear combination has the greatest correlation. Then, look for pairs of linear combinations among pairs that are uncorrelated in the pair of sections at the start of the selection. The pairs of these linear combinations are called canonical functions, and the correlations are called Canonical Correlations [5].

2.2 Determination of Canonical Correlation Coefficient Estimators and Canonical Functions

Suppose they want to measure a linear relationship between a set of dependent variables which is denoted by a random vector y, with a set of independent variables $x_1, x_2, \ldots, x_p$ which is denoted by a random vector x, Where $p \leq q$. For each sample on n vectors observation, then the average vector and its covariance matrix:

$$\begin{bmatrix} \overline{y} \\ \cdots \\ \overline{x} \end{bmatrix} = \begin{bmatrix} \overline{y_1} \\ \overline{y_2} \\ \vdots \\ \overline{y_p} \\ \cdots \\ \overline{x_1} \\ \overline{x_2} \\ \vdots \\ \overline{x_p} \end{bmatrix} ; S = \begin{bmatrix} S_{yy} & S_{yx} \\ S_{xy} & S_{xx} \end{bmatrix}.$$

The linear combination of the two sets of variables can be written as:

$$U = a^T y = a_1^T y + a_2^T y + \cdots + a_k^T y$$

$$V = b^T x = b_1^T x + b_2^T x + \cdots + b_k^T x$$

For:

$$var(U) = a^T \operatorname{cov}(y)a = a^T s_{yy} a$$

$$\operatorname{var}(V) = b^T \operatorname{cov}(x)a = b^T s_{xx} b$$

$$\operatorname{cov}(U,V) = a^T \operatorname{cov}(x,y)b = a^T S_{yx} b.$$

So, the canonical correlation:

$$r_{c(U,V)} = \frac{\operatorname{cov}(U,V)}{\sqrt{\operatorname{var}(U)\operatorname{var}(V)}} = \frac{a^T S_{yx} b}{\sqrt{a^T S_{yy} a}\sqrt{b^T S_{yy} b}}.$$

Eigenvalues can be obtained from the characteristic equation:

$$\left| S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} - \lambda \mathrm{I} \right| = 0$$

$$\left| S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} - \lambda \mathrm{I} \right| = 0$$

Vector coefficient a_k and b_k obtained in the canonical function $U_k = a_k^T y$ and $V_k = b_k^T x$ is a vector eigen of the same two matrices:

$$\left| S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} - \lambda \mathrm{I} \right| a = 0$$

$$\left| S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} - \lambda \mathrm{I} \right| b = 0$$

up to two matrices $S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy}$ and $S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx}$ have non-zero eigenvalues and are different.

Eigenvectors (2)

For the *k*-th canonical function pair:

$$\begin{array}{ccc} U_1 = a_1^T y & & V_1 = b_1^T x \\ \vdots & \text{and} & \vdots \\ U_k = a_k^T y & & V_k = b_k^T x \end{array}$$

where y and x are the values of the set of dependent and independent variables for the unit special observation [5].

2.3 Canonical Correlation Assumptions

2.3.1 Linearity

Linearity, namely the relationship between the set of independent variables x and variables the dependent y is linear. Linearity can be said to be important for canonical correlation analysis and affect two aspects of canonical correlation results. First, the canonical correlation coefficient between a pair of canonical variables is based on a linear relationship. If the variables is not linear, then the relationship will not be explained by the canonical correlation coefficient. Second, canonical correlation analysis maximizes the linear relationship between sets of variables [5].

2.3.2 Independent and Dependent Variables with Multivariate Normal Distribution

There are two ways to check the multivariate normal assumptions. Check the assumption of normality by making a Chi Square plot (for $p \geq 2$). The steps are as follows:

a. First, calculate value $d_j^2 = \left(x_j - \bar{x}\right)^T S^{-1}\left(x_j - \bar{x}\right), j = 1, 2, \cdots, n$ and then Sort d_j^2 according to ascending order $d_1^2 \leq d_2^2 \leq \cdots \leq d_n^2$, Couple Plots, $q_{c,p}\left(\left(j - \frac{1}{2}\right)/n, d_j^2\right)$, with $q_{c,p}\left(j - \frac{1}{2}\right)/n$ is $100/\left(j - \frac{1}{2}\right)/n$ quantiles of the Chi square distribution with degrees of freedom p, if the results of the plot are linear, then it can be assumed to be multivariate normal distribution.

b. Then the second is to look at the number of values d_j^2 which is less than the quantile value Chi square. The first thing to do is calculate the value d_j^2, $j = 1, 2, \cdots, n$ and then compare it to the quantile value χ^2. If there are half or more values $d_j^2 \leq q_{c,p}\left(0,50\right)$, then it can be said that the data is normally distributed multivariate [5].

2.3.3 NonMulticollinearity

Multicollinearity relates to situations where there is a definite linear relationship or close to certain among the independent variables. Multicollinearity occurs when several independent variables have a high correlation with other independent variables [16]. The multicollinearity is determined by the Variation of Inflation Factor (VIF); the VIF formula is as follows:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2},$$

where R_i^2 is the coefficient of determinant. If the VIF result exceed 10 indicate that there is multicollinearity.

2.4 Canonical Correlation Significance Test

There are two hypotheses to be tested in the canonical correlation analysis, namely, the correlation test canonical as a whole and the test in part.

2.4.1 Overall Canonical Correlation Test

The hypothesis for overall used in this paper will decide it is significant or insignificant is the canonical correlation. The hypotheses are as follow:

Hypothesis:

$H_0 : r_{c1} = r_{c2} = \cdots = r_{ck} = 0$ (all canonical correlations are not significant)

$H_1 : r_{ci} \neq 0$ (at least one significant canonical correlation, with $i = 1, 2, \cdots, k$)

Test Statistics: $F = \dfrac{1 - \Lambda_1^{1/t} df_2}{\Lambda_1^{1/t} df_1}$

with:

$$\Lambda_1 = \Pi_{i=1}^{k}\left(1-r_1^2\right); df_1 = pq; df_2 = wt - \frac{1}{2}pq + 1$$

$$w = n - \frac{1}{2}\left(p+q+3\right); t = \sqrt{\frac{p^2q^2-4}{p^2q^2-5}}$$

where:

n = number of observations

p = number of sets of variables y

q = number of sets of variables x

rejection area: H_0 rejected if $\text{F} > \text{F}_{\alpha;df_1;df_2}$ or $\Lambda_1 \leq \Lambda_{\alpha;p,q,n-1-q}$.

2.4.2 Partial Test

There are several tests can be used in CCA, after the overall test done, one of them is the partial test for CCA Hypothesis as follow:

$H_0 : r_{cj} = 0$ (non-significant canonical correlation)

$H_1 : r_{cj} \neq 0$ (at least one significant canonical correlation)

Test Statistics: $F = \frac{1-\Lambda_1^{1/t} df_2}{\Lambda_1^{1/t} df_1}$

with:

$$\Lambda_j = \Pi_{i=1}^{k}\left(1-r_1^2\right); df_1 = \left(p-j+1\right)\left(q-j+1\right)$$

$$df_2 = wt - \frac{1}{2}\left[\left(p-j+1\right)\left(q-j+1\right)\right]$$

$$w = n - \frac{1}{2}\left(p+q+3\right); t = \sqrt{\frac{\left(p-j+1\right)^2\left(q-j+1\right)^2-4}{\left(p-j+1\right)^2\left(q-j+1\right)^2-5}}$$

where:

n = number of observations

p = number of sets of variables y

q = number of sets of variables x

rejection area: H_0 rejected if $\text{F} > \text{F}_{\alpha;df_1;df_2}$ or $\Lambda_1 \leq \Lambda_{\alpha;p-j+1,q-j+1,n-j-q}$

2.5 Interpretation of Canonical Functions

2.5.1 Canonical Weight

Canonical weights, which are standardized canonical coefficients, can be interpreted as the magnitude of the closeness of the original variable to the canonical variable. The greater the coefficient value this states the higher the level of closeness of the variable concerned to the variable canonical and conversely the smaller the canonical weight value, the lower the level of closeness variable. Canonical weights are unstable due to multicollinearity in optimizing the results of canonical correlation calculations. It is more appropriate to use canonical payloads and canonical cross-loads to interpret the results of the canonical correlation analysis [6].

2.5.2 Canonical Load

Canonical loadings have been widely used for interpretation because of the lack of the nature of canonical weight. The canonical load can be called the correlation of the canonical structure, the canonical load is a simple linear correlation between the original variables and each variable it's canonical, describes the diversity of shared variables observed with the canonical variables, and can be interpreted like a factor loading in assessing the relative contribution of each variable to its canonical function [5].

2.5.3 Canonical Crossload

Canonical cross-load was suggested as an alternative to canonical load. Canonical cross-load provides a more precise measure of the relationship of dependent and independent variables, which can be calculated from the multiplication of the canonical correlation value with the payload value canonical. This calculation includes the correlation of each set of dependent variables with variables canonical of the set of independent variables and vice versa. The greater the cross-load canonical reflects the closer the relationship of canonical variables [6].

3. RESULTS AND DISCUSSION

This paper uses data from the Center of Statistics Bureau Indonesia in 2021. The data is the compilation of multiple data collected from https://www.bps.go.id/ and jointly together as a dataset named dataproject.xslx. The dataset contains variable dependent y_1 and y_2 respectively for economics growth and unemployment rate, whereas the variable independent x_1 (human index development), x_2 (wage minimum region), x_3 (poor citizens percentage), x_4 (investment) and x_5 (farmer rate). The first thing to do with the dataset is check the normality assumption, by using the plot normal Q-Q plot which can be seen in Figure 1:

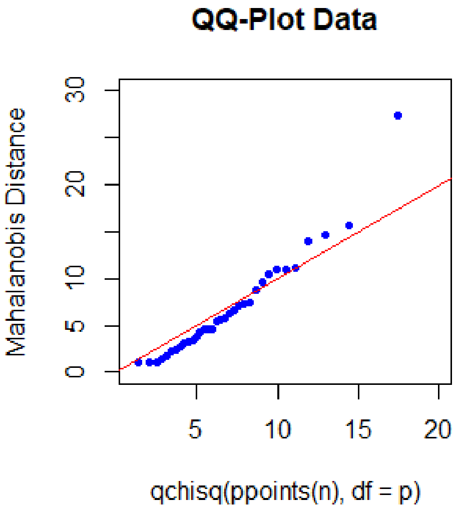

Figure 1. Plot Normality Data

From Figure 1, it can be concluded that the data spread normally because the data is spreading between the lines. The second assumption is multicollinearity, the test uses the VIF test to determine if there is any multicollinearity or not. The multicollinearity result can be seen in Table 1 as follows.

Table 1. VIF Score

x_1	x_2	x_3	x_4	x_5
1.296574	1.516801	1.274773	1.523834	1.443105

From Table 1, it can be concluded that there is no multicollinearity among the independent variables as the result shows that it is less than 10. However, the visualization correlation by using the R software can be shown in the figure below.

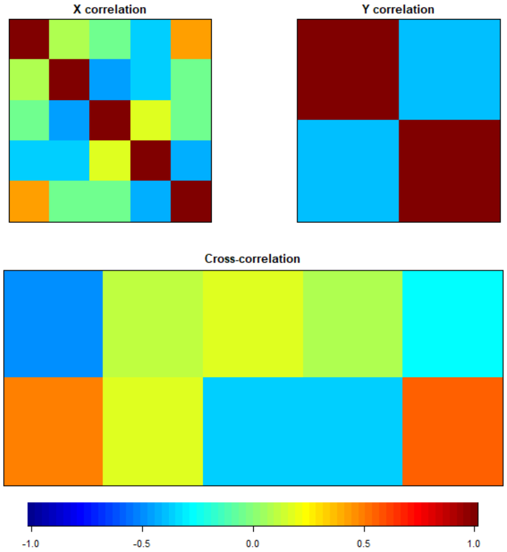

Figure 2. Coefficient Correlation Visualization

From Figure 2, it can be concluded that the correlation between the independent and the dependent variable each shows different relation. After the assumption checking on normality and multicollinearity, the next step is to determine the canonical correlation and the result is as follows.

Table 2. Canonical Correlation Analysis

	CanR	CanRSQ	Eigen	percent	cum	Pr(> F)
1st	0.7309	0.5343	1.1471	88.32	88.32	0.003536
2nd	0.3629	0.1317	0.1517	11.68	100.00	0.393692

1st : first canonical Function ; 2nd: first canonical Function

Based on Table 2, the Canonical correlation analysis of dependent variables y for y_1 (economic growth), y_2 (unemployment rate), and the independent variables x for x_1 (human index development), x_2 (wage minimum region), x_3 (poor citizens percentage), x_4 (investment) and x_5 (farmer rate). The first correlation between canonic pairs is 0.7309 and the squared canonical correlation is 0.5343, which means that the highest canonical correlation might happen between the combination linear from the dependent variables (y_1 and y_2) and some of the combination linear from the dependent variables (x_1, x_2, x_3, x_4, x_5).

The contribution of variation can be explained by the first canonical function as big as 88.32%, and the second canonical function gives 11.68%. Based on the proportion of both variation contributions, it is sufficient to use the first canonical function with 88% variation to explain the canonical correlation. From Table 2, the p-value 0.003536 (< 0.05) from the first canonical function is significant which means that the first canonical correlation can be used to describe the correlation between the dependent variable and the dependent variable. The graph for the best canonical correlation that can be seen from the figure as follows.

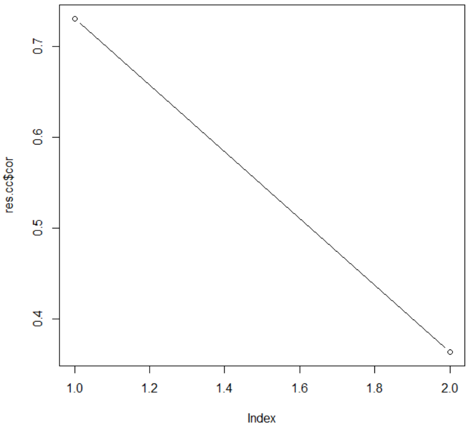

Figure 3. Percentage variation explained

Based on Figure 3, there is only one canonical function that can be used, that is the first canonical function. The next step is to determine the canonic coefficient from the canonical function, the result is seen in Table 3 as follows.

Table 3. Canonic Coefficient for Independent Variable x

	[,1]	[,2]
x_1	-1.466253e-01	1.901077e-01
x_2	1.027571e-01	-1.289701e+00
x_3	1.415788e-01	-4.526317e-02
x_4	-1.494413e-03	1.056959e-02
x_5	-3.041503e-05	-3.847616e-05

From Table 2, it can be seen that the variable that gives the highest contribution is the x_5 (farmer rate), followed by x_4 (Investment), x_1 (human index development), x_2 (poor citizen percentages) and x_3 (wage minimum region).

Table 4. Canonic Coefficient for Independent Variable y

	[,1]	[,2]
y_1	0.1569006	-0.4191974
y_2	-0.4497501	-0.4150017

From Table 4, the dependent variable that gives more contribution is y_1 (unemployment rate) and followed by y_2 (economic growth).

For further detailed correlation between the independent variables (y) and independent variables (x), the correlation result is as follows.

Table 5. Coefficient Correlation between y and x

	[,1]	[,2]
y_1	0.4956755	-0.2667336
y_2	-0.6845489	-0.1272241

Table 6. Coefficient Correlation between x and y

	[,1]	[,2]
x_1	0.5807001	-0.1270811
x_2	-0.1126921	-0.2324510
x_3	0.3402458	0.0755985
x_4	0.3124059	0.1641024
x_5	-0.5400955	-0.1208970

Table 7. Coefficient Correlation between x and x

	[,1]	[,2]
x_1	-0.7944700	0.3501449
x_2	-0.1541768	-0.6404691
x_3	0.4654987	0.2082955
x_4	0.4274101	0.4521491
x_5	-0.7389179	-0.3331059

Table 8. Coefficient Correlation between y and y

	[,1]	[,2]
y_1	0.6781458	-0.73449274
y_2	-0.9365482	-0.350538

From Table 8, it can be concluded that the dependent variable y in the first canonical function, which had a close relation is y_2 which is the unemployment rate at 93%. The independent variable x which has a close relation is x_1, the human index development at 79%, and x_5 (farmer rate) at 73%. The cross correlation between variables x and y can be seen in Table 5 gives the y_2 (Unemployment rate) at 68%, and Table 6 gives x_1 (human index development) at 58% as the variable that had the strong relation.

For further discussion, the research wishes to know if the province had the same canonical correlation among the dependent and independent variables, then the canonical correlation figure can be represented as follows.

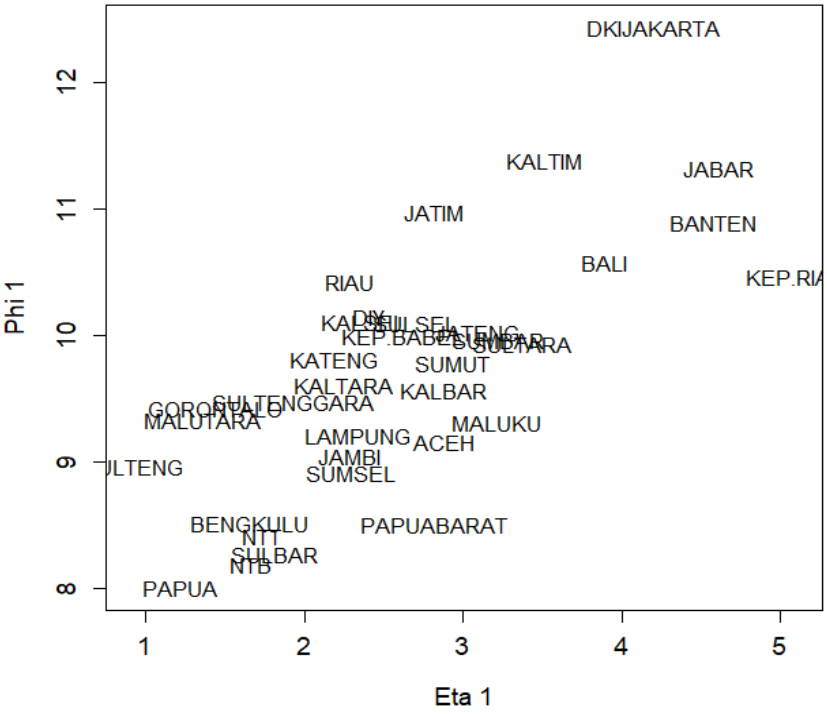

Figure 4. Canonical Correlation by Province

From **Figure 4**, the canonical correlation between provinces shows that DKI Jakarta had the weakest canonical correlation based on the variable used. It had the farthest distance among other provinces, and the other province that had the closest distance among others had a similarity to the canonical correlation between variable dependent y and variable independent x.

4. CONCLUSIONS

The result in the analysis of economic growth and employment rate using the canonical correlation shows that among independent variables, the human index development had the strongest relation at 73%, while the correlation between the dependent and independent variables the unemployment rate, gives the strongest influence it is 68%. and only the first canonical function that can be used to describe the variation percentage which can describe 88% variation. The canonical visualization shows that only one province had a different correlation with the other province. Almost all provinces had the same character as they had a close distance from each other.

ACKNOWLEDGMENT

Special thanks to BRIN Universitas Ahmad Dahlan for the financial support with contract number: 0-105/SP3/LPPM-UAD/XII/2023 and our sincere gratitude to Dedy Dwi Prastyo and Santi Putri Rahayu ITS for the knowledge and guidance so the paper finished.

REFERENCES

[1] P. T. Michael and S. C. Smith, Pembangunan Ekonomi, edisi 9, Jakarta: Airlangga, 2000.

[2] R. Nurkse and Hutagalung, Masalah pembentukan modal di negara-negara jang sedang membangun,, Djakarta: Bharata, 1964.

[3] M. R. Cili and Alkhaliq, "Economic Growth and Inflation: Evidence from Indonesia," *Signifikan: Jurnal Ilmu Ekonomi,* vol. 11, no. 1, pp. 145-160, 2022.

[4] W. K. Hardle and L. Simar, Applied Multivariate Statistical Analysis, New York: Springer, 2014.

[5] R. A. Johnson and W. W. D, Applied Multivariate Statistical Analysis 4th ed, New York: Prentice Hall, 1998.

[6] J. F. Hair, W. C. J. Black, B. J. Babin and E. A. R, Multivariate Data Analysis 7th ed, New York: Prentice Hall International, Inc., 2009.

[7] A. Rencher, Multivariate Statistical Inference and Applications, New York: John Wiley & Sons, Inc., 1998.

[8] Y. Akbas and C. Takma, "Canonical correlation analysis for studying the relationship between egg production traits and body weight, egg weight and age at sexual maturity in layers," *J. Anim. Sci.,* vol. 50, no. 4, pp. 162-168, 2005.

[9] D. Chenfeng and W. Dongrui, "Canonical Correlation Analysis (CCA) Based Multi-View Classification: An Overview," *arXiv:1907.01693v,* pp. 1-4, 2021.

[10] L. Sun S and J. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation extensions and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 33, no. 1, pp. 194-200, 2011.

[11] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika,* vol. 58, no. 3, pp. 433-451, 1971.

[12] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Proc. Conf. Data Mining Data Warehouses*, 2010.

[13] P. L. Lai and C. Fyfe, "Canonical correlation analysis using artificial neural network," in *Proc. 6th Eur. Symp. Artif. Neural Netw*, 1998.

[14] S. Szedmak, D. R. Hardoon and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput,* vol. 16, no. 12, pp. 2639-2664, 2004.

[15] X. Chen, L. Han and J. Carbonell, "Structured sparse canonical correlation analysis," in *Proc. Int. Conf. Artif. Intell. Statist*, 2012.

[16] D. Gujarati, Ekonometrika Dasar, Jakarta: Erlangga, 1978.

[17] F. H. Joseph, C. B. William, J. B. Barry and E. A. Rolph, Multivariate Data Analysis, Seventh Edition, New Jersey: Pearson Prentice Hall International Inc, 2010.