Cluster analysis in term of sustainability and digitalization of Vietnamese regional economies

Trần An Quân

Faculty of Development Economics, University of Economics and Business, Vietnam National University, Ha noi, Vietnam

Corresponding author: Trần An Quân, (email: anquan@vnu.edu.vn)

Abstract

The development of Vietnam in Covid-19 period has received the good reviews from economists around the world. The economic growth is still positive, the digital application is gradually popularized and the economy demonstrates its sustainable resilience towards instabilities. However, within Vietnamese territories, the differences between 63 provinces are still remarkable in term of sustainability and digitalization. This paper applies empirical method of K-Means Clustering Analysis to classify provinces into 4 clusters with 4 input variables: Regional economic growth rate, Digital Transformation Index, Public Administration Reform Index and Public Corruption Control Index. The paper shows that cluster 2 with 11 provinces has brilliant performance compared with other clusters, they get a lot of motivation for further development. Meanwhile, cluster 4 with 24 underdeveloped provinces has difficulties at starting point, but they have large potential to develop in the future. Based on the characteristic analysis for each cluster, the research proposes relevant policies to create favorable condition to encourage the sustainable and digitalized development.

Key words: *K-Means Clustering, digital transformation, economic growth, development*

I) Introduction

1.1 Overview

Vietnam has demonstrated to become a fast, stable and dynamic economy in Asia. Even in Covid 19 pandemic period, Vietnam achieved 2.6% economic growth in 2021, maintained the reamarkable growth rate among countries. Also, Vietnam has gradually changed into digitalized economy with national plans of digital transformation. Around the world, many countries also target 4.0 industrial revolution with many high-tech industries and business to boost up the economy growth and to raise the Total Factor Productivity of the country. Therefore, the combination between fast-growing economy and digitalization progress is inevitable trend in the world, suggests the research to build comprehensive set of indicators reflecting the sustainability and digitalization of Vietnamese economy.

In the scope of regional economics, Vietnam shows the remarkable differences between high and low-income provinces. In 2021, province with highest monthly income per capita was Binh Duong, equaling 296,7 USD; meanwhile, the lowest province was Dien Bien, only reached 75.8 USD. In term of economic growth rate, during Covid 19 period, Hai Phong province achieved the highest rank with 12.38%; by contrast, the Ho Chi Minh city fell to the last position at -6.78% due to severe damage of Covid pandemic. Overall, the whole national economy still achieved 2.6% growth rate, the positive figure compared with other countries. However, the differences of regional economies urge the authorities and researchers to propose relevant policies for each province in accordance with their development conditions.

About the progress of digital transformation, there are the incredibly differences between provinces. The paper collects the data from Digital Transformation Index (DTI) listed by Vietnamese Ministry of Information and Telecommunication in 2021. Specifically, in the range from 0 to 1, Da Nang scored 0.64 ranked at the highest position with abundant changes in the technology, whereas Bac Lieu only scored 0.25, at lowest point in the list. These indexes are comprehensively synthesized from 3 categories: digital government, digital society and digital economy, cover all sectors within a nation. Although, there are arguments about limitations and relevance of these indexes, they are the first measured indexes about regional digital transformation in Vietnam.

The differences between regional economies in term of economic growth and digital transformation degree encourage the research to classify these economies into clusters. With the scope of data science, the cluster analysis is necessary to take comprehensive view about development level between regions. Specifically, the paper applies K-means clustering algorithm to classify the regions into groups with optimal number of clusters. About variable in this paper, sustainability is represented by economic growth rate, Public Corruption Control Index, Public Administration Reform (PAR) index, and digitalization progress is measured by the DTI index. Actually, besides economic growth, the public administration and corruption control also incredibly contribute to the sustainability and stability of the development. With clear clustering work, the policy maker might propose appropriate solutions to promote the development of each region.

1.2 Literature review

Many researchers have completed their academic paper relating to cluster analysis in many fields of economics and business, including regional economics. Karakoc.O et Al (2019) applied clustering for provinces of Turkey. They evaluated the development level of Turkish provinces through grey cluster analysis with indexes of life expectancy, per capita income and education. On the other hand, Hana. R (2014) summarized methods of cluster analysis including Hierarchical clustering, K-means clustering and other transformed types, summarized some applications for classifying and analyzing economies of Eastern European countries. She also mentioned other techniques, such as cluster tree and two-step clustering. However, the author did not focus deeply on any kind of analysis because the research scope is relatively large. Thus, the reviews on all methods are insufficient, and the number of papers mentioned are small. In Eastern Asian area, Alexandre.R (2012) implemented research on cluster analysis to classify East Asian economies and evaluated on the similarity between them. He used 2 methods including hierarchical clustering and K-means clustering to classify East Asian economies into 4 groups with similar characteristics. The author approved 4 groups of variables to insert to the model: Structural shares, International trade, Economic development and Economic size; each variable group contains some specific variables. In analytic work, the cluster tree technique is also added to make linkage between groups.

In Vietnam, Lich H. K et Al (2021) applied cluster analysis to classify countries in the world in term of government expenditure. They clustered countries into 4 groups with similar characteristics; 5 variables for classification included: compensation of employees, goods and services expense, merchandise trade, government effectiveness and GNI per capita. The author's results show that only 5 countries ranked at Group 1 with highest performance. By contrast, most of the developing countries performing worst in term of public expenditure, ranked at Group 4. In financial market, Binh.T.T.D and Khanh H.B (2022) applied hierarchical cluster analysis for 33 Vietnamese commercial banks with 12 variables for each banks. Based on the operational and financial indicators, the research aimed to recognize the vulnerable banks to prevent from the bankruptcy. In business and administration field, Tam.P.T et al (2021) applied cluster analysis to fulfill the data-driven customer segmentation for Vietnamese SMEs. They realized many clustering techniques in their research including K-means clustering to fulfill customer segmentation for SMEs based on given dataset. The research also showed the superiority of data science to analyze the big dataset compared with traditional methods.

With the reviews on cluster analysis, they show that K-means clustering has solved a lot of problems in economics and business administration fields. Additionally, this method has been used to analyze and classify regional economies in other countries. However, there is lack of research applying cluster analysis for regional economies in Vietnam. Especially, the research which relates to digitalization factor is essential for Vietnamese government guideline. Therefore, the paper not only completes the shortage of clustering analysis for regional economies, but also refers to the digitalization progress for each province integrating with Vietnamese governmental programs.

II) Methodology

The empirical analysis is completed with the application of K-means Clustering algorithm, refers to vector quantization. It is also called a centroid-based algorithm or distance-based algorithm. The paper finds the centroid for each cluster and measures the distance of each data point within a cluster to the neighborhood. The target is to minimize the distance between all data points to their centroid within a cluster. Notwithstanding, the algorithm needs to ensure each data points has the similar characteristic reflecting the common features of the cluster.

Suppose N data points: $\mathbf{X= [x_1, x_2,...,x_N]} \in \mathbf{R}^{dxN}$, with K is the number of clusters. It is necessary to find centroids: $\mathbf{m_1, m_2,...m_k} \in \mathbf{R}^{dx1}$ and labels of each data point. Mathemetically, contraint of yi is possibly written in the type of

$$y_{ik} \in \{0,1\}, \quad \sum_{k=1}^{K} y_{ik} = 1 \tag{1}$$

Loss function and optimization are fundamental problems to deal with K-means Clustering. Suppose that $\mathbf{m_k}$ is centroid for each cluster, data point $\mathbf{x_i}$ scattered in $\mathbf{m_k}$ would have the error equaling $\mathbf{(x_i - m_k)}$.
The error for all dataset has type of

$$\mathcal{L}(\mathbf{Y}, \mathbf{M}) = \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2 \tag{2}$$

Where $\mathbf{Y}= [y_1, y_2,...,y_N]$, $\mathbf{M}= [m_1,m_2,...,m_K]$ are matrices created by labelled vectors of each data points and centroids, respectively. The Loss function of K-means clustering problem is $\mathscr{L}\mathbf{(Y, M)}$. The equation is rewritten with the optimization for Loss function through argument of minimum (arg min)

$$\mathbf{Y}, \mathbf{M} = \arg\min_{\mathbf{Y},\mathbf{M}} \sum_{i=1}^{N} \sum_{j=1}^{K} y_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2 \tag{3}$$

This equation is solved by giving the differential of the function to equal zero. The research finds the $\mathbf{m_j}$ solution as follow:

$$\mathbf{m}_j = \frac{\sum_{i=1}^{N} y_{ij} \mathbf{x}_i}{\sum_{i=1}^{N} y_{ij}} \tag{4}$$

Where $\mathbf{m_j}$ is an average of number of data points in cluster j

The determination about the number of centroids is also necessary to enhance the accuracy of clustering. Therefore, the paper applies the Elbow method to define the optimal number of clusters for analysis. In data science, number of clusters fits the model with the value of K, and K commonly ranges from 0 to 10. The elbow method plots the value of Loss function generated by different K values. When the K value raises, each cluster will have fewer constituent instances, the instances then go closer to the centroid. Elbow in the illustrated graph exists when improvement of average distortion declines the most. Mathematically, the research calculates the sum of squared errors (SSE) of distance between data points to with any given K value. It is represented by Silhouette coefficient as follow.

$$d = \sum_{i=1}^{k} \sum dist(x, c_i)^2 \tag{5}$$

Both Cluster Analysis and Elbow method are implemented through Jupyter Notebook program with Python programing language. Researchers in data science immensely applied this program in their work and it has enough libraries to support analyze and export intuitive graphs.
However, some analysis work is supported by SPSS, such as the ANOVA analysis for each variable.

III) Empirical research and Results

3.1 Data

The research has collected data in the many resources for total 63 administrative provinces in Vietnam. It is updated from newest statistics in 2021; however, the macroeconomic data might be distorted due to impact of Covid

19 pandemic. The provincial growth rate is synthesized from the General Statistics Office of Vietnam. Besides, the paper has achieved corruption control indexes from PAPI, a research program from Centre for Community Support and Development Studies (CECODES) and United Nation Development Program (UNDP) in Vietnam. Continuously, The PAR indexes are collected from the Report on Public Administration Reform of Ministry of Home Affairs. With DTI indexes, the paper collects them from the ranking list of Ministry of Information and Telecommunication. Generally, the data is synthesized from the reliable sources of public offices and organizations in Vietnam. About the symbol used in the model, beside DTI and PAR, the research uses COR standing for public corruption control index and GROW represents regional economic growth rate. The statistics of 4 variables are illustrated in the following descriptive statistics table.

Descriptive Statistics

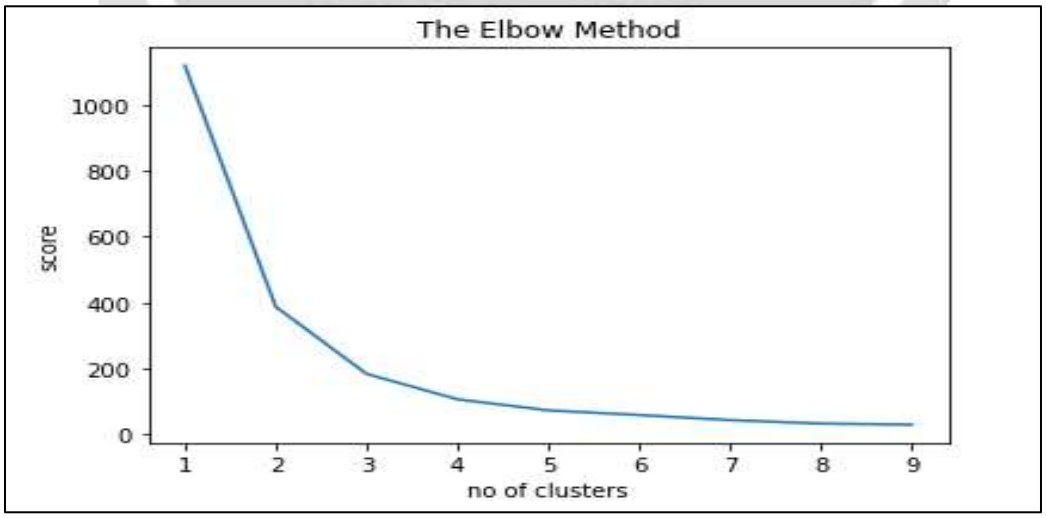
	N	Minimum	Maximum	Mean	Std. Deviation	Variance
DTI	63	.25	.64	.4011	.08932	.008
PAR	63	79.97	91.80	86.3697	2.33629	5.458
COR	63	5.42	8.29	6.8667	.52981	.281
GROW	63	-6.78	12.38	3.8195	4.21306	17.750
Valid N (listwise)	63					

Table 1: Descriptive Statistics for 4 variables

Source: Author's result from SPSS software

3.2 Results and analyses

Before clustering the dataset, research applies Elbow method to determine the optimal number of clusters. Through computer programming, the paper receives the Elbow method graph showing maximum of 10 groups. The graph shows Elbow curve is strongly kinked in value of 2 and 3. The curve finally distorts at number of 4 groups, helps to determine the number of clusters and after that value, it becomes horizontal. Besides, in the graph, scores of SSE decreases along with the increase in number of clusters. Hence, the paper chooses the value of 4.



Graph 1: Elbow curve for defining optimal number of clusters

Source: Author's result from Python programming

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
DTI	.024	3	.007	59	3.371	.024
PAR	34.837	3	3.964	59	8.788	.000
COR	1.120	3	.238	59	4.705	.005
GROW	313.732	3	2.700	59	116.202	.000

Table 2: ANOVA of 4 variables with significance level

Source: Author's result from SPSS software

The paper also applies ANOVA analysis to evaluate whether variables in the model help to discriminate across the different clusters. All variables have significance level at 5%, suggest all clusters in the model are different. The paper concludes that all variables imported to the model are appropriate for clustering work.

For the next step, cluster analysis is implemented to separate provinces into 4 groups. The computer programs show the results as table 3 in the Annex. The research groups 8 provinces in the cluster 1; 11 provinces in the cluster 2; cluster 3 includes 20 provinces and the last one contains 24 provinces. In general, cluster 1 includes big provinces and cities in Southern Vietnam, whereas most of the provinces in cluster 2 are dynamic and fast-growing economies in Northern Vietnam. It is also reasonable to regard provinces in cluster 4 as underdeveloped regional economies with lower starting point; however, they have a lot of room and opportunities to develop in the future.

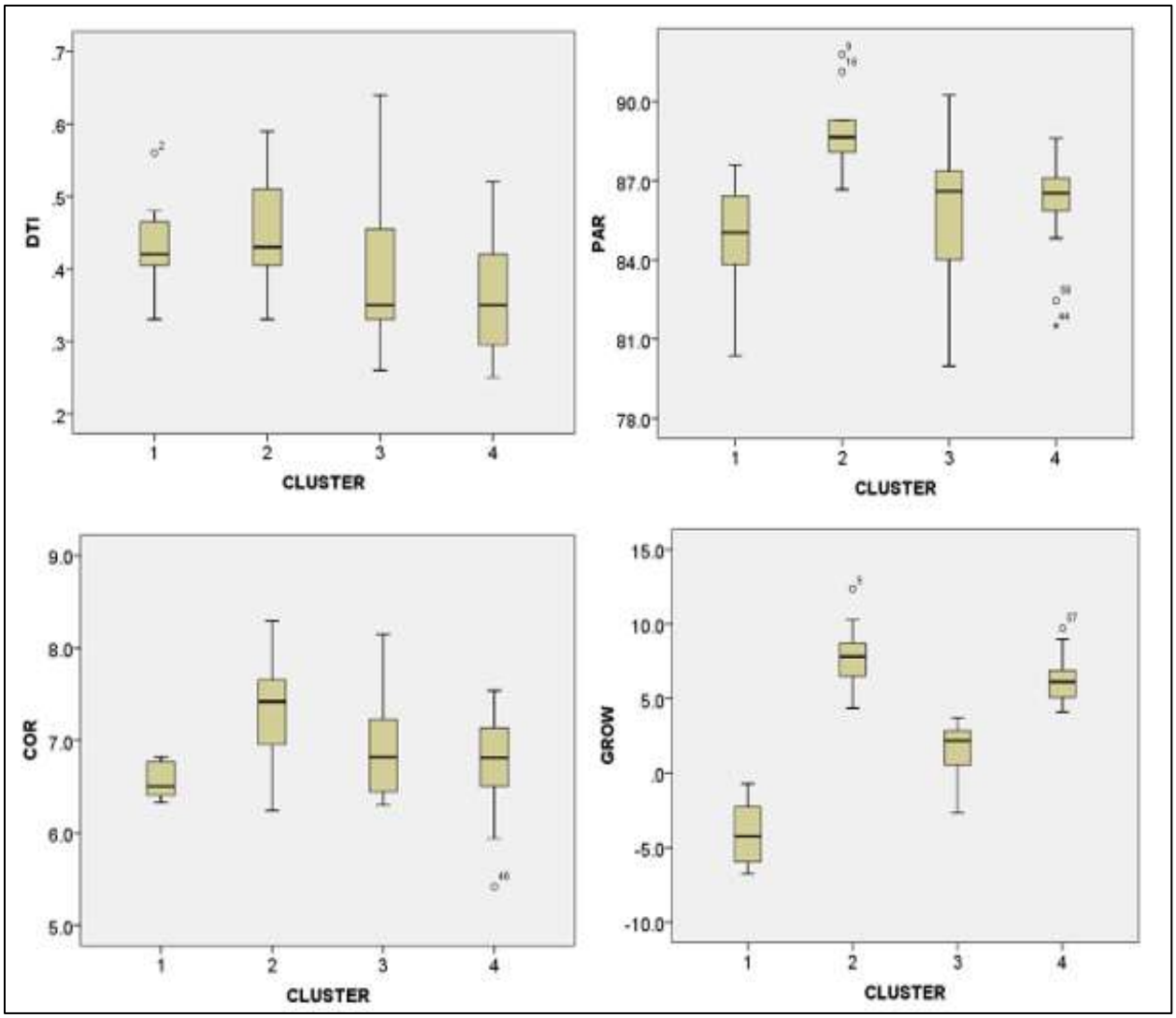
To determine the characteristic of clustering, the paper visualizes indicators of clusters by box plot diagram. It help to differentiate the particular characteristics of 4 clusters based on the performance of indicators. Similar to number of clusters, the research also demonstrates 4 box plot diagrams combined in graph 2. Specifically, the PAR variable contains large differences between data values in each cluster extending the whisker of box plot. Meanwhile, the COR variable shows smaller differences between data values in each cluster, and differences between clusters are also small. The GROW variable demonstrates really big differences between clusters, but in each cluster, distance between quartile 1 and quartile 3 is short proving the data values are close each other.

In the detailed analyses of each cluster, the cluster 1 shows high DTI variable but medium values of PAR and COR; this cluster also has the low growth rate. This is actual situation for big cities or provinces in Southern Vietnam in 2021. They own relatively good infrastructure and conditions to develop, and their development degree is good, but they had to face severe Covid 19 pandemic that slumped the economic growth. The box plot shows median value of DTI is high, above 0.4; meanwhile, the median value of growth rate equals -4%, relatively low compared with other groups.

The cluster 2 includes fast-growing and industrial provinces, mostly located in Northern area. This cluster scores well at digital transformation, reaches the same DTI level with the cluster 1. PAR, COR and GROW graphs also show cluster 2 achieves the highest positions. The results demonstrate the fast growth, improved technology and sustainability in term of economics and administration. Also, these provinces did not heavily suffer from Covid-19 consequences, so that they might maintain their development motivation. Hence, this group of provinces has the most favorable condition to continue improve their performance in all indicators

Medium degree of all indicators is the main characteristic of cluster 3. They include provinces and cities scoring medium or below-average in all DTI, PAR, COR and GROW indicators. In comparison with other clusters, they have low DTI indexes, the economic growth rate is low, yet much more improved than cluster 1. Most of them suffered from remarkable damage of Covid 19 in 2021, dragged down the whole economic performance; however, they still maintained the economic growth in positive value. This cluster does not perform well in general; therefore, the comprehensive reformation is necessary to improve all of their indicators.

At last, cluster 4 contains provinces performing low in DTI, PAR and COR indicators, but scores well in economic growth. These provinces are currently underdeveloped, located in mountainous or rural areas that face a lot of difficulties to develop the regional economy. However, they have a lot of room to improve the development in long-run. The evidence is that they scored well in term of economic growth and resources for development are abundant in the future. They also avoided the damage of Covid 19 due to their geographic location and population density. Therefore, despite the low starting point, these provinces own ample potential to achieve higher development level.



Graph 2: The plot box diagrams of 4 variables for 4 clusters

Source: Author's results from SPSS software

IV) Conclusions and policy implications

4.1 Conclusion

The paper applies cluster analysis with K-means method to classify Vietnamese provinces in term of economic growth, digital transformation, public administration and public corruption control. With elbow method

and ANOVA analysis, the research classifies provinces into 4 clusters. Specifically, the results show that cluster 2 with many northern industrial provinces achieves high ranking with fast economic growth and good technology bases. These provinces do not suffer from heavy damage of Covid-19 creating good motivation for further development. Meanwhile, cluster 1 with many big southern provinces faces severe impacts of the pandemic, slump down in economic growth with negative values despite of good digital transformation base. Cluster 3 shows the sluggish movement in growth rate and public administration reform, needs much more improvement in all indicators. Finally, Cluster 4 contains underdeveloped regional economies achieve high ranking in term of growth rate despite of slow digital transformation progress. This group has a lot of potential to reach higher development degree in the future.

4.2 Policy implication

In order to improve the sustainable development as well as digitalized regional economy, the research suggests policies to each cluster of provinces as follow. These recommendations are based on the above analyses.

At first, cluster 1 including HCM city and other big cities or provinces in Southern Vietnam has suffered from heavy damage of Covid 19, needs to receive policies to improve economic growth, through foreign investment and resident consumption encouragement. The government should continue to offer incentives and tax reduction for this cluster, such as tax extension and preferable loans for damaged economic sectors. Expansion of credit room also support necessary amount of capital for commodity producers and key industries to maintain the production activities

The cluster 2 including Hai Phong, Quang Ninh, Bac Ninh are big industrial zones in Northern Vietnam. The government should maintain the current policies to ensure growth motivation. In this cluster, the province's starting point in term of digital transformation and development degree are relatively good compared with other areas. Furthermore, the pandemic has not impacted on these provinces, so that they can continuously utilize convenient conditions and opportunities to develop.

Hanoi and other rural provinces are ranked in cluster 3, faces moderate difficulties in the pandemic. Similar to cluster 1, they should receive many supports from government to improve economic performance; however, most of them have positive economic growth that might be rapidly improved. The corruption control and public administration simplifying are important goals targeted to enhance sustainability. Besides, in scope of digital transformation, investment for technological infrastructure and ICT platforms also needs to be fulfilled to facilitate the economic growth.

The last cluster contains almost mountainous and remote areas achieved high economic growth due to the large room for development. Notwithstanding, these provinces have low starting points; therefore, they need receive public investment policies concentrating on particular regional industries, digital infrastructure as well as ICT platforms to modernize the economy. Public administration reform is also essential issue to support their development. Finally, owing to natural conditions, the strict control of resources use also leads these provinces to sustainable development pathway.

References

General Statistic Office of Vietnam (2022), Statistical yearbook of Vietnam 2021, Statistical Publishing house

CECODES, VFF-CRT, RTA & UNDP (2022), The Vietnam Provincial Governance and Public Administration Performance Index: Measuring citizens's experiences 2021, a Joint Policy Research Paper by Centre for Community Support and Development Studies (CECODES), Centre for Research and Training of the Viet Nam Fatherland Front (VFF-CRT), Real-Time Analytics (RTA), and United Nations Development Programme (UNDP), Hanoi, Vietnam.

Ministry of Home Affairs (2022), Report of Public Administration Reform Index 2021

Ministry of information and Telecommunication (2022), Digital Transformation Index Ranking 2021, https://dti.gov.vn/

Alexandre R (2012), How Similar Are the East Asian Economies? A Cluster Analysis Perspective on Economic Cooperation in the Region, Journal of International and Area Studies, Vol 19, 2012, pp.27-44

Karakoc.O et Al (2019), Evaluation of the development level of provinces by grey cluster analysis, Journal of Procedia Computer Science, Vol.158, pp. 135-144, ISSN 1877-0509

Hana.R (2014), Cluster Analysis of Economic data, Journal of Statistika, Vol.94, pp.73-86

Lich K.H et Al (2021), Classifying Countries in Terms of Government Expenditure: A Multi-criteria Approach, HSE Economic Journal, Vol 25, No.4, pp.610-627

Binh.T.T.D (2014), Cluster Analysis of Vietnamese Banks, SSRN Electronic Journal, 10.2139/ssrn.2543094

Tam.P.T et Al (2021), Data Driven Segmentation for Vietnamese SMEs in Big Data Era, Journal of Macro Management & Public Policies, Vol.3, Iss.2, pp.33-43

Shehroz S.Khan & Amir Ahmad (2004), Cluster center initialization algorithm for K-means clustering, Journal of Pattern Recognition Letters, Vol.25, Iss.11, pp.1293-1302

Ian H.Witten et Al (2016) Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publisher, 4th edition

Trevor.H, Robert.T, Jerome.F (2016), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics), Spriger Publisher, 2th editon

Pratap Dangeti (2017), Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R, Packt Publisher, pp.305-320

Alberto.V (2017), Vietnam's Key Regions and Economic Zones, Vietnam Briefing Consultancy, iss.31, www.vietnam-briefing.com

ANNEX

Cluster table

Cluster 1		Cluster 2		Cluster 3			Cluster 4
No	Province	No	Province	No	Province	No	Province
1	Khanh Hoa	9	Hai Phong	20	Ha noi	40	Tuyen Quang
2	HCM city	10	Bac Ninh	21	Cao Bang	41	Yen Bai
3	Ba Ria -Vung Tau	11	Hai Duong	22	Bac Kan	42	Thai Binh
4	Vinh Long	12	Phu Tho	23	Hoa Binh	43	Nam Dinh
5	Dong Thap	13	Vinh Phuc	24	Lai Chau	44	Ha Nam
6	Tra Vinh	14	Thai Nguyen	25	Son La	45	Ninh Binh
7	Can Tho	15	Bac Giang	26	Da Nang	46	Dien Bien
8	Tien Giang	16	Quang Ninh	27	Phu Yen	47	Ha Giang
Sum: 8		17	Hung Yen	28	Lam Dong	48	Lang Son
		18	Thanh Hoa	29	Binh Thuan	49	Lao Cai
		19	Hue	30	Dong Nai	50	Nghe An
		Sum: 11		31	Tay Ninh	51	Ha Tinh
				32	Binh Duong	52	Quang Binh
				33	Long An	53	Quang Tri
				34	An Giang	54	Quang Nam
				35	Ca Mau	55	Quang Ngai
				36	Hau Giang	56	Ninh Thuan
				37	Soc Trang	57	Binh Dinh
				38	Ben Tre	58	Gia Lai
				39	Kien Giang	59	Kon Tum
				Sum: 20		60	Dak Lak
						61	Dak Nong
						62	Binh Phuoc
						63	Bac Lieu
						Sum: 24	

Table 3: Cluster table for 63 provinces in Vietnam

Source: Author's results from Python programming and SPSS software