ISSN No. 0976-5697

Volume 6, No. 1, Jan-Feb 2015

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Analysis of Regional Development Disparity with Clustering Technique Based Perspective

Tb. Ai Munandar
Information Tech. Faculty – Informatics Eng. Dept
Universitas Serang Raya (UNSERA)
Banten Province - INDONESIA

Azhari, SN
Faculty of Math. & Natural Sciences
Universitas Gajah Mada (UGM)
Yogyakarta - INDONESIA

Abstract: The disparity in development is a situation where there is a difference between the achievements of development within the region, with a view of a particular indicator. Many of the techniques are can be used for the determination of the disparity. This paper discusses the determination of regional development disparities seen from the point of view of data mining using the cluster approach to the data. This paper put forward the discussion of literature on several research studies conducted by scientists associated with the determination of disparity development using cluster technique. Case studies related to the grouping level of development in order to see the disparity in regional development are also discussed in this paper for a complete description of the study of literature is done. K-means clustering method is used with the help of MATLAB applications to analyze the data of gross regional domestic product (GDP) of the province of Central Java. The analysis showed the K-means cluster has an accuracy rate of 80% on the level of development of the region grouping. Of the fifteen districts are analyzed, only three counties that do not fit the classification results when compared to the results of the cluster with Klassen typology.

Keywords: Disparity in development, clustering technique, k-means, MATLAB, GDP, Klassen typology

I. INTRODUCTION

The disparity in regional development is a concept that relates to how a region or country has a level difference to the economic well-being and its power (OECD 2003 in [1]). Determination of the development disparity regions is important to look at the extent of achievement of the development of a region. In addition, the determination of disparity is also able to show the development gap that occurs over time. There are many techniques that can be done to determine the regional development disparities, see reference [2], [3]. Other approaches related to data mining, one of which is by using a clustering technique.

Clustering is able to divide the data into specific groups by nature, despite not having a class as the classification technique [4]. Data objects are grouped based on information available to the data and relationships that exist in it. The goal is, the data objects that have similar characteristics are grouped into the same group [5]. Clustering technique itself has been used for a variety of data analysis needs. Some of them for example, to identify genetic expression patterns to classify cancer [6], the identification of the user session on the current Web search high web user activity [7], grouping documents [8], the classification of pixels flame to improve the quality of digital image analysis [9] , and many more use of cluster techniques to help solve real-world problems.

This paper discusses how clustering technique is used for the determination of regional development disparities based on certain indicators. The discussion in this paper is a literature review of the activities of some of the results of research conducted by scientists associated with the use of clustering techniques in the field of regional development. It is important to address in order to provide new perspectives on the approach that is widely clustering techniques have been widely used for determining the achievement of regional development disparities. Furthermore, the study of the literature study is expected to contribute to the development of science in the field of informatics and computer science to other disciplines.

The discussion paper is divided into six main sections. The first part is an introduction which includes the background paper topic selection. The second and third section outlines the theoretical basis related to the disparity in the development and clustering techniques. The fourth section outlines some of the relevant literature study clustering technique for the determination of regional development disparities. This paper also discussed on a case study of regional development level classification using cluster technique, discussed in the fifth section. The last section is a discussion of the cover containing the conclusions and suggestions for future research.

II. REGIONAL DEVELOPMENT DISPARITIES

The disparity in terms of development is closely related to regional disparities. This term indicates a difference in the level of economic achievement and well-being of a region compared to other regions (OECD 2003 in [1]).

Geographic disparities are generally divided into two perspectives. First of vertical viewing angles is where disparities region is seen as a change based on the size of the geographical area concerned. Second, is the horizontal viewing angle, where the disparity seen by differences in the level of social, economic and conditions of a region [1]. Some cases (especially in the literature of this review), the second perspective is more widely adopted by some scientists to measure disparities in the region to see the difference in the achievement of its development.

Wishlade-Yuill (1997) in [1] suggests that the structure of the disparity of a region is divided into three main areas:

a. Physical Disparities. Related to environmental conditions and geographical possessed a certain region. Assessment disparity is usually quite difficult and

sometimes put forward the opinion of those who do the assessment.

b. Economic Disparities. Associated with differences of a region in quality and quantity. Assessment is usually done using indicators Gross National Product (GNP), which combined with the analysis of tax revenues, the growth of industry, demographic trend, infrastructure and services.

c. Social Disparities. Related to income and living standards of the population. Most countries usually pay more attention to indicators of unemployment.

But in this paper, the disparity in development will be assessed based Klassen typology. This typology divides the disparity of development into four quadrants Based on the construction of indicators of regional gross domestic product (GDP). The fourth quadrant are Quadrant I (areas with indications of advanced and fast-growing), Quadrant II (regions with advanced indication yet depressed), Quadrant III (with an indication of the fast growing area) and Quadrant IV (areas with relatively low indication)..

III. CLUSTER TECHNIQUE

Clustering is a process that is performed to determine a set of objects fit into certain groups based on the similarity of objects from one another [5]. Mathematical explanation of the definition of clustering is described as follows [6]: Let X = $\{x_1, x_2, x_3 \ldots x_{m-1}, x_m\} \subset R^n$ which is a compilation of a number of data representations of the set *m* to x_i in R^n, where $x_i = \{x_{i1}, x_{i2}, x_{i3} \ldots x_{in}\}$. The goal is to divide X into a number of groups *k* $\{C_i: 1 \leq i \leq k\}$, and k groups is called clusters. The results of this cluster algorithm mapping of data items x_i, to a group C_k.

In general, clustering techniques are divided into two types, namely Hierarchical and Partitioned clustering [10]. Specifically in this paper, the discussion on the fifth section describes the classification level of development of the region using the technique partitioned clustering with K-means clustering method. Use of partitioned clustering techniques is because it is able to handle large amounts of data, and is able to produce a number of output clusters as expected [10]. The reasons behind the selection of K-means clustering technique in this paper, considering the number of clusters to be adapted to the type of development in accordance with the classification level Klassen typology. Here is a step-by-step grouping using K-means clustering technique [10]:

a. Determine the number of clusters
b. Allocate the data into clusters randomly
c. Calculate the centroid / average of the data in each cluster.
d. Allocate each centroid data to / average nearest
e. Go back to Step 3, if there is still data to move the cluster or if the change in value of the centroid, there are above a specified threshold value or if the change in value of the objective function used, above a specified threshold value

IV. DEVELOPMENT DISPARITIES UNDER CLUSTERING PERSPECTIVE

This section discusses some of the studies related to the determination of regional development disparities using cluster technique. Many scientists analyze the determination of regional development disparities using clustering techniques. K-means clustering method is used to classify the socio-economic indicators in order to classify an area of development disparities that have been achieved [11], [12], [13]. The study, conducted by scientists at the Portuguese against 33 indicators of socio-economic region, was able to show differences in the level of development based on the grouping of the five regions in Portuguese. In addition, based on the results of cluster analysis, it was found that the coastal regions have a better level of development than in the urban area [11]. Classification disparity in regional development also applied scientists to analyze grouping 11 socio-economic indicators of the state-owned Croatia. Cluster process performed on the data entity similarity of each region according to the indicators that have been determined. The goal is to formulate a regional development policy measurement standards based on the results of the cluster to the achievement of development through the determination of geographic disparities [12]. The use of K-means clustering is also done by [13] for grouping changes disparity of a region using regional indicators of competitiveness and spatial proceses. Results of cluster forming clusters of three regions according to the indicators of regional competitiveness owned.

Other cluster methods used for grouping geographic disparities is Ward technique [14], [15], [16]. Research conducted [14] to classify the West German state and the East Germans based on 12 economic indicators that exist, indicate that, after 12 years together, East Germany is still not able to catch up to the West German economy. Nevertheless, Ward cluster analysis also found there were some areas in East Germany, have economic development that is almost on par with some of the other regions in West Germany. Grouping of regional development disparities in the Czech Republic also performed using Ward technique and form three clusters. The first cluster contains regions with very low living standards, the second cluster contains regions with average living standard and third cluster contains regions with a very high standard of living. Significant differences between the visible region with other regions based on the results of cluster analysis which formed [15].

Similarly, the state of Germany and the Czech Republic, grouping disparity development using Ward technique was also applied to the Ukrainian state [16]. Grouping is done based on the indicators of Gross Regional Product (GRP), the industrial production index, the index of fixed capital investment, foreign direct investment per capita, employment, number of Organizations, conducting scientific research and the total value of innovation costs. The results of the cluster using this technique, is dividing the territory into three clusters. The first cluster shows the estimated economic rate is relatively higher compared to the second and third cluster. The second cluster contains regions with the estimated level of development that is unstable, while the third cluster contains regions with the estimated level of development that is often fickle. This study also proposes some policy development can be done by the government of Ukraine based on the results of its GRP data classification. Regions that belong to the first cluster for example, became a major locomotive area for future development innovation. Region into the second cluster, an area which requires increased per capita income and other infrastructure

development. As for the third cluster, an area that still requires special attention in the field of increasing capital investment, education, subsidies and so on.

Several other studies, combining the method of K-means clustering with Ward to strengthen clusters formed outcomes [17], [18], [19]. The first stage carried out using Ward techniques to obtain the expected number of clusters. The number of clusters of Ward techniques is later used as the cluster centroid using the K-means. In grouping regional development disparities in the EU are carried out [17], obtained three cluster regions. Each cluster labeled in accordance with the level of achievement of the development of an area by disparities that occur. The first cluster is an area with better development. The second cluster is the region with the construction of the transition. While the third cluster, an area with a low level of development. The results of the different clusters, is shown in studies [18] which classifies regions into four clusters. The first cluster and the second are showing the areas with good economic performance. The third cluster is a region with a large educational gap, while the fourth cluster showed a positive value of the factor levels of unemployment, education and demographics. In different countries, namely Pakistan, especially Punjab region [19], socio-economic indicators of the Multiple Indicator Cluster according to data from Survey (MICS) is used to see the achievement of development results in accordance with the MDG targets. Cluster analysis results also showed that there are inequities in the spread of the sources driving the development so that the level of development in the Punjab region can be uneven.

Other methods such as within-group linkage, complete linkage [20] and medoid partitioning algorithm [21] can also be used to classify the regional development disparities. The similarity of the characteristics of the economic development of every region in the country of Romania, grouped into five clusters by region [20] using three techniques at the same cluster. The first cluster is a region with a moderate level of economic development. The second cluster is a region with a low level of economic development. The third cluster is a region with a better level of development. The fourth cluster is the region with the construction of infrastructure and public facilities are growing, while the fifth cluster is an area that is affected by the position of the neighboring region.

Not only the light of the economic level has been achieved, the disparity can also be seen with the grouping of the indicators of the level of income in agriculture. This is done by [21] in the EU countries and revealed a number of countries that have a high level of income in agriculture, such as Belgium and the Netherlands. In addition, there are also countries that have low levels of income in agriculture such as Romania, Poland, Portugal, Hungary, Bulgaria, Slovenia, Greece and Italy. The results of cluster analysis can certainly be used as an evaluation tool to see the difference in incomes in agriculture in the countries of the European Union.

Several studies above demonstrate how cluster technique provides an important contribution to the determination of regional development disparities. The results of the cluster are able to categorize outline statistical data to regional development in a particular group. Cluster formation can also be adapted to the needs of a region grouping adapted to non-cluster method. However, not all the results of the cluster can form a group and incorporate the area into the same group.

Especially when compared to the results of cluster-cluster method no, as the results of research carried out [17] and [18]. Where the results of the cluster analysis groups the region has a different member of the group when compared with the development disparity classification established by the European Commission (EC). The difference in these results on the other hand, it could be used as a tool for the re-evaluation of non-cluster method is used as a classification method regional development disparities. Or perhaps, on the contrary, that it is also two methods, both cluster and non-cluster can be combined to give the results of grouping disparity better development.

In short it can be said, the results of cluster analysis can be used as an alternative method that can help grouping of a region based on indicators of economic, social, territorial and other indicators. Analysis of the results of the cluster can of course also be used as consideration for the custodian of government policy to determine the future direction of development rate. As well as with formulating policy direction of development, according to the grouping is done based on the similarities and characteristics of the data using cluster technique.

V. CASE STUDY: GROUPING REGIONAL DEVELOPMENT USING CLUSTER TECHNIQUE

This section discusses the level of development of the region grouping analysis using the K-means method. The data used is statistical data Gross Regional Domestic Product (GDP) of the province of Central Java by taking samples of the data as much as 15 districts for 2011 and 2012. The GDP data were then calculated the average growth rate and the average contribution to the development of both the district and provincial levels. Once the data is then classified using Klassen typology. This was done to compare the results of the cluster data classification techniques commonly used by statistical agencies of development, both national and local.

Cluster analysis was performed using MATLAB applications and utilizes k-means clustering functionality already present in it. Determination of the number of clusters adapted to the classification level of development according to Klassen typology, which is about 4. A total of 135 datasets were tested using the K-means method. Then the cluster results compared with the results Klassen typology to analyze differences in results between the two methods. Testing the cluster to 135 datasets performed 10 times to see the difference in the level of accuracy and cluster results.

The amount of the percentage of GDP indicator data is classified into a particular cluster to 135 datasets, according to Klassen consecutive typology is, the first cluster of 28.89%, the second cluster of 4:44%, the third cluster of 0.74% and 65.93% for the fourth cluster. In the first cluster test, the percentage of data indicators grouped into clusters each row is, the first cluster of 22:22%, the second cluster of 14.81%, the third cluster of 11:11% and 51.85% for the fourth cluster. The results of the cluster in the first test until the tenth are show a fluctuating level of accuracy when compared with the results Klassen typology. Percentage of the data cluster with 10 times the test is shown in Table 1. Table 1 is also shows the magnitude of the percentage of data simultaneously entered into a certain cluster to 135 datasets were used.

Table I. Differences in percentage 10 times testing the typology Klassen

	C – 1	C – 2	C – 3	C – 4
Klassen	28,89%	4,44%	0,74%	65,93%
Tes-1	22,22%	14,81%	11,11%	51,85%
Test-2	3,70%	20,74%	64,44%	11,11%
Test-3	21,48%	35,56%	30,37%	12,59%
Test-4	3,70%	31,85%	34,07%	30,37%
Test-5	30,37%	3,70%	31,85%	34,07%
Test-6	35,56%	30,37%	21,48%	12,59%
Test-7	21,48%	22,22%	43,70%	12,59%
Test-8	22,22%	64,44%	2,22%	11,11%
Test-9	14,07%	31,85%	21,48%	32,59%
Test-10	4,44%	31,85%	22,22%	41,48%

Table II. Comparison of the average test between K-means and Klassen

	Klassen	Cluster
Clust – 1	28,89%	17,93%
Clust – 2	4,44%	28,74%
Clust – 3	0,74%	28,30%
Clust – 4	65,93%	25,04%

If in Table 1 illustrates the magnitude of the percentage of the data is classified into a particular cluster, then Table 3 shows the percentage of suitability grouping results of the classification results Klassen cluster. The results of the percentage calculation suitability grouping results in cluster 2 and 3 show a very significant difference. This percentage is obtained by comparing the results of the cluster on the n-th test against Klassen classification results (see Figure 1 for the visualization of the distribution of the data to the cluster for each test).

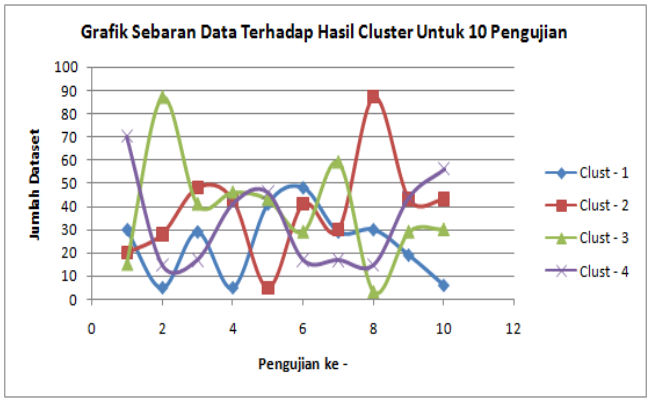

Figure 1. Distribution of data on the results of the cluster

Grouping accuracy rate of development of the region results cluster results, achieved by 80%. This means that, of the 15 districts that have been classified using typologies Klassen, when compared with the results of K-means cluster, there are only three districts that do not fit. This means that, grouping the results of K-means cluster technique is still acceptable (see Figure 2).

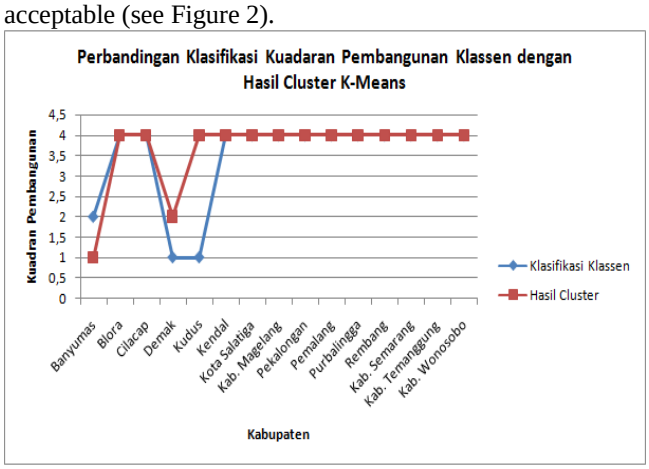

Figure 2. Comparison of the results of the cluster with Klassen

Table III. Percentage distribution of the data to the cluster

	C - 1		C - 2		C - 3		C - 4	
Klassen	39		1		89		39	
	The distribution of data to the cluster				Percentage distribution of the data to the cluster			
	C - 1	C - 2	C - 3	C - 4	C - 1	C - 2	C - 3	C - 4
Tes-1	30	20	15	70	76,9%	333,3%	1500,0%	78,7%
Test-2	5	28	87	15	12,8%	466,7%	8700,0%	16,9%
Test-3	29	48	41	17	74,4%	800,0%	4100,0%	19,1%
Test-4	5	43	46	41	12,8%	716,7%	4600,0%	46,1%
Test-5	41	5	43	46	105,1%	83,3%	4300,0%	51,7%
Test-6	48	41	29	17	123,1%	683,3%	2900,0%	19,1%
Test-7	29	30	59	17	74,4%	500,0%	5900,0%	19,1%
Test-8	30	87	3	15	76,9%	1450,0%	300,0%	16,9%
Test-9	19	43	29	44	48,7%	716,7%	2900,0%	49,4%
Test-10	6	43	30	56	15,4%	716,7%	3000,0%	62,9%

VI. CONCLUSION

Based on the study of literature and studies clustering level of development, it can be seen that, essentially cluster technique can be used as an alternative approach to determining regional development disparities according to certain indicators. Cluster results obtained sometimes have significant differences when compared with the non-cluster classification method in general. It is very reasonable, because the concept of the cluster itself departs from the activities of grouping data based on common characteristics of the data are then calculated and dissimilarity its similarity distance.

In contrast to non-cluster classification techniques, such as Klassen, the classification is done not by patterns and also the characteristics of the data. The results of the regional development level grouping studies using K-means clustering indicates the level of accuracy of 80%, where only

three of the fifteen counties that do not fit the current classification compared between Klassen and K-means clustering. The third district is Banyumas, Demak and Kudus.

VII. REFERENCES

[1] Kutscherauer, A., Fachinelli, H., Hučka, M., Skokan, K., Sucháček, J., Tománek, P. and Tulej, P. 2010. Regional Disparities : Disparities in country regional development - concept, theory, identification and assessment. Faculty of Economics. VŠB-Technical University of Ostrava.

[2] Yunisti, Trias D. 2012. Analisis Ketimpangan Pembangunan Antar Kabupaten/Kota Di Provinsi Banten, Tesis Program Magister Perencanaan dan Kebijakan Publik Fakultas Ekonomi Universitas Indonesia, Jakarta

[3] Dhyatmika, Ketut Wahyu. 2013. Analisis Ketimpangan Pembangunan Provinsi Banten Pasca Pemekaran, Skripsi Fakultas Ekonomi dan Bisnis Universitas Diponegori, Semarang.

[4] Witten H, Ian and Frank, Eibe. 2005. Data Mining : Practical Machine Learning Tools and Techniques, United Kingdom (UK) : ELSEVIER

[5] Sharma, N., Bajpai, Aman and Litoriya, R. 2012. Comparison the various clustering algorithms of weka tools. International Journal of Emerging Technology and Advanced Engineering. Volume 2, Issue 5, May 2012

[6] Kumar, Parvesh and Wasan, Krishan, S. 2011. Comparative Study of K-Means, Pam and Rough K-Means Algorithms Using Cancer Datasets. Proceeding of 2009 International Symposium on Computing, Communication, and Control (ISCCC 2009).

[7] Murray, G. C., Lin, J. and Chowdhury, A. 2006. Identification of User Sessions with Hierarchical Agglomerative Clustering. Proceedings of the 2006 Annual Meeting of the American Society for Information Science and Technology (ASIST 2006), November 2006, Austin, Texas.

[8] Patidar, G., Singh,Anju and Singh, D. 2013. An Approach for Document Clustering using Agglomerative Clustering and Hebbian-type Neural Network. International Journal of Computer Applications (0975 – 8887). Volume 75– No.9, August 2013.

[9] Souza, K.J. F., Araujo, Arnaldo de A. and Cousty, J. 2011. A simple hierarchical clustering method for improving flame pixel classification. Proceeding of 2011 23rd IEEE International Conference on Tools with Artificial Intelligence.

[10] Jain, A.K., Murty, M.N. and Flynn, P.J. 1999. Data Clustering: A Review. ACM Computing Surveys, Vol. 31, No. 3, September 1999

[11] Soares, J.O., Marques, M.M.L and Monteiro, C.M.F. 2003. A Multivariate Methodology To Uncover Regional Disparities: A Contribution To Improve European Union And Governmental Decisions. European Journal of Operational Research 145 (2003) 121–135

[12] Bakaric, I.R. 2005. Uncovering Regional Disparities – the Use of Factor and Cluster Analysis. Economic Trends and Economic Policy. No. 105, 2005, pp. 52-77

[13] Lukovics, M. 2009. Measuring Regional Disparities on Competitiveness Basis. JATEPress, Szeged, pp. 39-53

[14] Kronthaler, F. 2003. A Study of the Competitiveness of Regions based on a Cluster Analysis: The Example of East Germany. Laporan Penelitian Institute for Economic Research Halle (IWH)

[15] Vydrová, H. V., and Novotná, Z. 2012. Evaluation Of Disparities In Living Standards Of Regions Of The Czech Republic. Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis. Volume LX 42 Number 4, 2012

[16] Nosova, O. 2013. The Innovation Development in Ukraine: Problems and Development Perspectives. International Journal Of Innovation And Business Strategy. Vol. 02/August 2013

[17] Poledníková, E. 2014. Regional classification: The case of the Visegrad Four. Ekonomická revue – Central European Review of Economic Issues. Volume 14: 25–37 (2014)

[18] del Campo, C., Monteiro, Carlos M. F., and Soares, J. O., The European regional policy and the socioeconomic diversity of European regions: A multivariate analysis. European Journal of Operational Research 187(2).

[19] Ramzan, Shahla., Khan, M.I, Zahid F.M and Ramzan, S. 2013. Regional Development Assessment Based on Socioeconomic Factors in Pakistan Using Cluster Analysis. World Applied Sciences Journal 21 (2): 284-292

[20] Spicka, J. 2013. The Economic Disparity in European Agriculture in the Context of the Recent EU Enlargements. Journal of Economics and Sustainable Development. Vol.4, No.15, 2013.

[21] Jaba, E., Ionescu, A.M., Iatu, Corneliu and Balan, C.B. 2009. The Evaluation Of The Regional Profile Of The Economic Development In Romania. Analele Ştiintifice Ale Universităţii„ Alexandru Ioan Cuza" Din Iaşi. Tomul LVI ŞtiinŃe Economice 2009.