Advances in Social Science, Education and Humanities Research, volume 474
Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)

Application of Clustering Algorithm and Spatial Analysis for Industrial Optimization

Achmad Fauzan[1,*], Ginanjar Wiro Sasmito[2], Sekti Kartika Dini[1]

[1]*Universitas Islam Indonesia, Indonesia.*
[2]*Politeknik Harapan Bersama, Indonesia*
**Corresponding author: achmadfauzan@uii.ac.id*

ABSTRACT
Technological advances and increasingly diverse human needs, making one of the challenges for policymakers in planning future design. One of them is the determination of policies in industrial equality. The even distribution of the industry itself is in line with the objectives of the country's first field of Development Goals, namely, to end poverty in all forms. Based on this, this study aims to analyze industrial clusterings and continue with spatial analysis of each region and its visualization with the R programming language. The case study of this uses data from the Department of Industry and Trade in Tegal City; this is because Tegal City is one of the tourist destinations and industrial centers that are developing, especially in the field of food or drink in Central Java Province. K-Means cluster method is used in clustering, as well as spatial autocorrelation, to determine whether there is influence from each region. Based on the research results, obtained two optimal groups in the food industry grouping. The determination of the optimal number of groups is based on the Within-Cluster-Sum of Squared Errors (WSS) and Silhouette evaluation methods. The two food industry producer groups formed have the characteristics of the first group consisting of twenty-three food industry producers with average investment value and production value relatively lower compared to the second group consisting of two food industry producers. Moran's index is used to check spatial autocorrelation where the results obtained will be visualized in the form of a geographic information system which is expected to facilitate future policy makers.
Keywords: *K-Means Clustering, Silhouette, Spatial Autocorrelation.*

1. INTRODUCTION

Industrial equalization of each region is needed every year to improve the welfare of the people in the area. Industrial equalization is basically in line with the first objective of the Indonesian Sustainable Development Goals (SGDs), namely, to end poverty in all forms of any kind [1]. One of the cities that is undergoing development in the field of industry in Central Java Province is Tegal City. Tegal City is located in the northern part of the island of Java, which is one of the main routes in transportation and trade. In addition, Tegal City is also one of the tourist destinations and industrial centers that are developing, especially food and beverages.

Due to the growing development of the industry, it needs to be supported by the existence of a strong system, especially related to spatial data in Tegal City. This is done with the aim of creating even industrial distribution so that later it will not overlap or be able to level the development in each location. Optimal industrial equalization can later help policymakers to determine the rules, facilities, and infrastructure as well as other media to be able to support the SGDs and minimize the risk of potential impacts.

Based on this, stages need to be designed to analyze the distribution of each industry in Tegal City. In the initial stage, cluster analysis is used based on data from the Department of Industry and Trade (DISPERINDAG) of Tegal City. Cluster analysis is used here to determine which areas have high, medium, or low levels so that it makes it easy to prioritize future policies, in accordance with the Long Term Development Plan (RPJMD) of Tegal city 2005-2025 [2]. From the cluster analysis followed by spatial analysis, this is because the data held has spatial dimensions. Furthermore, the spatial analysis will be visualized later to make it easier to determine areas that have high or low clusters. In the future, it can also be developed with the presence of online-based publication media, as an illustration of the use of online systems in mapping in Tegal City [3].

2. MATERIALS AND METHODS

2.1. Data Source

The data used in the study was data of small and medium food industries in Tegal City obtained from DISPERINDAG Tegal City in 2016. The industrial data that is owned is spatial data (district level) as many as 23 districts covering the types of industrial centers, the number

of workers, investment value, production value, types of products for one (1) year.

2.2. Research Variables

The variables in the research are as follows:

1. Value of investment
   Investment or capital can also be called the initial costs incurred by a business unit or industry to carry out production activities. The amount of investment is an important factor for the running of a business. With the investment, it is expected that an industrial unit will continue to grow and make a profit. In this study, the value of an investment is expressed in units of the rupiah.
2. Value of production
   Production value is the level of production or the total amount of goods produced by an industry. The size of the production value can be an indication of the size of the industry. In this study, the value of production is expressed in rupiah units

2.3. Synthesis

The research method used in this study broadly consisted of three (3) stages, namely (1) K-Menas Cluster Analysis, (2) Spatial Analysis, (3) Visualization of results, visualized the following flowchart.

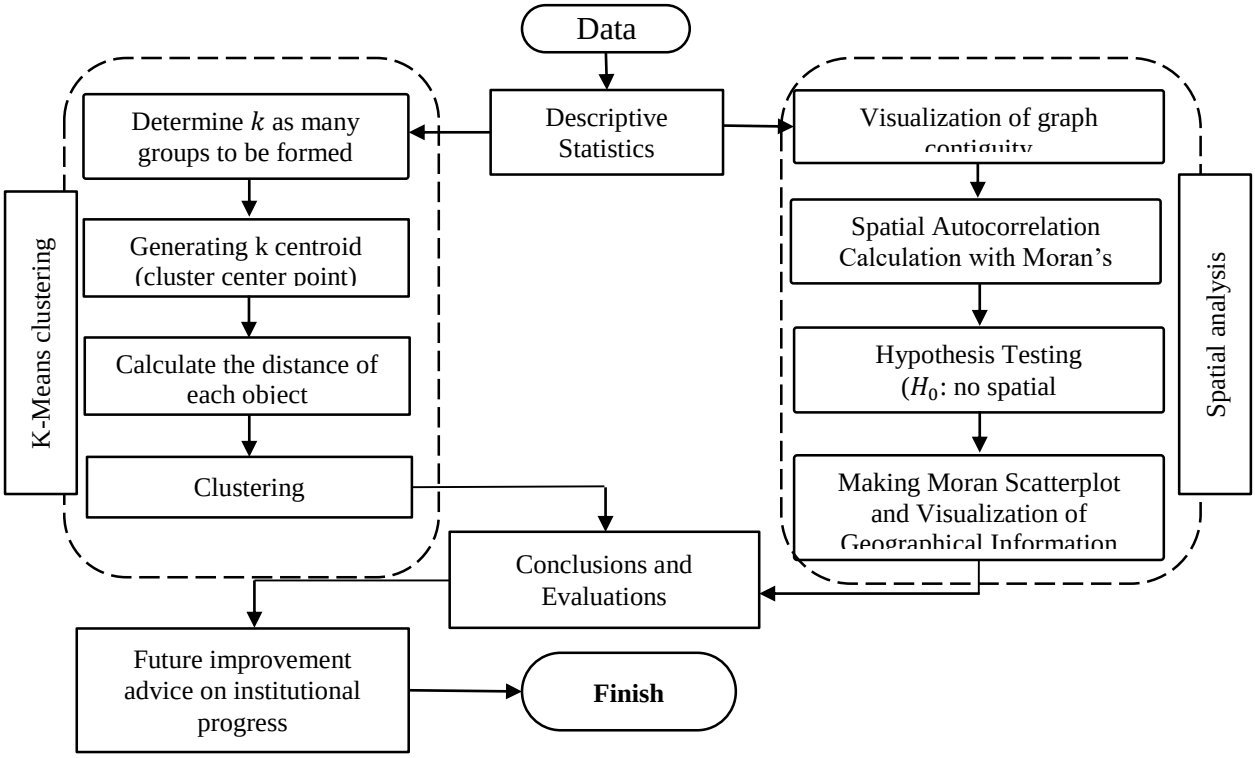

Figure 1. Research Flowchart

2.4. Cluster Analysis

Cluster Analysis is a multivariate statistical analysis used to cluster a group of objects based on the similarity of the characteristics of the object [4]. The aim of clustering is to group objects that have high similarity, while objects that are in different groups will have high dissimilarities. If there are n objects and p variables, then observations with $i = 1,2,3,\ldots,n$ and $j = 1,2,3,\ldots,p$ can be illustrated in Table 1 [5].

Table 1. Illustration of an arrangement of observations in cluster analysis.

	Variable 1	Variable 2	⋯	Variable p
Object 1	X_{11}	X_{12}	⋯	X_{1p}
Object 2	X_{21}	X_{22}	⋱	X_{2p}
⋮	⋮	⋮		⋮
Object n	X_{n1}	X_{n2}	⋯	X_{np}

The measure used in cluster analysis is a measure of similarity, which in this case, is approximated by distance measurement. The most commonly used measure of distance is the Euclidean Distance, written in Equation 1 [4].

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \qquad (1)$$

with d_{ij} : euclidean distance of the i-th data object and the j-th data object, p : number of variables, x_{ik} : i-data object in the k-th variable, and x_{jk} : j-data object in the k-th variable.

2.5. K-Means Cluster Method

K-Means cluster method is one of the non-hierarchical cluster methods used with the aim of dividing existing data into one or more clusters [6]. The algorithm used in the K-Means cluster method is as follows [6] and [7].

a. Set k as the number of groups to be formed
b. Generates a k centroid (cluster center point) randomly based on available objects as many as cluster k. Next, to calculate the next i-centroid cluster, used Equation 2.

$$v = \frac{\sum_{i=1}^{n} x_i}{n} \qquad n = 1,2,3,\ldots,n \quad (2)$$

v: centroid in the cluster, x_i: i-th object, and n: the number of objects that are members of the cluster.
c. Calculate the distance of each object to each centroid in each cluster using Euclidean distance (Equation 1).
d. Group each data according to the closest distance to the centroid.
e. Determine the new centroid position (C_k) by calculating the average value of objects in the same centroid, written in Equation 3.

$$C_k = \left(\frac{1}{n_k}\right)\sum d_i \qquad (3)$$

n_k : number of members in cluster k and d_i : member in cluster k.

The assumption that needs to be done for K-Means cluster analysis is the absence of multicollinearity on the research variables. Multicollinearity is a linear relationship between research variables. To find out whether there is multicollinearity among research variables based on the VIF value. If the VIF value of the research variable is less than 10, then there are no symptoms of multicollinearity on the research variable.

2.6. Spatial Analysis

Before conducting spatial analysis, it can be seen in graph contiguity. Graph contiguity (picture of neighborliness) is a visualization of the neighborliness of each region. This study focused on the food industry in the city of Tegal. Generally, graph contiguity is used as the basis for making spatial weighting matrices denoted by **W** of size $n \times n$ where n represents the amount of data (area) studied. This study uses the Queen Contiguity matrix. The Queen Contiguity matrix is a side-angle intersection matrix that defines the value of $w_{ij} = 1$ if the common side or vertices meet with other regions, while if there is no intersection of the edges then $w_{ij} = 0$. w_{ij} is the value in the i-th row matrix and the j-th column [8].

After the Queen Contiguity matrix is obtained, it is continued with the standardization for each value obtained so that later the number of weights in each row is equal to one. Calculations for standardizing the Queen Contiguity weighting are written in Equation 4 [9].

$$w_{ij} = \frac{c_{ij}}{c_i} \qquad (4)$$

w_{ij} : matrix weight value between the i-th location and j-th row / i-th column (standardized), c_{ij} = value in the i-th row in j-column. Standardized Queen Contiguity weighting matrices are used in testing and modeling methods in spatial analysis.

The spatial analysis used begins with checking the presence or absence of spatial autocorrelation. Spatial autocorrelation is one of the spatial analyzes to determine the pattern of relationships or correlations between locations (observations). The Moran Index (Moran's I) is one method to calculate spatial

autocorrelation. Moran's I can be used to find the outset of spatial randomness. This spatial randomness can represent the presence of patterns that cluster or form trends in space [10]. Calculation of Moran's I with a weighting matrix in the form of normality or a standardized matrix can be calculated with Equation 5 [11].

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left((y_i - \bar{y})(y_j - \bar{y})\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \quad (5)$$

n: number of observations (location), y_i : observation value at the i-th location, y_j : observation value at the j-th location, $\bar{y}$: the average value at n locations, w_{ij} : matrix weight value between i-th and j-location (standardized). Moran's I coefficient is used to testing spatial dependencies or autocorrelation between observations or locations. Pattern identification uses Moran's I index value criteria, if the value of $I > E[I]$ then has a clustering pattern, if $I < E[I]$, then it has a spread pattern, if $I = E[I]$ then it has an uneven spread pattern [12]. E [I] is the expectation of I formulated in Equation 1.

$$E[I] = \frac{-1}{(n-1)} \quad (6)$$

The null hypothesis for the Moran Index is that there is no spatial autocorrelation. The test statistic used can be seen in Equation 7 through Equation 9.

$$Z(I) = \frac{I - E[I]}{\sqrt{Var\ (I)}} \approx N(0,1) \quad (7)$$

$$Var\ (I) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n^2-1)S_0^2} - [E(I)]^2 \quad (8)$$

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,, S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij} + w_{ji}\right)^2\,, S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2 \quad (9)$$

H_0 is rejected or there is autocorrelation [13], [14] between locations if $|\ Z_{count}\ | > Z_{\left(\frac{\alpha}{2}\right)}$, $Z_{\left(\frac{\alpha}{2}\right)}$ is 1.96 or p_{value} <5%.

3. RESULTS AND DISCUSSION

3.1. Data Exploration

Descriptive statistics are performed to find out the basic description of the data held, a statistical summary of the variables, and data size [15]. Descriptive statistics for the research variables are presented in Table 2.

Table 2. Description of data based on investment value and production value variables

Data Summary	Investment Value of IDR (000)	Production Value of IDR (000)
Average	853093	24306789,84
Minimum value	1500	15000
Maximum value	8307950	478626004

At a glance, the data possessed have a fairly large range when viewed from the minimum and maximum values of investment data and production value.

3.2. Cluster Analysis

Prior to the cluster analysis, the multicollinearity assumption was tested, and the VIF value obtained for the investment value variable was 2.309, and the VIF value for the production value variable was 2.309. Because the VIF value for each variable is less than 10, it can be concluded that there is no multicollinearity between variables.

Based on the analysis using the K-Means method, the optimal number of groups is two groups. The determination of the number of groups is based on the Within-Cluster-Sum of Squared Errors (WSS) and Silhouette criteria.

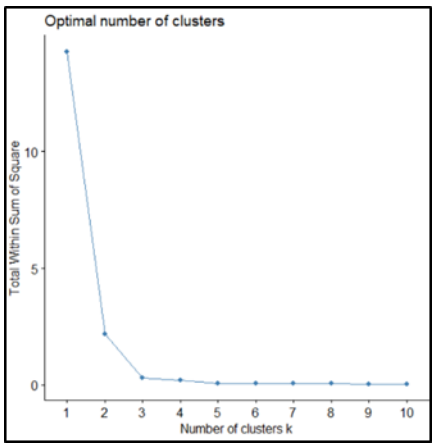

Figure 2. WSS Criteria

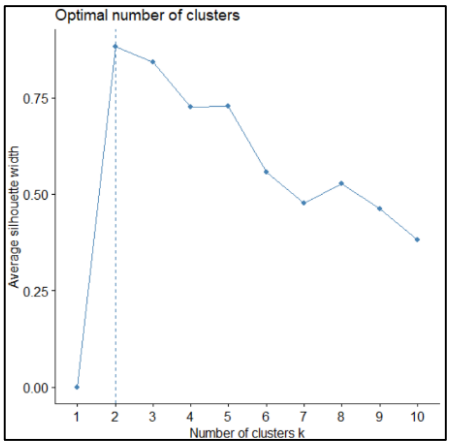

Figure 3. Silhouette Criteria

Obtained cluster characteristics are presented in Table 3.

Table 3. Characterization Summary

Cluster	Number of members	Average Investment Value (IDR 000)	Average Production Value (IDR 000)
1	23	337720,6522	4736119,348
2	2	6779875	249369501

Based on Table 3, it is known that cluster two has characteristics with higher investment value and production value when compared to cluster one. Cluster one members consist of the Village of Kradon, Cebawan, Sumurpanggang, Kalinyamat Kulon, Pesurungan Lor, Margadana, Kaligangsa, Pesurungan Kidul, Pekauman, Kraton, Muara Reja, Bandung, Keturen, Tunon, Kalinyamat Wetan, Debong Kulon, Debong Tengah, Randugunting, Debong Kidul, Kejambon, Slerok, Stage and Muspusuman. Meanwhile, cluster two consists of the villages of Tegal Sari and Mintaragen. Based on the results of the cluster, it can be used as a basis for decision making by the government in order to optimize the industrial area in the city of Tegal

3.3. Spatial Analysis

After clustering using the K-Means cluster and its evaluation, it is followed by spatial analysis to determine whether there are dependencies or spatial dependencies between regions. As an initial illustration, the following is a neighboring region matrix presented in Figure 4. While Figure 5 is an illustration of Moran's Scatterplot. Spatial analysis code can be seen Spatial Data Analysis with R [16], while the basics of basic spatial visualization can be seen in Spatial Data in R [17]

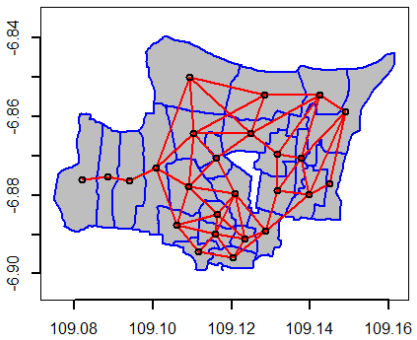

Figure 4. Graph contiguity

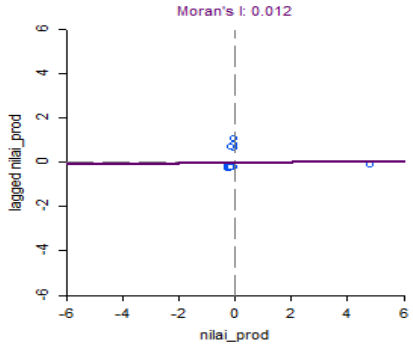

Figure 5. Moran's Scatterplot

In Figure 4, the red line shows that the two regions are neighbors, the neighbor in this neighborhood using the queen type. While in Figure 5 shows Moran's Scatter Plot or distribution of each region into the quadrant (I, II, III, or IV). Based on the Graph Contiguity visualization, it is obtained that each region is neighboring with other regions, with at least one neighbor. Then the spatial weighting matrix is calculated and standardized for the Moran's Index calculation. Based on Equation 5, the morans index value is obtained as follows.

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left((y_i - \bar{y})(y_j - \bar{y})\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} = 0.012$$

Based on the significance test of the normal approach, according to Equation 7, a value is obtained $S_0 = 25$, $S_1 = 13.376$, and $S_2 = 102.02$. Then based on Equation 8, the value of Var (I) is obtained

$$Var\ (I) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n^2 - 1)S_0^2} - [E(I)]^2 = \frac{25^2(13.376) - 25(102.02) + 3(25^2)}{(25^2 - 1)(25^2)} - [-0.042]^2 = 0.0179$$

$$Z(I) = \frac{I - E[I]}{\sqrt{Var\ (I)}} = \frac{0.012 - (-0.042)}{\sqrt{0.0179}} = 0.4$$

Earned value $|Z(I)| = 0.4 < 1.96 = Z_{\frac{\alpha}{2}}$ until it fails to reject the null hypothesis. This is reinforced by value $p_{value} = 0.65 > \alpha$. Based on the Moran's Index test, it was concluded that at a significance level of 5%, there was no spatial autocorrelation of the number of food industries in Tegal City, Central Java. This is also reinforced from Moran's Index value of 0.012, which indicates a positive but small autocorrelation because it is more likely to approach zero than one.

Because the Moran's Index obtained illustrates that there is no spatial dependence, spatial regression modeling cannot be continued. As an alternative, you can use Moran's Scatter Plot. Moran's Scatterplot defined the relationship between the value of observations in an area (standardized) with the average value between observations of neighboring regions with the area concerned [12]. In Scatterplot, divided into four quadrants [18] namely Quadrant I (High-High) shows the location that has a high observation surrounded by a location that has a high observation value as well, with the same analogy Quadrant II is (Low-High), Quadrant III (High-III) Low-Low), and Quadrant IV (High-Low).

Based on Figure 5, obtained the distribution of points only spread to three (3) quadrants, namely quadrants II, III, and IV. In quadrant two (Low-High) obtained four (4) districts, namely Muarareja, Mintaragen, Pesurugan Lor, and Kraton. Most of the locations are located in Quadrant three (Low-low), namely the districts of Panggung, Margadana, Kaligangsa, Krandon, Pekauman, Mangkusuman, Cabawan, Pesurugan Kidul, Slerok, Sumurpanggang, Randugunting, Debong Kulon, Kejambon, Keturen, Kalinyamat Kulon, Central Debong , Tunon, Debong Kidul, Kalinyamat Wetan, and Bandung. In quadrant four (High-Low), there is only one district, namely Tegalsari. Visualization of the results of Moran's Scatterplot and Clustering is presented in Figure 6 and Figure 7.

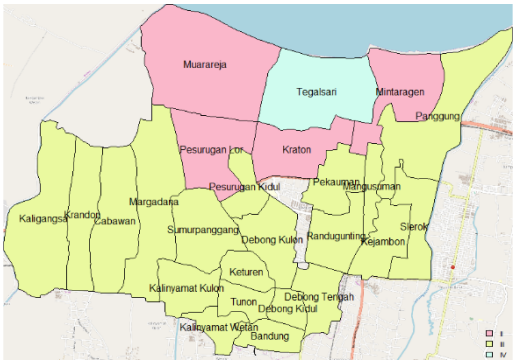

Figure 6. Map of the distribution of the food industry

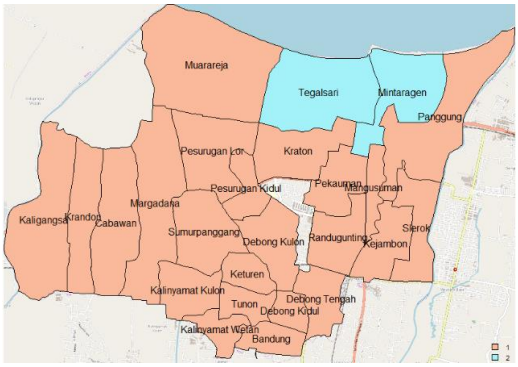

Figure 7. Food Industry Clustering Map

Based on the results of scatter plot moran and overall clustering, regions in Tegal City have almost even distribution in the food industry. There are only a few dominant locations in the food production field. This is because the location is strategic and easily accessible to the surrounding community, thereby increasing in terms of production and in terms of turnover received. In optimizing the food industry, policymakers can arrange futures development starting from the regions in Quadrant II (Low-High) in the hope that it will be easier to develop then Quadrant IV (High-Low) and finally the regions in Quadrant III (Low-Low).

4. CONCLUSION

The conclusions obtained from this study based on the results obtained in the discussion are as follows. Based on K-Means clustering, there are two clusters in the Food Industry in Tegal City. There is no spatial autocorrelation between adjacent locations, but after being analyzed by Moran Scatter, the industry plot is divided into three quadrants, namely Quadrant II (Low-High), Quadrant III (Low-Low), and the last Quadrant IV (High-Low). The regional government can be equal in terms of the policy by

focusing on the regions in quadrant II and then continuing in quadrant IV and, finally, the regions in Quadrant III.

REFERENCES

[1] Widoyono, S. B. e. (2016). Potret Awal Tujuan Pembangunan Berkelanjutan, Jakarta: Badan Pusat Statistik.

[2] P. K. Tegal, "Perpustakaan BAPPENAS," 2014. [Online]. Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=18&cad=rja&uact=8&ved=2ahUKEwjalvTRl-bmAhWd63MBHWnzAiYQFjARegQICBAC&url=http%3A%2F%2Fperpustakaan.bappenas.go.id%2Flontar%2Ffile%3Ffile%3Ddigital%2F194114-[_Konten_]-Konten%2520E2951.pdf&usg=AO.

[3] Sasmito, G. W., Nishom, M. (2018). Mapping of land use system in Tegal city. International Journal of Research in Engineering and Innovation (IJREI). 2(5): 498-502.

[4] Johnson, R. A., Wichern, D. W. (2007). Applied Multivariate Statistical Analysis sixth edition, New Jersey: Pearson Education. Inc.

[5] Soemartini, E., Supartini. (2017). Analisis K-Means Cluster untuk Pengelompokan Kabupaten /Kota Di Jawabarat Berdasarkan Indikator Masyarakat. in Konferensi Nasional Penelitian Matematika dan Pembelajarannya II (KNPMP II). Surakarta.

[6] Bangun, R. H. B. (2016). Analisis Klaster Non-Hierarki dalam Pengelompokan Kabupaten/Kota di Sumatera Utara Berdasarkan Faktor Produksi Padi. JURNAL AGRICA. 9(1): 54-61.

[7] Agusta, Y. (2007). K-Means–Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika. 3(1): 47-60.

[8] LeSage, J. P. (2008). An introduction to spatial econometrics. Revue d'économie industrielle. (123): 19-44.

[9] Kumboro, A. R., Martha, S., Prihandono, B. (2016). Identifikasi Autokorelasi Spasial pada Penyebaran Anak Terlantar di Kabupaten Ketapang dengan Indeks Moran. Buletin Ilmiah Mat. Stat. dan Terapannya (Bimaster).

[10] R. Kosfeld, Spatial Econometric, http://www.scribd.com, 2006.

[11] E. Paradis. (2008). Moran's Autocorrelation Coefficient in Comparative Methods. R De-velopment Core Team. R: A language and environment for statistical computing.R Foundation for Statistical Computing. APE Package vignette.

[12] Lee, J., Wong, D. W. (2001). Statistical analysis with ArcView GIS. John Wiley & Sons.

[13] Zhang, C., McGrath, D. (2004). Geostatistical and GIS Analyses on Soil Organic Carbon Concentrations in Grassland of Southeastern Ireland from Two Different Periods. Geoderma. 119(3-4): 261-275.

[14] Mathur, M. (2015). Spatial autocorrelation analysis in plant population: An overview. Journal of Applied and Natural Science. 7(1): 501-513.

[15] Lani, J. (2010). StatisticsSolutions. [Online]. Available: https://www.statisticssolutions.com/wp-content/uploads/kalins-pdf/singles/descriptive-statistics.pdf. (Accessed 3 January 2020).

[16] Hijmans, R. J., Ghosh, A. (2019). Spatial Data Analysis with R.

[17] Hijmans, R. J. (2019). Spatial Data in R, United States: GFC for the Innovation Lab for Collaborative Research on Sustainable Intensification.

[18] Perobelli, F., Haddad, E. (2003). Brazilian interregional trade (1985-1996): An Exploratory Spatial Data Analysis. ANPEC-AssociaÃ § Ã£ o Nacional dos Centros de PÃ³s-GraduaÃ § Ã£ o em Economia (Brazilian Association of Graduate Programs in Economics).