Application of *K*-Means Clustering in Mapping of Central Java Crime Area

by Retno Tri Vulandari

Submission date: 13-Jan-2020 09:47AM (UTC+0700)
Submission ID: 1630829781
File name: Revisi_IJAS_UNS.docx (369.54K)
Word count: 2921
Character count: 15800

Indonesian Journal of Applied Statistics ISSN 2621–086X
Volume 3 No. 1 May 2020

Application of *K*-Means Clustering in Mapping of Central Java Crime Area

Retno Tri Vulandari[1], Wawan Laksito Yuly Saptomo[2], and Danar Wijaya Aditama[3]
[1, 2, 3] Informatics Engineering, STMIK Sinar Nusantara, Surakarta, Indonesia

[1]retnotv@sinus.ac.id

Abstract. Crimes occur in many places and cause complex problems that have widespread impacts on all levels of society. Crime is related to several factors including crime index, the ratio of the number of police to the population, population density and poverty rates. In this study trying to develop an information system that is able to display and map crime-prone areas in Central Java. Based on these factors, it is used to classify regions in Central Java, namely the category of safe, quite vulnerable, vulnerable and very vulnerable. *K*-Means clustering method, is very suitable to be used in predicting and grouping which areas are included in the 4 categories. The formulation of the problem is to find out areas prone to crime in Central Java. Based on the results, there are 11 regions with safe categories, 4 areas with quite vulnerable categories, 13 regions with vulnerable categories and 6 regions with very vulnerable categories.

Keywords : *K*-Means clustering, mapping, Central Java, criminality, crime area.

1. Introduction

Crime is a complicated problem that has wide impact on all levels of society. Crime is a common problem everywhere. Crimes often occur in various places with different time events, making it difficult to determine which areas have a degree of vulnerability to crime. Information about the number of crimes is needed by the community and law enforcement in this case the police. For the wider community, this information is very useful for anticipatory actions. For the police, this information helps in making decisions about whether an area needs extra supervision or not. In addition, this information is needed to determine the intensity of the crime.

The solution to these problems is how to create a mapping application for crime-prone areas. This application provides information about crime-prone areas in Central Java. The mapping of crime-prone areas is influenced by crime index, the ratio of the number of police to the population, population density and poverty [1]. Some algorithms that can be used to classify crime-prone areas are Hamming *K*-Means, Fuzzy *C*-Means, *K*-Medoid and *K*-Means. Hamming *K*-Means can only be used non-numeric data [2]. Fuzzy *C*-Means is a method of grouping data where the existence of each data in a group is determined by a certain value or degree of membership [3]. *K*-Medoid works effectively for small datasets but does not work well for large datasets [4].

K-Means does not require complicated mathematical operations and *K*-Means minimizes the objective functions set in the clustering process, generally tries to minimize variations within a cluster, the weakness is that the centroids values given at the beginning can affect the results of the clustering if the values are different (sensitive to initial centroids value) [1]. Therefore the mapping of crime-prone areas in the Central Java region is based on the crime index criteria, the ratio of the number of police to the population, population density, and poverty. This study uses the *K*-Means method to cluster crime-prone areas in the Central Java.

2. Literature Review

2.1. Clustering Method. Clustering can be considered the most important unsupervised learning problem, so as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way" [5]. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

For the clustering algorithm to be advantageous and beneficial some of the conditions need to be satisfied. In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering that two or more objects belong to the same cluster if this one defines a concept common to all those objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

Data clustering is a method of data mining that is unsupervised. *K*-Means is a non-hierarchical data clustering method that attempts to partition existing data into one or more clusters/groups. This method partitioned data into clusters/groups so that data that has the same characteristics are grouped into the same cluster and data that has different characteristics are grouped into other groups. Data clustering using the *K*-Means method is generally done with the basic algorithm as follows [6, 7]:

a. Determine the number of clusters.
b. Allocate data into clusters randomly.
c. Calculate the distance of each existing data against each cluster center with the following formula (1):

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (1)$$

d. Allocate each data to the closest centroid/average.

e. Return to step 3, if there is still data that moves the cluster.

If there is a large enough number of data between one variable with another can be difficult in the process of grouping. One solution used to reduce the amount of numbers between variables is to normalize the numbers in the variables using the following equation (2):

$$\text{Normalization value} = \frac{(\text{initial value} - \text{minimum value})}{(\text{maximum value} - \text{minimum value})} \quad (2)$$

2.2. Crime Index. Crime index is the percentage increase or decrease in crime during the year compared to one particular year (which is used as a base year). The higher the crime index of an area indicates the lower level of security in the community of the region. By comparing the 2016 crime rate with the crime rate that occurred in 2014 (as a reference year) [8]. In 2014 the crime rate was relatively high. In that year the circulation of firearms was very high which increased social conflicts in the community such as brawl between villages that occurred in various regions. Conflicts between the TNI-Polri were recorded seven times as well as high rates of drug fraud and drug trafficking in 2014.

The formula for determining the crime index is as follows :

$$I_t = \frac{\text{Number of crimes in the t} - \text{year}}{\text{Number of crimes in the t}_0 - \text{base year}} \text{ x } 100 \quad (3)$$

with t : year and t_o : base year

2.3. Silhouette Coefficient. Silhouette coefficient is a combination of two methods, namely the cohesion method that functions to measure how close the distance between objects in a cluster, and the separation method that functions to measure how far a cluster is separated from another cluster [9]. It can be seen in Table 1.

Table 1. Silhouette Values

Silhouette Values	Structure
$0.7 < SC \leq 1$	Strong Stucture
$0.5 < SC \leq 0.7$	Medium Stucture
$0.25 < SC \leq 0.5$	Weak Stucture
$0 < SC \leq 0.25$	No Stucture

Stages of Silhouette coefficient calculation:

a. Calculate the average distance with all other objects in one cluster with the equation :

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{r=1 \\ r \neq i}}^{m_j} d(x_j^i, x_r^j) \quad (4)$$

$d(x_j^i, x_r^j)$ is the distance of the *i*-th data with the *r*-th data in one *j* cluster, whereas m_j is the number of data in the *j*-th cluster.

b. Calculate the average distance of an object with other objects in another cluster, then take the minimum value with the equation :

$$b_i^j = \min\left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r \neq 1}}^{m_n} d(x_j^i, x_r^n) \right\} \quad (5)$$

c. Calculate the Silhouette coefficient value with the equation :

$$SI_i^j = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \quad (6)$$

The following Table 2 gives literature review on *K*-Means and its modifications.

Table 2. *K*-Means and Its Enhancement

No	Author(s)	Description
1	(Hilman, 2015)	Research on mapping crime-prone areas in the Semarang District Police jurisdiction in 2013. The variables studied were population density, poverty percentage and police sector performance and the number of crime. The results of this study are that several regions in Semarang City have different levels of vulnerability to crime [1].
2	(Handoko, 2014)	Research on web-based geographic information systems for mapping alumni distribution. Variables studied biodata, questionnaire and Polines Commerce Administration majors. The results of this study were obtained that information from 8 alumni of Informatics Engineering study program in Central Java province obtained information on cluster 1 consisting of 1 alumni, cluster 2 consisting of 1 alumni, while for cluster 3 consisted of 6 alumni [10].

No	Author(s)	Description
3	(Astuti, 2016)	Research on mapping street crime in the Semarang City. The research variables are abuse, theft, mugging, robbery, and extortion. The results of this study are the results obtained entropy value for each cluster attribute time events. Clusters/groups 1 are 0.177091, clusters 2 are 0.165072, and clusters 3 are 0.185785 with a total of 0.527947 and an average of 0.175982 [11].
4	(Nengsih, 2016)	Research on web-based Geographic Information Systems (GIS) for land mapping using the classifier model. The purpose of this research is to develop an application in the form of a detection of vacant land based on web-based GIS that provides information on the position of land suitable for construction, the size of available land and access to land to the nearest public facilities [12].
5	(Iswari, 2015)	Research on the use of *K*-means algorithm for mapping the results of traffic accident data clustering. The purpose of this study is to display the results of the analysis as a map showing the grouping of roads along with the status of the level of vulnerability. Cluster status is divided into 3, namely: Not prone, prone, and very prone [13].

3. Results and Discussion

3.1. System Design. Context diagram is the input or output relationship which becomes a unity in a system. In the context diagram, the data is described globally which illustrates the flow of data sourced from the admin user which is then processed in the data processing to produce information such as Figure 1.

Data Flow Diagrams (DFD) are a data logic model or process created to describe where the data originated from and where the data is coming out of the system, where the data is stored, what processes produce the data and the interactions between the data saved and the processes that are subject to in that data. DFD shows the relationship between data in the clustering system using the *K*-Means algorithm can be illustrated in Figure 2.

 ISSN 2621–086X

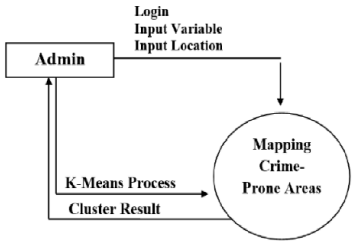

Figure 1. Diagram Context of Mapping Crime-Prone Areas

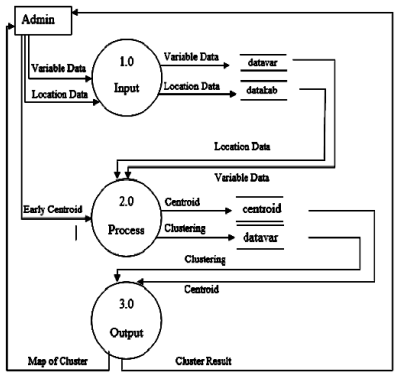

Figure 2. DFD Mapping

3.2. Case Study. Case studies are an appropriate application to provide an understanding of clustering with the *K*-Means algorithm. The following is an overview of the calculation of clustering using the *K*-Means algorithm using district/city data in Central Java. Crime index is the percentage increase or decrease in crime during the year compared to a certain year (which is used as a base year). Obtained a crime index in districts/cities in Central Java using the formula (3). It can be seen in Table 3.

Table 3. Crime Index Data

No	District/city	Crime of 2016	Crime of 2014	Crime Index
1	Cilacap District	283	471	60,085
2	Banyumas District	521	460	113,261
3	Purbalingga District	307	392	78,316
4	Banjarnegara District	259	177	146,328
5	Kebumen District	333	426	78,169
…	...	…	…	…
34	Tegal City	199	368	54,076

District/city data taken in the form of 34 districts/cities in Central Java. Data taken in the form of crime index data, data on the ratio of the number of police to the population, population density and poverty. It can be seen in Table 4.

Table 4. *K*-Means Attribute

No	District/ City	Crime Index	Ratio (Police/ Resident)	Population Density	Percentage of Poverty
1	Cilacap District	60,085	1703	796,531	14,12
2	Banyumas District	113,261	1252	1243,324	17,23
3	Purbalingga District	78,316	1157	1166,986	18,98
4	Banjarnegara District	146,328	1406	848,253	17,46
5	Kebumen District	78,169	1341	926,613	19,86
...	…	…	…	…	…
34	Tegal City	54,076	448	7167,643	8,2

In the clustering process using the *K*-Means method, it will be conducted on 34 districts/cities in Central Java.

a. Determine the initial random centroid

   C1 = (0.3618, 0.4404, 0.0381, 0.4158)

   C2 = (0.3578, 0.4868, 0,0.5274)

   C3 = (0.3017, 0.3746, 0.1122, 0.6129)

   C4 = (0.1611, 0.2244, 0.368, 0)

b. Calculate the shortest distance, it can be seen in Table 5.

Table 5. Cluster Process in the First Iteration

No	C1	C2	C3	C4	Shortest Distance
1	0,3644	0,2894	0,3573	0,8287	C2
2	0,4635	0,3843	0,3924	1,0018	C2
3	0,4871	0,3825	0,3014	0,9908	C3
4	0,6901	0,6314	0,6774	1,2031	C2
5	0,5493	0,4344	0,3840	1,0693	C3
...	…	…	…	...	…
34	0,7225	0,8089	0,7058	0,3356	C3

c. Repeat if there are still changes in the cluster or object. Repeat the above steps until no data has changed.

d. Cluster Results

From the research clustering process that has been manually calculated, that the repetition is carried out until the 4th iteration with the results shown in Table 6. Table 6 is the result of distance calculation in the 4th iteration. Then determine the members of each cluster by selecting the smallest value. To determine the cluster center by counting the number of members in each cluster, divide the number of cluster members to produce a cluster center. The results of these calculations are in Table 7.

Table 6. Cluster Process in the 4th Iteration

No	C1	C2	C3	C4	Shortest Distance
1	0,3726	0,7424	0,2296	0,8918	C3
2	0,4909	0,2795	0,4291	0,9393	C2
3	0,5624	0,6009	0,2380	0,9373	C3
4	0,6799	0,0808	0,7183	1,1350	C2
5	0,6090	0,6047	0,2817	1,0232	C3
…	…	…	…	…	…
34	0,7642	1,1849	0,8368	0,1627	C4

Based on the results of the *K*-Means clustering, it can be concluded that there are 11 regions categorized as safe, 4 areas categorized as quite vulnerable, 13 areas categorized as vulnerable, and 6 regions categorized as highly vulnerable.

Table 7. Cluster Center on the 4th Iteration

C1	0,4109	0,5442	0,0593	0,3552
C2	0,8855	0,5868	0,0441	0,7404
C3	0,2123	0,4970	0,0439	0,7005
C4	0,2674	0,1160	0,5558	0,1784

System implementation, the mapping system by importing data from the database into the application and shown in Figure 3. This page is used to see the results of clustering and the results of the clustering page will be shown in Figure 4.

Black box testing or commonly known as functional testing is a software testing method used to test software without knowing the internal structure of the code or program. Tests carried out on this application system are shown in Table 8.

 ISSN 2621–086X

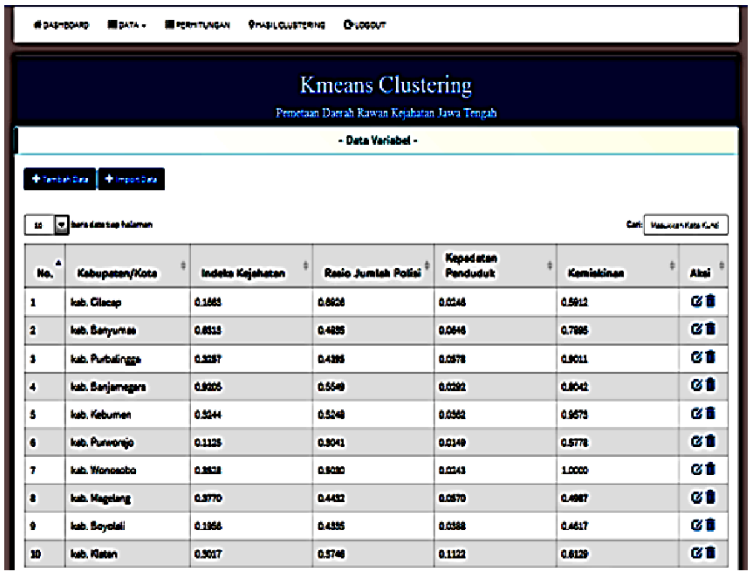

Figure 3. Import Data Page

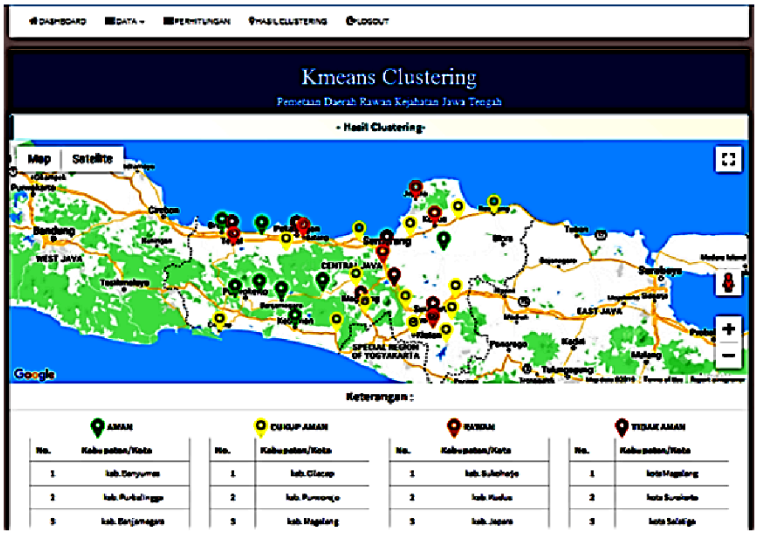

Figure 4. Clustering Results

 ISSN 2621–086X

Table 8. Black-Box Test

No	System test component	Scenario test	Expected results	Test result
1	Login Page	▪ Enter the user and password correctly ▪ Enter the user and password incorrectly	▪ Appear to enter the dashboard menu ▪ Error alerts appear	Received Received
2	Variable Data Page	▪ Add records to the data table in the database ▪ Only .xls file extensions can be imported ▪ Editing existing data ▪ Wipe data	▪ Stored completely in the database ▪ Files with the extension .xls can be imported properly ▪ Data changes based on new input. ▪ The selected data has been successfully erased	Received Received
3	Location Page	▪ Add data to the datacab table in the database ▪ Only .xls extension files can be imported ▪ Editing data ▪ Delete selected data	▪ Stored completely in the database ▪ Files with the extension .xls can be imported properly ▪ Data successfully changes according to the new changes entered ▪ Data has been erased successfully	Received Received Received
4	Cluster Result Page	▪ Accurate results with balanced manual calculation results ▪ Maps can display the location of objects. ▪ The map displays markers by crime category	▪ Accurate results with balanced manual calculation results ▪ Maps can display the location of objects. ▪ The map displays markers by crime category	Received Received Received

Validity testing uses Silhouette coefficient. The results of the Silhouette coefficient test are as follows:

a. Calculate the average distance with all other objects in one cluster with equation (4). It can be seen in Table 9.

b. Calculate the average distance of an object with other objects in another cluster, then take the minimum value with equation (5). It can be seen in Table 10.

 ISSN 2621–086X

Table 9. Calculation of Distances in one cluster

No	Districts/City	$a(i)$
1	Cilacap District	0,365018
2	Banyumas District	0,396732
3	Purbalingga District	0,364933
4	Banjarnegara District	0,25504
5	Kebumen District	0,39216
6	Purworejo District	0,358792
7	Wonosobo District	0,427464
8	Magelang District	0,306219
9	Boyolali District	0,348697
…	…	…
34	Tegal City	0,378542

c. Calculate the Silhouette coefficient value with equation (6).

From the results of Table 11, it can be concluded that the results of the clustering were tested by calculating the silhouette coefficient validation resulting in the value of SC in each cluster, there are 5 medium structure data and 29 weak structure data. It can be seen in Table 11.

Table 10. Calculation of Distance in other clusters

No	Districts/City	$b(i)$
1	Cilacap District	0,651892
2	Banyumas District	0,619805
3	Purbalingga District	0,719931
4	Banjarnegara District	0,833744
5	Kebumen District	0,768948
6	Purworejo District	0,636772
7	Wonosobo District	0,796647
8	Magelang District	0,506456
9	Boyolali District	0,532456
…	…	…
34	Tegal City	0,888726

Table 11. Silhouette Coefficient Calculation

No	Districts/City	$s(i)$	Structure
1	Cilacap District	0,440063	Weak Structure
2	Banyumas District	0,359909	Weak Structure
3	Purbalingga District	0,4931	Weak Structure
4	Banjarnegara District	0,694103	Medium Structure
5	Kebumen District	0,490004	Weak Structure
6	Purworejo District	0,436545	Weak Structure
7	Wonosobo District	0,463421	Weak Structure
8	Magelang District	0,395369	Weak Structure
9	Boyolali District	0,345116	Weak Structure
…	…	…	…
34	Tegal City	0,574062	Medium Structure

4. Conclusions

The regions included in the safe category are Magelang, Sukoharjo, Wonogiri, Pati, Kudus, Jepara, Semarang, Temanggung, Kendal, Batang and Tegal districts. The districts that are included in the enough prone category, namely Banyumas, Banjarnegara, Grobogan and Pemalang. Areas included in the prone category are Cilacap, Purbalingga, Kebumen, Purworejo, Wonosobo, Boyolali, Klaten, Karanganyar, Sragen, Rembang, Demak, Pekalongan and Brebes districts. The regions that are included in the very vulnerable category are Magelang, Surakarta, Salatiga, Semarang, Pekalongan and Tegal. This mapping application uses the *K*-Means method, can map crime-prone areas in Central Java. Users can find out information on the location of crime-prone areas in the system in the form of maps. In the black box testing functionality, the Pamsimas mapping system is free of error syntax and functionally displays the expected results. In testing the validity using silhouette coefficients produce SC values in each cluster, there are 5 medium structure data and 29 weak structure data.

This system can be developed by adding or changing the factors that influence the level of crime vulnerability in the area of Central Java. This system can be developed by providing additional features of map data print out and clustered data as print data output, so that this system can provide data recapitulation.

References

[1] Hilman, G. Y. Pemetaan Daerah Rawan Kriminalitas di Wilayah Hukum Poltabes Semarang tahun 2013 dengan Menggunakan Metode Clustering. *Jurnal Geodesi UNDIP*. 4(1): 32-42. 2015.
[2] Murti, D. H. Clustering Data Non-Numerik dengan Pendekatan Algoritma K-Means dan Hamming Distance Studi Kasus Biro Jodoh. *JUTI: Jurnal Ilmiah Teknologi Informasi.* 4(1): 46-53. 2015.
[3] Hardiyanti, M. Pemetaan Daerah Berpotensi Transmigran di Kecamatan Kartasura dengan Metode Fuzzy C-Means Clustering. *Jurnal Teknologi dan Komunikasi (TiKomSin).* 6(1): 13-20. 2018.
[4] Pramesti, D. F. Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.* 1(9): 723-732. 2017.
[5] Siska, S. T. Analisa dan Penerapan Data Mining untuk Menentukan Kubikasi Air Terjual Berdasarkan Pengelompokan Pelanggan Menggunakan Algoritma K-Means Clustering. *Jurnal Teknologi Informasi dan Pendidikan*. 9(1): 86-93. 2016.
[6] Nasari, F. Penerapan Algoritma K-Means Clustering untuk Pengelompokan Penyebaran Diare di Kabupaten Langkat. *COGITO SMART JOURNAL.* 2(2): 108-119. 2016.
[7] Vulandari, R. T. *Data Mining: Teori dan Aplikasi Rapidminer.* Yogyakarta: Gava Media. 2017.
[8] Latief. Trends Kriminal di Pekanbaru 2012-2016. *Sisi Lain Realita.* 2(1):1-19. 2017.
[9] Prasetyo, E. *Data Mining - Mengolah Data menjadi Informasi Menggunakan Matlab.* Yogyakarta. Andi Publisher. 2014.
[10] Handoko, S. Sistem Informasi Geografis Berbasis Web untuk Pemetaan Sebaran Alumni Menggunakan Metode K-Means. *Jurnal Sistem Informasi Bisnis*. 1(2): 81-86. 2011.
[11] Astuti, W. Pemetaan Tindak Kejahatan Jalanan di Kota Semarang Menggunakan Algoritma K-Means Clustering. *Jurnal Teknik Elektro.* 8(1): 5-7. 2016.
[12] Nengsih, W. GIS Berbasis Web untuk Pemetaan Lahan Menggunakan Classifier Model. *Jurnal Komputer Terapan.* 2(1): 1-6. 2016.
[13] Iswari, L. Pemanfaatan Algoritma K-Means untuk Pemetaan Hasil Klasterisasi Data Kecelakaan Lalu Lintas. *Jurnal Teknologi Industi (TEKNOIN).* 21(1): 1-13. 2015.

Application of K-Means Clustering in Mapping of Central Java Crime Area

ORIGINALITY REPORT

13%	12%	2%
SIMILARITY INDEX	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.tandfonline.com Internet Source	12%
2	Murni Marbun, Muhammad Zarlis, Zulkifli Nasution. "Analysis of Application of the SAW, WP and TOPSIS Methods in Decision Support System Determining Scholarship Recipients at University", Journal of Physics: Conference Series, 2021 Publication	2%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		