Spatial Distribution of Crime in Sumatra and Java Using Unsupervised Learning Algorithm

Yose Indarta[1,a)], Andy Hakim[2,b)], Joli Afriany[3,c)], Imam Ahmad[4,d)], M Mesran[5,e)]
Ronal Watrianthos[6,f)]

[1]*National Police Headquarters of the Republic of Indonesia, Jakarta, Indonesia*
[2]*STAIN Mandailing Natal, Panyabungan, Indonesia*
[3]*Universitas Nahdlatul Ulama Sumatera Utara, Medan, Indonesia*
[4]*Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia*
[5]*Universitas Budi Darma, Medan, Indonesia*
[6]*Universitas Al Washliyah, Rantauprapat, Indonesia*

[a)]yose_11@yahoo.co.id, [b)]andyhakim@stain-madina.ac.id, [c)]joliafriani@gmail.com, [d)]imamahmad666@gmail.com,
[e)]mesran.skom.mkom@gmail.com, [f)]Corresponding author: ronal.watrianthos@gmail.com

Abstract. Indonesia has a low Human Development Index, indicating that the country still has a long way to go in terms of improving its general quality of life and health. Furthermore, Indonesia is afflicted by a number of socioeconomic difficulties, such as overpopulation, poverty, excessive unemployment, and a deficient education system. This issue has the potential to have negative consequences for our society, such as an increase in crime rates. In the subject of criminal statistics, many indicators are commonly used to measure crime from a broader perspective as well as the level of severity. This study aims to characterize the distribution of overall crime rates among Indonesia's provinces, with a special focus on Sumatra and Java. Statistics supplied by the Central Bureau of Statistics for the years 2010-2020 about the number of crimes recorded by regional police officers. The unsupervised learning technique is used, with K-means as the classification algorithm, to categorize objects observed in each state into groups that are related to one another. As a result, the provinces of Bengkulu, the Bangka Belitung Islands, and Banten became the lowest clusters; as a result, it is reasonable to state that these three provinces had the lowest crime rates in contrast to the other provinces in Sumatra and Java between 2010 and 2020.

Keywords: spatial, mapping, crime, Indonesia, K-Means

INTRODUCTION

The Human Development Index for Indonesia is quite low, which indicates that the country still has a long way to go in terms of improving its overall quality of life and health. In addition, Indonesia is plagued by a variety of societal issues, including overpopulation, poverty, high unemployment, and a poor education system. This issue has the potential to have adverse effects on our society, such as an increase in the rate of crime. The Indonesian government does not incorporate social issues that have been recognized as being able to effect crime and instead relies solely on descriptive data for this step of the identification process of social problems and crime[1].

All behaviors that are economically and psychologically destructive to the law, societal standards, and religion are considered criminal. Various community characteristics, including education, population, and economics, might be linked to the prevalence of criminal activities in a given location. Each region in Indonesia must possess unique traits, so that each location's propensity for crime is distinct[2]. According to the calculations of the Central Statistics Agency, there was one act of criminality committed in Indonesia every 1 minute and 32 seconds. Meanwhile, there are 100,000 individuals living in Indonesia, and out of those, there are 140 persons who are vulnerable to being victims of crime. The high crime rate is caused by a number of variables, some of which include a lack of education, lax enforcement of laws, a high unemployment rate, and low earnings[3].

In the field of criminal statistics, there are various indicators that are typically utilized to quantify crime from a more global viewpoint as well as the level of severity. In the broader context, there are indicators that measure the overall number of crimes committed (total crime), the number of crimes committed for every 100,000 people in the population (crime rate), and the amount of time that passes before a crime takes place (crime clock). When analyzing the crime rate in the larger context, extreme caution is required due to the fact that it is an aggregate of all different sorts of crimes that occur at the same time without taking into consideration the level of importance[4].

The most recent information reveals that the pattern observed in the data pertaining to registration tends to be followed by the proportion of the population that was a victim of crime in the year 2019–2020. This pattern shows a decreasing trend. From 1.01 percent of the population in 2019 to 0.78 percent of the population in 2020, the percentage of people who are victims of crime has declined. In the meanwhile, the number of reports to the police each year is still rather modest. In the 2019-2020 period, the percentage of the Indonesian population who had a criminal incidence and subsequently reported it to the police was no more than 25 percent. The proportion was at 23.46 percent in 2020, representing a little rise in comparison to 2019's numbers (22.19 percent)[4] [5].

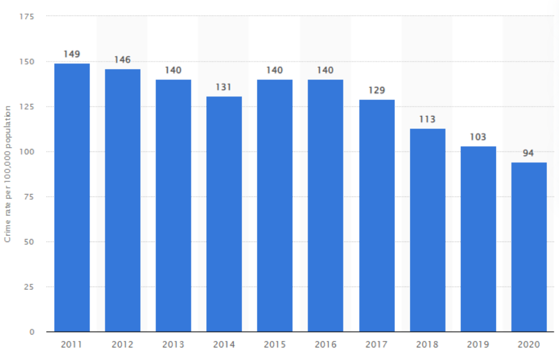

FIGURE 1. Crime Rate in Indonesia from 2011 to 2020 (per 100,000 Population)

Figure 1 shows in the year 2020, the rate of crime in Indonesia stood at 94 for every one hundred thousand residents in the country. Since reaching a seven-year high of 149 crimes committed per one hundred thousand inhabitants in 2011, Indonesia's crime rate has been steadily declining, making the country just moderately secure.

This research endeavors to categorize the dispersion of the total number of crimes that occur across Indonesia's provinces, with a particular focus on Sumatra and Java. Statistics provided by the Central Bureau of Statistics for the years 2010-2020 pertaining to the number of crimes reported by the regional police for the years 2010-2020[6]. The unsupervised learning strategy, with K-means functioning as the classification algorithm, is applied in order to classify things found in each state into groups that are relevant to one another. The core of this methodology is unsupervised learning techniques and the K-means clustering algorithm. These methods have garnered a lot of attention recently. The K-Means method is one illustration of partitioning clustering, which refers to a type of clustering that consists of the process of breaking data down into its individual component parts. K-Means is a well-known method because of how easy it is to put into practice and how quickly it can cluster together vast amounts of data and outliers[7] [8].

In Bangladesh, studies on the spatial distribution of crime were carried out by researchers. In the course of this research, crime mapping and analyses for the spatial distribution of crime are carried out. Mapping criminal activity with a Geographic Information System (GIS) in Bangladesh is just in its infancy at this point. This research also provides an illustration of an examination of the trend of crime over the last three years, from 2016 to 2018, in order to get an understanding of the patterns of crime distribution according to the existing administrative divisions. In addition to this, it focuses on the incidence rate of criminal activity throughout all of Bangladesh's different areas[9].

Using unsupervised learning techniques and data sources collected from the Central Statistics Agency (BPS), another research assessed the number of criminal cases in Indonesia. This study also employed the data mining approach that was used to map. The data that was utilized is comprised of 34 records and is data on the number of criminal activities that were reported to the regional police from 2017-2019. As a direct consequence of this, the six provinces of North Sumatra, South Sumatra, Metro Jaya, West Java, East Java, and South Sulawesi are the ones that have the highest concentrations of criminal cases[10].

The clustering process is carried out with the assistance of RapidMiner, and then, in order to gain spatial visualization, a comprehensive map of all of Indonesia's provinces is constructed with the assistance of Quantum GIS utilizing data obtained during the clustering process[11]. This is done so that a visual comprehension of the number of crimes that occur in Indonesia, particularly those that occur in the provinces that are located on the islands of Java and Sumatra, may be achieved.

METHODS

The K-Means method is one that has been utilized by practically everyone. Multiple strategies for extending K-Means have been proposed in the literature. Unsupervised learning is used by the K-Means algorithm and its variants to conduct clustering for use in pattern recognition and machine learning, however the number of clusters used in the initialization of these algorithms is fixed. That holds true whether or not the training is overseen. Put differently, the K-Means algorithm is not an automated solution for classifying data without human intervention[12][10].

K-Means Algorithm

An example of a fundamental iterative clustering method that is straightforward to grasp is provided by the K-means algorithm. Calculate the distance mean, which will provide the first centroid by using distance itself as the metric and the K classes contained in the data set as inputs. This will result in the calculation of the first centroid. You will now be able to determine the first centroid using this information. As a direct result of this, each class will be differentiated by its very own centroid in its own right. The Euclidean distance has been selected as the method of choice for use as the similarity index for a given data collection X that already contains n multidimensional data points and a category K that needs to be differentiated. This decision was reached after considering a number of potential solutions. Adjustments are made to the clustering objectives in order to limit the total number of possible combinations to the greatest extent that is practically possible while at the same time bringing the sum of squares for each of the distinct sorts down to their absolute minimum[13][14][15].

$$d=\sum_{k=1}^{k}\sum_{i=1}^{n} \| \left(x_i - {}^{u}k \right) \|^2 \qquad (1)$$

where *k* is the total number of cluster centers, u_k is the index of the $_{k}th$ center, and x_i is the index of the *ith* point in the data set, *k* represents the total number of cluster centers. *K* might represent any positive integer that you choose it to represent. The following is one potential solution to the problem with the centroid of the u_k that has been identified:

$$\begin{aligned}\frac{\partial}{\partial u_k} &= \frac{\partial}{\partial u_k}\sum_{k=1}^{k}\sum_{i=1}^{n}\left(x_i - u_k\right)^2 \\ &= \sum_{k=1}^{k}\sum_{i=1}^{n}\frac{\partial}{\partial u_k}\left(x_i - u_k\right)^2 \qquad (2) \\ &= \sum_{i=1}^{n} 2\left(x_i - u_k\right)\end{aligned}$$

The use of an algorithm is predicated on the notion of selecting *k* points at random from the collection of sample points in order for them to function as the center of the first cluster: To analyze the data, first divide each sample point into the cluster whose center point is the closest to the point being analyzed. The average of all of the sample points that are contained inside a given cluster is used to calculate the location of each cluster's center point. Repeat the steps from the preceding section until the point at which the cluster is centered does not shift and the maximum number of repetitions has been reached, whichever comes first. The conclusions reached by the algorithm are susceptible to change depending on the center point that is chosen.

As a consequence of this, the findings cannot be trusted. The value of *k*, which serves as the major focus of the algorithm, is responsible for determining the precise location of the center point. It has an immediate and significant bearing on the outcomes of the clustering, such as determining whether the clusters are local or global[16][15].

RapidMiner

RapidMiner is one of the open-source data mining systems that is the most popular and widely utilized throughout the world. The University of Dortmund was the birthplace of the project in 2001, and Rapid-I GmbH has been responsible for its further development since 2007. RapidMiner continues to cater not just to customers in the commercial world, but also to academic institutions and academics working in a wide variety of fields, thanks to its academic heritage. On the one hand, this category contains computer scientists, statisticians, and mathematicians who are interested in the processes of data mining, machine learning, and statistical approaches. On the other hand, this category does not include anyone. Implementing new types of research and procedures, as well as contrasting them with other options, is made simple and straightforward by RapidMiner[17].

On the other hand, RapidMiner is implemented in a wide variety of application domains, including physics, mechanical engineering, medicine, chemistry, linguistics, and the social sciences. Today, many subfields of research are driven by data, and as a result, versatile analytical tools are required. RapidMiner is a tool that can be put to use in this capacity because it offers a comprehensive set of methodologies, ranging from straightforward statistical evaluations like correlation analysis to more complex procedures like regression, classification, and clustering, as well as dimension reduction and parameter optimization. Analysis of text, images, audio, and time series are only few of the application areas that might benefit from the utilization of these technologies. It is possible to fully automate all of these studies, and the findings may be shown in a variety of different ways[17].

Quantum GIS

A Geographic Information System, or GIS for short, is any system that stores, saves, processes, and shows data that is tied to a specific area. The term "geographic" refers to the data itself, not the location (i.e., are georeferenced). The word "GIS" refers to a wide variety of operations, some examples of which include the administration, processing, and utilization of digital data that is georeferenced. Over the past ten years, Geographic Information Systems (GIS) have experienced significant expansion. This expansion has resulted in a wider and more diverse range of applications across a variety of sectors, including business and government, extensive consumer use, and an increasing role in information systems and business schools. The design of Geographic Information Systems (GIS) falls within the category of qualitative writing, and it employs descriptive research methods. Within the scope of this investigation, the application known as Quantum GIS acts as a tool for the processing and clustering of data on the number of criminal offenses committed in the provinces of Sumatra and Java[18].

Dataset

The Central Statistics Agency of the Republic of Indonesia (abbreviated BPS) is responsible for data collection and analysis. We examined crime data for Sumatra and Java province from 2010 to 2020[6]. As part of its data visualization process, Rapid Miner employs a clustering technique. Table 1 provides both raw data and processed results:

TABLE 1. Number of Crimes According to the Regional Police of Each Province in 2010-2020

Province	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Aceh	9244.00	9114.00	9200.00	9150.00	7569.00	8048.00	9646.00	8885.00	8758.00	7483.00	7745.00
North Sumatera	33227.00	37610.00	33250.00	40709.00	35728.00	35248.00	37102.00	39867.00	32922.00	30831.00	32990.00
West Sumatera	10819.00	11695.00	13468.00	14324.00	14955.00	16277.00	14921.00	13205.00	12953.00	11064.00	7992.00
Riau	10129.00	8323.00	12533.00	9399.00	9644.00	9595.00	8520.00	6869.00	7246.00	6570.00	8194.00
Jambi	3586.00	4450.00	6099.00	6510.00	7643.00	10564.00	9424.00	9531.00	6313.00	6848.00	4709.00
South Sumatera	18288.00	19353.00	21498.00	22882.00	22708.00	20575.00	20368.00	15728.00	13558.00	12861.00	12189.00
Bengkulu	2717.00	3498.00	3943.00	4550.00	3847.00	4463.00	5904.00	4867.00	3389.00	3453.00	3333.00
Lampung	4813.00	6052.00	4383.00	4812.00	7755.00	9218.00	10485.00	11089.00	8963.00	8534.00	7594.00
Bangka Belitung	2642.00	2732.00	5197.00	2515.00	1796.00	1875.00	2094.00	1931.00	2048.00	1953.00	1931.00
Riau Island	4141.00	3643.00	3626.00	4278.00	4633.00	4892.00	4885.00	3673.00	3409.00	3159.00	2843.00
Jakarta	60989.00	53324.00	52642.00	49498.00	44298.00	44461.00	43842.00	34767.00	34655.00	31934.00	26585.00
West Java	16869.00	29296.00	27247.00	24843.00	27058.00	27805.00	29351.00	25183.00	16209.00	13145.00	11256.00
Centre Java	15479.00	15205.00	11079.00	14859.00	15993.00	15958.00	14353.00	12033.00	9127.00	10317.00	10712.00
Yogyakarta	17622.00	6326.00	8987.00	6727.00	7135.00	9692.00	8348.00	7251.00	6731.00	6650.00	7721.00
East Java	16948.00	28392.00	22774.00	16913.00	14102.00	35437.00	28902.00	34598.00	26295.00	26985.00	17642.00
Banten	3832.00	3205.00	3804.00	4259.00	5741.00	5002.00	4570.00	3692.00	3623.00	3287.00	4250.00

RESULTS AND DISCUSSION

In this investigation, the data from Table 1 were analyzed using the K-means clustering method. On the third, fifth, and seventh clusters, the grouping is done with the use of mapping labels. Figure 2 presents an illustration of the mapping procedure performed with Rapid Miner.

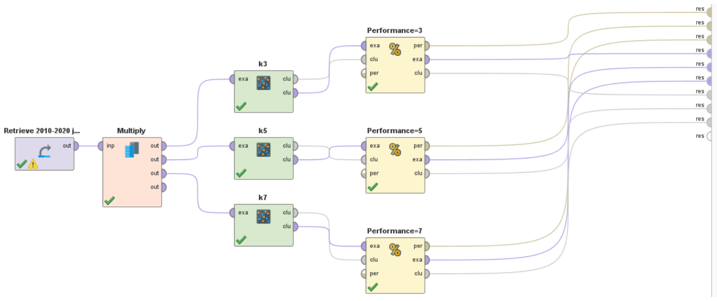

FIGURE 2. Three Different Clusters Were Mapped With the Help of the Rapidminer Modeling Mapping Tool

Figure 2 illustrates how the read Excel method can be used to enter data based on the information that is provided in Table 1. The K-means model is responsible for mapping out the amount of crimes committed in Sumatra and Java as one of its responsibilities. According to what is shown in table 2, there are three clusters that are used to improve performance measures that are based on the Davies Bouldin Index (DBI). This is done so that it is possible to determine which cluster is the most accurate for mapping information technology abilities by province.

TABLE 2. The Third Performance According to the Davies-Bouldin Index

Cluster	DBI
Performance K=3	-0.584
Performance K=5	-0.554
Performance K=7	-0.432

The results of each table's evaluation according to the Davies-Bouldin Index are presented in Table 2. The Davies-Bouldin Index (DBI) is a tool that measures the efficiency of cluster evaluation by making use of the clustering approach. The DBI number should be non-negative and greater than zero. The smaller this value is, the more robust the cluster (k) that the method employed discovers. With a DBI score of -0.432, grouping the data from the 16 provinces of Sumatra and Java into seven clusters proved to be one of the most efficient ways to arrange the data.

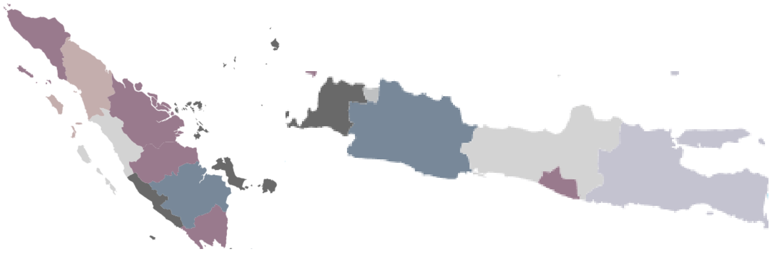

FIGURE 3. Distribution of number crime by cluster in Java and Sumatra, 2010-2020

Seven distinct clusters were created when the data set of Indonesian provinces from Open Map was analyzed with Quantum GIS. This allowed for the creation of the dataset. Figure 3 illustrates the outcomes of the procedure by doing a QGIS visualization on each cluster.

TABLE 3. Color Distribution in Each Cluster Every Province

Cluster	Province	Color
Cluster 0	Bengkulu	⬛
	Bangka Belitung Island	
	Banten	
Cluster 1	South Sumatra	🟦
	West Java	
Cluster 2	Jakarta	⬜
Cluster 3	West Sumatra	⬜
	Centre Java	
Cluster 4	East Java	⬜
Cluster 5	North Sumatra	🟫
Cluster 6	Aceh	🟪
	Riau	
	Jambi	
	Lampung	
	DI Yogyakarta	

According to the information presented in table 3, out of a total of seven clusters, Bengkulu province, the Bangka Belitung islands, and Banten are the clusters with the lowest scores. This demonstrates that when compared to other provinces, this one has the lowest amount of criminal activities that occurred during the years of 2010 and 2020. However, the provinces of Aceh, Riau, Jambi, Lampung, and di Yogyakarta, which had the highest crime rates throughout that time period, have the highest concentrations of violent crime. The fact that various provinces with significant populations, like Jakarta, East Java, and South Sumatra, are located in their own distinct clusters is an intriguing discovery. This demonstrates that the province is home to a number of criminal acts that, when added together, prevent it from being mixed with those of other provinces throughout the period of 2010-2020.

CONCLUSIONS

On the basis of the results of the research, the K-means method can be utilized to map the number of acts of criminality that have taken place in Indonesia, particularly in the regions of Sumatra and Java. In order to make the Davies Bouldin Index (DBI) a more accurate performance measurement tool, three clusters have been introduced. This is done so that it may be determined which cluster gives the mapping by province that is the most accurate. The seventh cluster out of the three has a DBI score that is 0.432 points lower than the other two clusters, making it the superior option. These seven clusters have been processed with QGIS in order to make it possible to view them based on the open map data that is available for each province in Indonesia. As a consequence of this, the province of Bengkulu, the Bangka Belitung Islands, and the province of Banten became the lowest clusters; consequently, it is possible to assert that these three provinces had the lowest crime rates in comparison to the other provinces in Sumatra and Java during the period of 2010-2020. It is anticipated that the progress of this research would result in more methods, particularly in data clustering with a data set that is more comprehensive and covers all of Indonesia's provinces.

REFERENCES

[1] R. A. Pratama, T. Shafira, F. Ardiansyah, and R. F. Hakim, “Characteristics and segmentation of social problems with kohonen self-organizing maps,” *Bulletin of Social Informatics Theory and Application*, vol. 1, no. 1, 2017, doi: 10.31763/businta.v1i1.19.

[2] L. Ghiffari, N. Gusriani, and K. Parmikanti, “Pemetaan Jenis Tindak Kriminal di Indonesia Berdasarkan Karakteristik Wilayah Menggunakan Canonical Correspondence Analysis (CCA),” *Jurnal Statistika dan Aplikasinya*, vol. 5, no. 2, 2021, doi: 10.21009/jsa.05202.

[3] R. Khairani and Y. Ariesa, “Analisis Faktor-Faktor Yang Mempengaruhi Tingkat Kriminalitas Sumatera Utara (Pendekatan Ekonomi),” *Jurnal Kajian Ekonomi dan Kebijakan PUBLIK*, vol. 4, no. 2, 2019.

[4] Margo Yuwono, *STATISTIK KRIMINAL 2021*. Badan Pusat Statistik, 2021.

[5] Statista Research Department, “Crime rate in Indonesia from 2011 to 2020(per 100,000 population),” *statista.com*, Feb. 2022.

[6] BPS, “Jumlah Tindak Pidana Menurut Kepolisian Daerah 2018-2020,” *Badan Pusat Statistik*, 2022.

[7] B. Setio and P. Prasetyaningrum, “PENERAPAN DATA MINING DALAM MENGELOMPOKKAN KUNJUNGAN WISATAWAN DI KOTA YOGYAKARTA MENGGUNAKAN METODE K-MEANS,” *Journal of Computer Science and Technology (JCS-TECH)*, vol. 1, no. 1, 2021, doi: 10.54840/jcstech.v1i1.9.

[8] R. Watrianthos, Ambiyar, Syahril, Fadhilah, A. Dwinggo, and Samala, “A PROMETHEE-GAIA Method-Based Appraisal of Higher Vocational College in Indonesia,” 2021.

[9] Md. B. Khan and Md. I. Talukder, “Spatial Distribution of Crime in Bangladesh: An Analysis,” *Journal of Penal Law and Criminology / Ceza Hukuku ve Kriminoloji Dergisi*, vol. 0, no. 0, 2021, doi: 10.26650/jplc2021-931625.

[10] N. Azis *et al.*, “Mapping study using the unsupervised learning clustering approach,” *IOP Conf Ser Mater Sci Eng*, vol. 1088, no. 1, p. 012005, Feb. 2021, doi: 10.1088/1757-899X/1088/1/012005.

[11] R. Watrianthos, M. Bobbi Kurniawan, Kusmanto, S. Budiman, and B. Ulya, “Mapping of Traffic Accidents in Labuhanbatu Regency using GIS Support,” *J Phys Conf Ser*, vol. 1566, no. 1, 2020, doi: 10.1088/1742-6596/1566/1/012104.

[12] K. P. Sinaga and M. S. Yang, “Unsupervised K-means clustering algorithm,” *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2988796.

[13] M. Nasution, D. Irmayani, R. Watrianthos, S. Suryadi, and I. R. Munthe, “Comparative analysis of data mining using the rought set method with K-means method,” *International Journal of Scientific and Technology Research*, vol. 8, no. 5, 2019.

[14] Samsir *et al.*, “Naives Bayes Algorithm for Twitter Sentiment Analysis,” *J Phys Conf Ser*, vol. 1933, no. 1, p. 012019, Jun. 2021, doi: 10.1088/1742-6596/1933/1/012019.

[15] C. Yuan and H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” *J Multidisciplinary Scientific Journal*, vol. 2, no. 2, pp. 226–235, Jun. 2019, doi: 10.3390/j2020016.

[16] F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo, and A. P. Windarto, “Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country,” *IOP Conf Ser Mater Sci Eng*, vol. 835, no. 1, p. 012058, Apr. 2020, doi: 10.1088/1757-899X/835/1/012058.

[17] S. Land and S. Fischer, *RapidMiner 5: RapidMiner in academic use*. Rapid-I GmbH, 2012.

[18] Muttaqin, M. Zuhril, and A. Irfan, “Perancangan sistem pemetaan dan pendataan populasi penduduk miskin di Kota Banda Aceh menggunakan aplikasi Quantum GIS,” *Informatics and Computer Science*, vol. 5, no. 1, 2019.