

Discovering the Optimal Number of Crime Cluster Using Elbow, Silhouette, Gap Statistics, and NbClust Methods

Noviyanti T. M. Sagala^{1*} and Alexander Agung Santoso Gunawan²

¹Statistics Department, School of Computer Science, Bina Nusantara University
Jln. K. H. Syahdan No. 9, Jakarta Barat 11480, Indonesia

²Computer Science Department, School of Computer Science, Bina Nusantara University
Jln. K. H. Syahdan No. 9, Jakarta Barat 11480, Indonesia

¹noviyanti.sagala@binus.edu; ²aagung@binus.edu

Received: 31st March 2021/ Revised: 24th May 2021/ Accepted: 24th May 2021

How to Cite: Sagala, N. T. M., & Gunawan, A. A. S. (2022). Discovering the Optimal Number of Crime Cluster Using Elbow, Silhouette, Gap Statistics, and NbClust Methods. *ComTech: Computer, Mathematics and Engineering Applications*, 13(1), 1–10. <https://doi.org/10.21512/comtech.v13i1.7270>

Abstract - In recent years, crime has been critical to be analyzed and tracked to identify the trends and associations with crime patterns and activities. Generally, the analysis is conducted to discover the area or location where the crime is high or low by using different clustering methods, including k-means clustering. Even though the k-means algorithm is commonly used in clustering techniques because of its simplicity, convergence speed, and high efficiency, finding the optimal number of clusters is difficult. Determining the correct clusters for crime analysis is critical to enhancing current crime resolution rates, avoiding future incidents, spending less time for new officers, and increasing activity quality. To address the problem of estimating the number of clusters in the crime domain without the interference of humans, the research carried out Elbow, Silhouette, Gap Statistics, and NbClust methods on datasets of Major Crime Indicators (MCI) in 2014–2019. Several stages were performed to process the crime datasets: data understanding, data preparation, cluster modelling, and cluster validation. The first two phases were performed in the R Studio environment and the last two stages in Azure Studio. From the experimental result, Elbow, Silhouette, and NbClust methods suggest a similar number of optimum clusters that is two. After validating the result using the average Silhouette method, the research considers two clusters as the best clusters for the dataset. The visualization result of Silhouette method displays the value of 0,73. Then, the observation of the data is well-grouped. It is placed in the correct group.

Keywords: Crime Clustering, Elbow method, Silhouette method, Gap Statistics method, NbClust method

I. INTRODUCTION

In recent years, crime has been one of the major problems in which the intensity and complexity have grown exponentially. With the increasing number of crimes and the availability of information technologies, crime analysis is required to reduce the probability of crime. The potential crime analysis can give significant data about the crime patterns and find locations where criminal activities are high or low (Prabakaran & Mitra, 2018).

Crime is critical to be analyzed and tracked to identify trends and associations with crime patterns and activities. The appropriate technique must be chosen to solve crime cases briefly due to a large amount of data. Previous research claims that crime analysis may be figured out by a clustering-based model (Hajela, Chawla, & Rasool, 2020). It is required to determine or identify the right crime patterns by conducting the proper clustering approaches. It is essential to identify the right clusters for crime analysis to allow people to examine past crime patterns and improve current crime resolution rates. In addition, using preventive measures based on observed trends take steps to avoid potential accidents. Furthermore, officers who are assigned to a new position with no previous knowledge of site-specific crimes will spend less time in preparation. The last is to increase operating efficiency by redeploying limited resources to the most efficient locations at the most effective times.

Clustering is one of the techniques in data mining which is fundamentally a collection of objects based on closeness and difference between them. In short, clustering is equal to classification. The only difference is that the classes are not defined and

determined in advance, and the grouping of the data is done without supervision (Han, Kamber, & Pei, 2012). Because of its simplicity, convergence speed, and high efficiency, the k-means algorithm is the most popular method in clustering (Berkhin, 2006; Li, Yu, Lei, & Tang, 2017; Jain & Dubes, 1988).

Partitioning clustering, such as k-means clustering, allows users to specify the number of k-clusters to be created. However, determining the optimal number of clusters in a data set is a fundamental problem. Unfortunately, this question has no definitive response. There is no prior information on the correct approach for k. The best possible choice of k is somewhere in the middle ground based on the features of the dataset, such as its size, variance, and characteristics. The optimum number of clusters is somewhat subjective and determined by the method for calculating similarity and the partitioning parameters. Furthermore, it is interesting to note that n of acceptable variety of clusters of specific types of datasets is rarely recognized in practice.

Thus, to solve the problem of determining the number of clusters without human interferences in the crime domain to help crime analysis, the researchers conduct four different approaches that are Elbow, Silhouette coefficient, Gap Statistics, and NbClust methods. Then, those methods are tested on a k-means algorithm to determine the optimal number of crime clusters in the Toronto crime dataset. To the best of researchers' knowledge, this is the first research that works with four different approaches to select the best optimum k-means value on the datasets of Major Crime Indicators (MCI) 2014–2019.

Kingrani, Levene, and Zhang (2018) researched estimating the number of clusters using diversity. They demonstrated that the proposed technique was powerful for groups of diverse sizes, variances, and shapes. Then, Maheswari (2019) proposed a k-means algorithm to build clusters on customers' datasets and validated them using Elbow, Silhouette, and Gap Statistics methods.

Yuan and Yang (2019) implemented k-means on the iris datasets. From the experimental result on small datasets, the four methods, Elbow, Silhouette coefficient, Gap Statistics, and Canopy algorithm methods, met the research requirements. However, the Canopy algorithm was the best choice to work on large and complex datasets.

Nath (2006) proposed the concept of crime detection as a machine learning task by implementing a k-means clustering algorithm on real crime data from a sheriff's office. It implemented analysis, investigation, and discovery of patterns for the occurrence of distinct crimes. Then, Bokde, Kakade, Tumsare, and Wadhai (2018) utilized data mining extensively. They enforced a theoretical model based on clustering and classification of police recorded in England and Wales between 1990 and 2011 to real crime datasets. Joshi, Sabitha, and Choudhury (2017) also used k-means clustering to identify areas

with excessive crime rates and the most frequent type of crimes.

Moreover, Subbalakshmi, Krishna, Rao, and Rao (2015) proposed a Fuzzy Silhouette method to discover an optimal number of clusters. In contrast, Saleh and Khan (2019) analyzed the prediction and visualization of the patterns and trends of different crimes in Chicago by using the k-means algorithm in a Python environment. They discovered that robberies were at their supremacy, and most criminals were not arrested for their crimes.

By considering the result of previous works using the Elbow, Gap Statistics, and Silhouette methods, the researchers propose those approaches together with NbClust on MCI datasets. Since the experiment in an R environment is conducted, the research would also like to know the performance of the R package for clustering, such as NbClust, on the datasets. The results are expected to help to determine which method is the best at finding the optimum number of clusters for the dataset.

II. METHODS

The datasets used in the research are open-source and updated data from Toronto Police Service Public Safety Data Portal (n.d.). The datasets include MCI from 2014 to 2019 occurrences by reported date and related offences. The datasets comprise 206.435 rows and 29 columns. Table 1 displays the description of attributes used. It should be noted that the description of attributes is from other resources, such as Chicago and San Francisco datasets and crime articles, due to no unavailability description in the data portal.

It is proposed that the location of crime incidents is intentionally offset to the nearby highway intersection node to ensure the safety of the parties involved with the case. The location information should be regarded as an estimated location of the event. It is advisable for users not to perceive any of these locations to be directly linked to a particular address or person. In Figure 1, the researchers present the proposed framework for the research.

The framework employed in the research is visualized in Figure 1. The experiment is conducted in an R environment. It starts from data understanding. The data understanding process involves gathering information about where the data comes from, how it is handled, what choices are taken, where it is stored, and how it flows to downstream systems. The process also consists of inspecting the missing values, how they are displayed, and how common they are. Then, the researchers load the data into an R Studio environment as the tool for data preprocessing. Then, deep dive into the data to examine is done to see if it will be necessary to use additional recognized external data sources to strengthen decision-making or conclusion about the clustering models.

Next, the method of cleaning and converting raw data prior to processing and interpretation is known as

Table 1 Attributes Description

No	Name	Description
1	I..X	Unknown
2	Y	Unknown
3	Index	Unknown
4	event_unique_id	Unique identifier for the record
5	Occurrence date	Date when the incident occurs
6	reporteddate	Date when the incident is reported
7	premisetype	Location of incident
8	ucr_code	Uniform Crime Reporting Code
9	ucr_ext	Uniform Crime Reporting Extension (Ext. number to report the incident)
10	offence	Incident subcategory
11	reportedyear	Year when the incident is reported
12	reportedmonth	Month when the incident is reported
13	reportedday	Day when the incident is reported
14	reporteddayofyear	Day of the year of the reported incident
15	reporteddayofweek	Day of week of the reported incident
16	reportedhour	Time of reported incident
17	Occurrenceyear	The year of the incident occurs
18	Occurrencemonth	The month of the incident occurs
19	Occurrence day	The day of the incident occurs
20	Occurrence dayofyear	The day of year incident occurs
21	Occurrence dayofweek	The day of week incident occurs
22	Occurrence hour	Time of incident occurs
23	MCI	Major Crime Indicator/ Incident category
24	Division	Police division code that handles the incident
25	Hood_ID	City of Toronto's neighbourhood identifier
26	Neighborhood	City of Toronto's neighbourhood
27	Lat	The latitude of the location where the incident occurs. Approximate location of the occurrence.
28	Long	The longitude of the location where the incident occurs. This coordinate is shifted from the actual coordinate to protect the privacy of parties involved in the occurrence.
29	ObjectId	Internal feature number

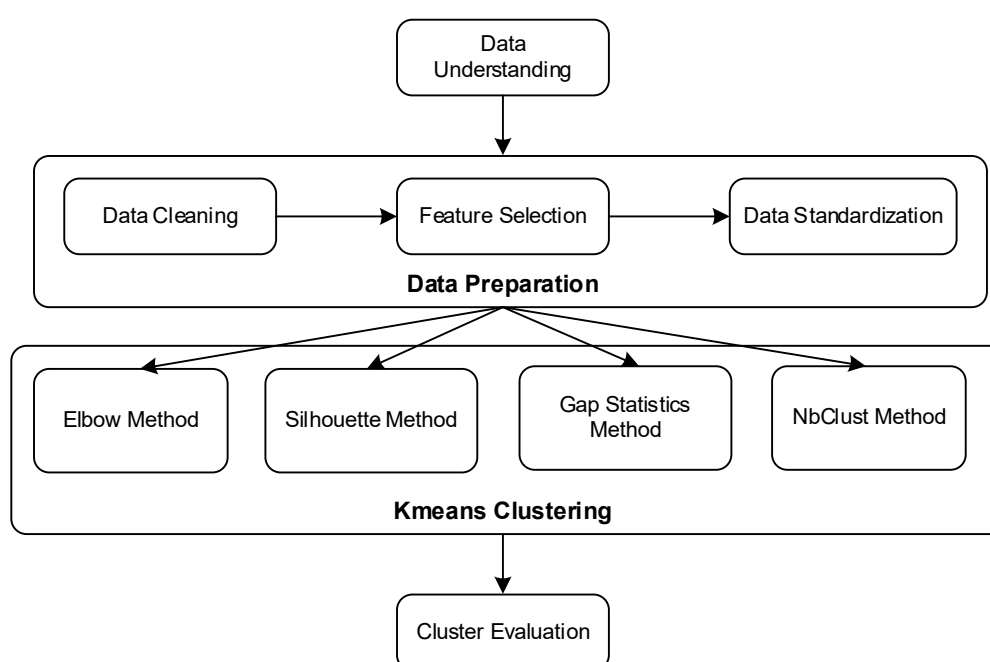


Figure 1 The Proposed Framework

data preparation. It is a critical stage before processing that always includes reformatting data, making data corrections, and merging datasets to enrich data. Data processing may be time-consuming, but transforming data into observations and removing bias caused by low data quality are necessary to bring data into action. Data preparation consists of three phases: data cleaning, feature selection, and data standardization. The result of this phase is acceptable datasets with enhanced quality for the cluster modelling stage.

Similarly, cleaning up the data takes the most time, but it is essential for eliminating inaccurate data and filling in gaps. The task commences by removing or modifying incomplete, incorrect, duplicated, or irrelevant data. Data must be checked after it has been cleansed by searching for errors in the data preparation phase up to this stage. Sometimes, an error in the method is found during this stage and needs to be corrected before continuing. Then, the process is followed by the feature selection phase. The objective of selecting features is to remove non-informative or redundant features' contribution to the target variable or output to achieve better performance for the clustering model. The data features to train machine learning models to have a significant impact on the achieved results. The presence of irrelevant features in the datasets will reduce model accuracy and cause the built model to learn based on irrelevant features. For the research, features are selected by specifying the unimportant features and implementing the Boruta package. The researchers check the description and format of the features. Then, if there are any unimportant features, the researchers will rid them. After that, the next phase is done by running the Boruta package in R Studio to remove the remaining meaningless features completely. Boruta is a wrapper built around the classification method of random forests. It manages to capture all the crucial, unique features that the datasets have about an outcome variable.

The significant problem is that the number of variables can be very different. When the original scale is used, the variables with a wide range receive more weight. Hence, the research uses the feature rescaling technique to scale independent variables or data features during the data preprocessing stage to address this problem. Feature scaling or standardization aims to ensure that features are on a nearly equal scale, each feature is equally critical, and it is easier for most clustering algorithms to process. The model built in this work is distance-based algorithms that measure similarities between findings. Therefore, features with wider ranges will have a greater impact on clustering. Variables that are calculated at various scales will contribute to the analysis unequally. The research utilizes the Z-score normalization technique to standardize or rescale the values. It ensures the mean and the standard deviation to be 0 and 1, respectively. The formula of Z-scores is described in Equation (1). It shows χ_i as raw score, μ as mean, and σ as standard deviation.

$$z = \frac{X_i - \mu}{\sigma} \quad (1)$$

A k-means algorithm is conducted on cleaned data in the cluster modelling phase. K-means clustering algorithm is adopted, where the clustering process starts from choosing the valid initial cluster centers randomly from datasets. Then, it randomly assigns each data point to a cluster, followed by calculating the distance between the residual data items and the Euclidean Cluster Center (C_i ($1 \leq i \leq k$)). After that, finding the closest cluster centers of C_i to the destination data item and designating the specified data item to the class of cluster i are done. Then, the following steps determine the mean value of each cluster class data item as the new cluster center and repeat the process until no improvements are possible. It indicates that the objective function has converged. The Euclidean distance formula between data items and space cluster centers is shown in Equation (2). Then, the minimum error sum of squares is adopted by k-means as the objective function of f . It is described in Equation (3). It shows k as number of classes for clusters, χ_i as the set of all the points in the sample, and C_i as the cluster center.

$$d(x, C_i) = \sqrt{\sum_{j=1}^m (x_j - C_{ij})^2} \quad (2)$$

$$f = \min \sum_{i=1}^k \sum_{x \in X_i} |d(x, C_i)| \quad (3)$$

The researchers carry out three methods to find the best possible k for the k-means algorithm. It includes a direct method, statistical testing method, and R. Direct approaches include maximizing a criterion, such as the number of squares inside a cluster or the average Silhouette. Elbow and Silhouette methods are the names of the related methods. Then, the use of statistical techniques involves comparing evidence to the null hypothesis. The Gap Statistics is an example. Meanwhile, R offers the NbClust function for determining the best number of clusters. To demonstrate the direct and statistical methods in the R environment, the researchers need to install a package of factoextra.

The underlying principle of the Elbow rule is to use a square of the distance between the points of the sample in each cluster, and the center of the cluster delivers a set of k -values. The Sum of Squared Errors (SSE) is used as a performance indicator. The researchers iterate and compute the SSE over the k -value. Smaller values indicate that they are more convergent in each cluster.

The Silhouette method is first proposed by Kaufman and Rousseeuw (1990). It merges cohesion

and resolution factors. Cohesion is the closeness between the item and the group. Compared to other clusters, this is called separation. The value of the Silhouette, which is in the range of -1 and 1, is achieved by this comparison. The value of the Silhouette near 1 indicates that the item and the group have a strong relationship. If a data group with a high Silhouette value in a model is generated, the model is appropriate and justifiable.

Gap Statistics is an algorithm proposed by Tibshirani, Walther, and Hastie (2001) to determine the number of clusters of unknown classification numbers of datasets. The principal idea of Gap Statistics is to initiate reference metrics that can be acquired using the Monte Carlo sampling method and to compute the Euclidean distance squares between the two measurements in each class (Xiao & Yu, 2007). The clustering outcomes of the constructed reference zero-mean distribution are compared to determine the optimal number of clusters in the datasets.

The NbClust method determines the number of works based on varying combinations of several clusters, distance measures, and clustering algorithms (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). In short, NbClust computes about 30 methods at once to find the optimal number of clusters. The researchers adopt the function is for a single function call. Then, it can continuously generate specific indices and the number of clusters. Furthermore, various results give the best clustering scheme in work. Before using the package, the researchers need to install the package (NbClust). The difference between the four approaches is shown in Table 2.

In unsupervised learning, cluster evaluation, also known as cluster validation, is not as well-developed because of its very nature (Palacio-Niño & Berzal, 2019). It contributes to various assessment methods. The research focuses on internal validation methods that check the quality of the structure of clustering without access to external information. For the research purpose, the researchers also utilize the Silhouette coefficient.

III. RESULTS AND DISCUSSIONS

After checking the dataset, there is a considerable amount of duplication of “unique_event_id” feature after checking the dataset. Therefore, the researchers remove 26.622 event ids and leave 179.813 rows. In addition, there are 48 rows of standard missing values. Then, they are simply removed from the dataset since the number of cases is less than 5% of the data sets. The remains have plenty of records. Next, the total number of cleaned data is 179.765. The feature selection is carried out manually. It has been discovered that several numbers of features have identical descriptions, and some of them are unimportant to the output. Therefore, features of 1 to 6, 8, 9, 11 to 16, 20, 25, and 29 are deleted from the datasets (the number according to Table 1).

Automatically removing unimportant features for clustering tasks by implementing the Boruta function is offered by R. According to Kurasa and Rudnicki (2010), the top two reasons why Boruta is conducted are as follows. First, it considers multi-variable relationships. Second, it improves random forest for variable importance measure, which is a very popular method for variable selection. Boruta works by creating duplication of all independent variables, performing shadow features, combining the original ones with shuffled copies, running a random forest classifier, computing Z-score, finding the maximum Z-score among shadow features, tagging the variables by the level of importance, removing unimportant variables, and repeating all the process until all variables have been tagged into one of the categories (unimportant or important).

Performing Boruta as an automatic feature selector results in the feature of occurrence month. It is considered unimportant to the target variable. Thus, the variable is eliminated. The following phase is applying data standardization. The data summary before and after Z-score standardization is applied in Table 3. From Table 3, the range or interval between the minimum and maximum values of each feature is small and acceptable.

Table 2 Comparison of the Proposed Approaches

Criteria/Approaches		Elbow	Silhouette	Gap Statistics	NbClust Function
Method	Statistical	√			
	Direct		√	√	√
Works		Computing Within-Cluster Sum of Square (WSS)	Computing the average Silhouette of observations	Comparing the total within intracluster	Providing 30 indices
The optimal number of clusters		The smallest value of total WSS measurements	Based on the Silhouette score	The value which clusters compactness on the original data falls the farthest below this reference curve	Varying all combinations of the number of clusters, distance measures, and clustering methods

Table 3 Standardization Using Z-Score

	Assault	Auto Theft	Break and Enter	Robbery	Theft Over
BEFORE					
Min	98,0	14,0	62,0	10,0	7,00
Max	4160,0	1982,0	1461,0	691,0	326,00
AFTER					
Min	-0,8884	-0,7232	-1,1150	-0,9679	-0,72835
Max	5,5197	9,6640	5,3209	4,9959	5,08959

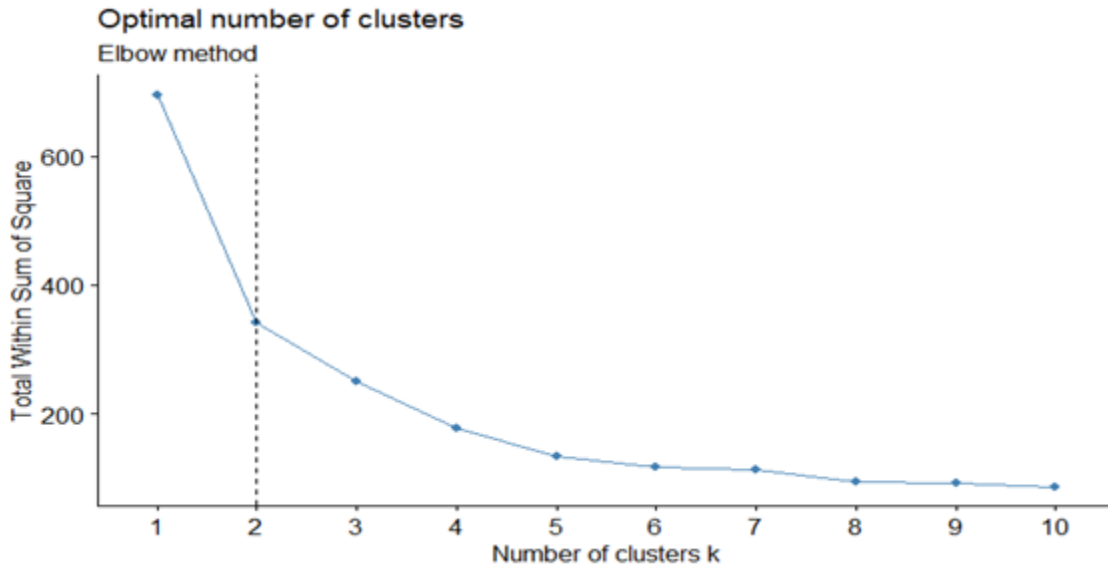


Figure 2 Result of Elbow Method

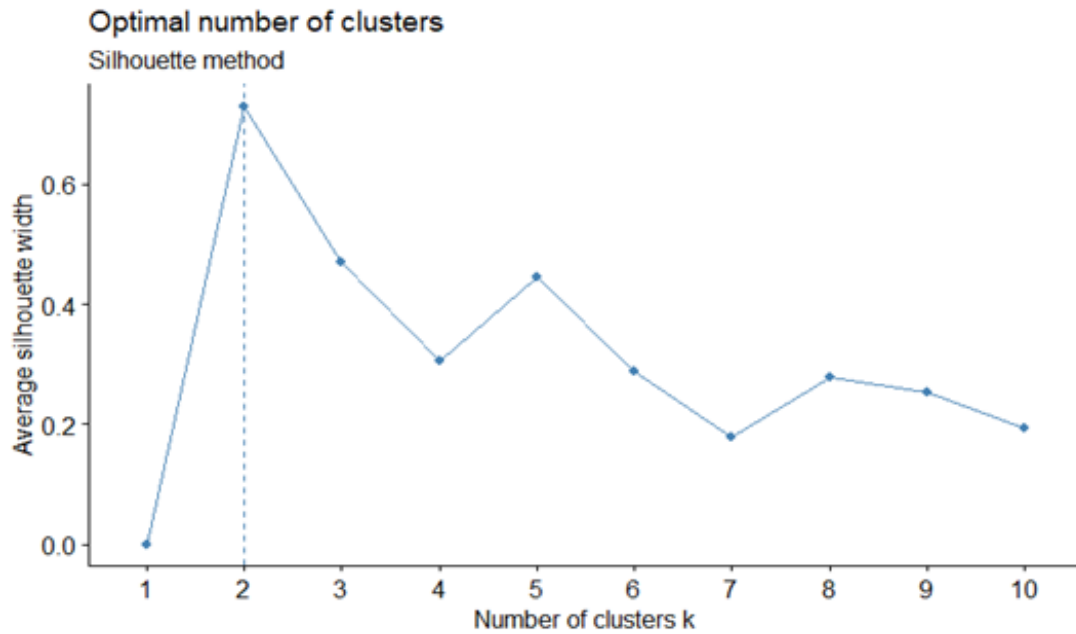


Figure 3 Result of Silhouette Method

The k-value can be determined clearly by plotting the k-SSE curve and finding the inflection point down using the Elbow method. As shown in Figure 2, the location of a knee in the plot is usually considered an indicator of the appropriate number of clusters. It means that adding another cluster does not improve the partition much better. The method seems to suggest two clusters for the dataset. From Figure 2, the larger the number of clusters is, the lower the total WSS. However, there is a slight increase in cluster 7.

The result of the Silhouette method from Figure 3 suggests having two clusters. There are rapid changes from the result of utilizing the Silhouette technique, an upward trend from 1 to 2; 4 to 5; and 7 to 8. However, there is a downward trend from 2 to 4; 5 to 7; and 8 to 10.

By using the Gap Statistics, an optimal number of clusters is the one that makes the best use of it. The method advises clustering the data into three clusters, as visualized in Figure 4. The Gap Statistics coefficient rises significantly to reach a high of 1,9 at cluster 3. Then, it falls to 1 at cluster 4 and steadily increases and reaches the highest coefficient of about 1,12 at cluster 10.

The NbClust function works by providing 30 indices for choosing the best number of clusters. The minimum and maximum numbers of the clusters are 2 and 15, respectively. The function runs a k-means algorithm to check what number of clusters is the best for the dataset. The result is displayed in Figure 5 and Figure 6.

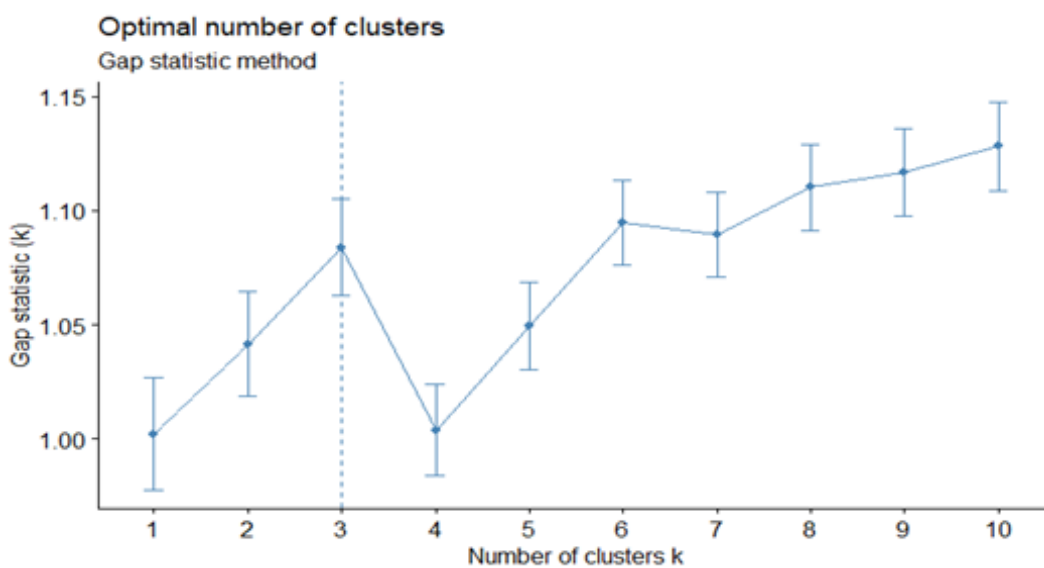


Figure 4 Result of Gap Statistics

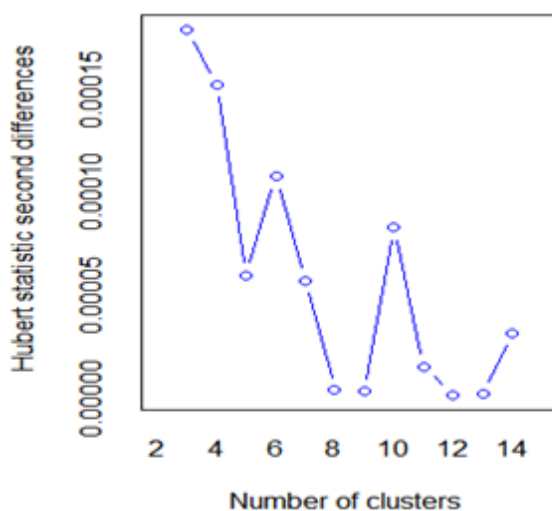


Figure 5 Result of NbClust (Hubert Statistics)

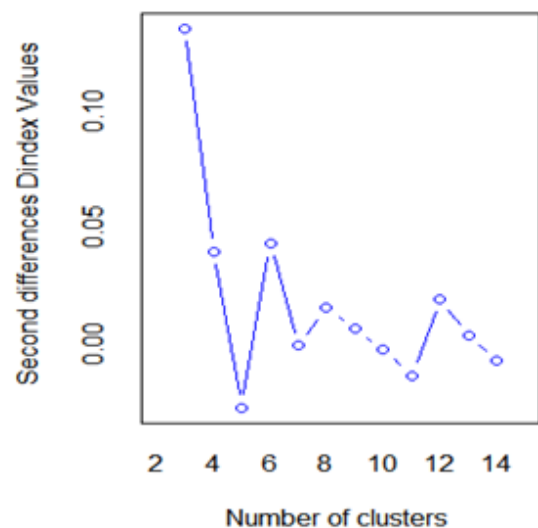


Figure 6 Result of NbClust (Dindex)

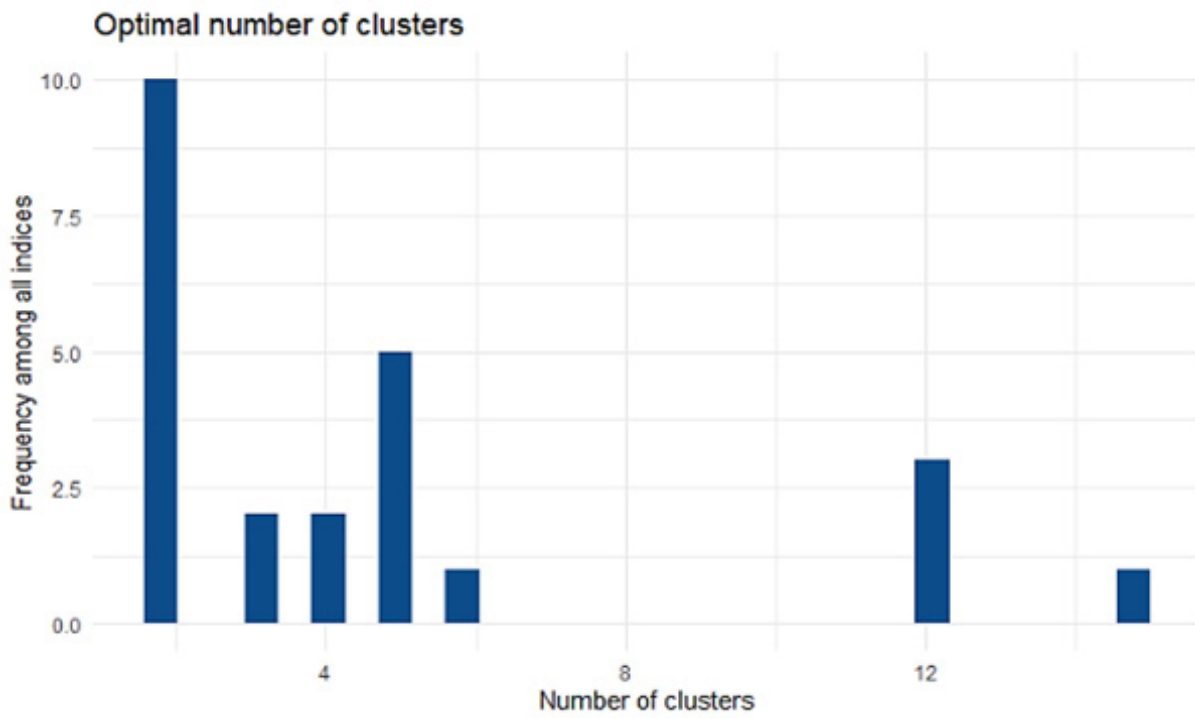


Figure 7 Bar Plot of NbClust Result

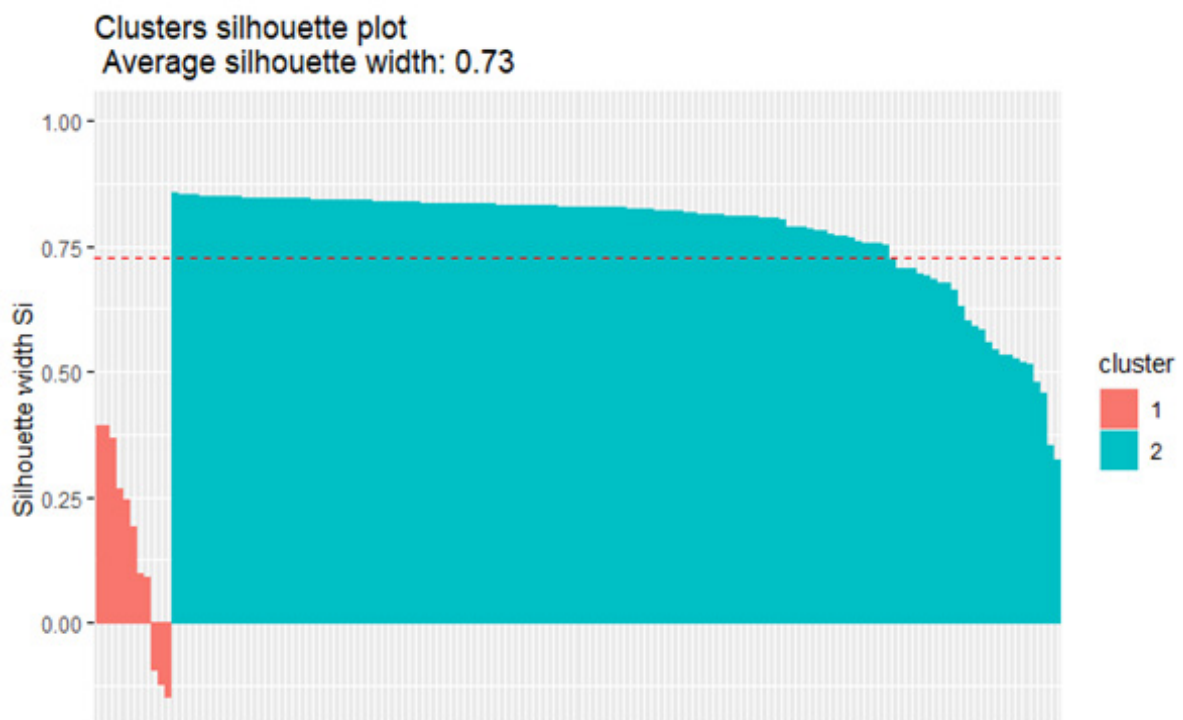


Figure 8 Internal Cluster Validation Using Silhouette Method

The Hubert and Dindex index are a graphical method to determine the number of clusters. In both plots, it seeks a significant knee that corresponds to a significant increase in the value of the measure. The summary of the NbClust is described as follows. Among all indices, 10 frequencies propose 2 as the best number of clusters. Then, 2 frequencies propose 3 as the best number of clusters. On the other hand, 2 frequencies also propose 4 as the best number of clusters, and 5 frequencies propose 5 as the best number of clusters. Then, there are results, such as 1 frequency with 6 as the best number of clusters, 3 frequencies with 12 as the best number of clusters, and 1 frequency with 15 as the best number of clusters. According to the majority rule, the best number of clusters is two. Furthermore, the optimal number of clusters from the NbClust method using a bar plot is visualized in Figure 7. It is clearly presented in Figure 7 that the optimal number of clusters is two.

It is interesting to note that three approaches (Elbow, Silhouette, and NbClust) suggest two as the optimal number of clusters for the crime dataset, while Gap Statistics proposes 3. Gap Statistics works quite different from the remaining approaches because the method has already prepared for its expected values and compared them to the total within intra-cluster for the range of k defined. On the other hand, Elbow, Silhouette, and NbClust work by computing solely the total or average within the intracluster and without the need to come up with their expected values. With the “best” set of relevant features as the result of feature selection phase, uncovering interesting natural groupings (clusters) from data according to the chosen criterion is easier. In other word, relevant features are helpful in finding clusters efficiently.

There is a way to evaluate the clustering quality via the Silhouette plot (which shows the Silhouette coefficient on the y-axis) to confirm that the suggested number of classes is optimal. Then, the researchers draw the Silhouette plot for two clusters, as displayed in Figure 8.

The interpretation of the Silhouette coefficient is as follows. The 0 means that the observation is well grouped. The closer the coefficient is to 1, the better the observation is grouped. Meanwhile, less than 0 means that the observation has been placed in the wrong cluster. If it is equal to 0, the observation is between two clusters. If a large majority of the Silhouette coefficient is positive, the result indicates that the observations are placed in the correct group.

IV. CONCLUSIONS

In recent years, crime has been important to be analyzed and tracked to identify the trends and associations with crime patterns and activities. To address the problem of estimating the number of clusters in the crime, the researchers study Elbow, Silhouette, Gap Statistics, and NbClust methods on datasets of Major Crime Indicators (MCI) in 2014–2019. The primary contribution of the research is the finding of

the best optimal number of clusters for Toronto’s MCI datasets by applying Elbow, Silhouette, and NbClust approaches. Cluster validation is performed to validate the number of recommended clusters. As a result, the best optimal number of clusters for the MCI datasets is two clusters with a 0,73 Silhouette coefficient.

To the best of the researchers’ knowledge, the research literature has not revealed the use of those approaches in this crime dataset previously. By getting the most optimal number of clusters, it may ease the crime analysis. It can focus on these two clusters or the cluster that has large number of criminal cases. The research limitation is that k-mean has trouble with clustering data. The clusters have varying sizes and densities. As shown in cluster validation, most observations belong to cluster 2. Hence, it will also be interesting to compare the approaches to estimate the number of clusters altogether with other clustering algorithms for future research.

REFERENCES

- Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25–71). Springer.
- Bokde, K. A., Kakade, T. P., Tumsare, D. S., & Wadhai, C. G. (2018). Crime detection technique using data mining and K-means. *International Journal of Engineering Research & Technology (IJERT)*, 7(02), 223–226.
- Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6), 1–36.
- Hajela, G., Chawla, M., & Rasool, A. (2020). A clustering based hotspot identification approach for crime prediction. *Procedia Computer Science*, 167, 1462–1470.
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Elsevier.
- Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.
- Joshi, A., Sabitha, A. S., & Choudhury, T. (2017). Crime analysis using K-means clustering. In *2017 3rd International Conference on Computational Intelligence and Networks (CINE)* (pp. 33–39). IEEE.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.
- Kingrani, S. K., Levene, M., & Zhang, D. (2018). Estimating the number of clusters using diversity. *Artificial Intelligence Research*, 7(1), 15–22.
- Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11), 1–13.
- Li, X. Y., Yu, L. Y., Lei, H., & Tang, X. F. (2017). The parallel implementation and application of an improved K-means algorithm. *Journal of University of Electronic Science and Technology of China*, 46(1), 61–68.

- Maheswari, K. (2019). Finding best possible number of clusters using K-means algorithm. *International Journal of Engineering and Advanced Technology*, 9(1S3), 533–538.
- Nath, S. V. (2006). Crime pattern detection using data mining. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops* (pp. 41–44). IEEE.
- Palacio-Niño, J. O., & Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. *ArXiv Preprint ArXiv:1905.05667*.
- Prabakaran, S., & Mitra, S. (2018). Survey of analysis of crime detection techniques using data mining and machine learning. *Journal of Physics: Conference Series*, 1000, 1–10.
- Saleh, M. A., & Khan, I. R. (2019). Crime data analysis in Python using K-means clustering. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 7(IV), 151–155.
- Subbalakshmi, C., Krishna, G. R., Rao, S. K. M., & Rao, P. V. (2015). A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set. *Procedia Computer Science*, 46, 346–353.
- Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411–423.
- Toronto Police Service Public Safety Data Portal. (n.d.) *Major Crime Indicators*. Retrieved from <https://data.torontopolice.on.ca/datasets/TorontoPS::major-crime-indicators-1/about>
- Xiao, Y., & Yu, J. (2007). Gap statistic and K-means algorithm. *J. Comput. Res. Dev*, 44, 176–180.
- Yuan, C., & Yang, H. (2019). Research on K-value selection method of K-means clustering algorithm. *J: Multidisciplinary Scientific Journal*, 2(2), 226–235.