A TWOSTEP CLUSTERING ALGORITHM AS APPLIED TO CRIME DATA OF SOUTH AFRICA

Ntebogang D Moroke*

Abstract

This study applied a TwoStep cluster analysis on the 29 serious crimes reported at 1119 police stations across South Africa for the 2009/2010 financial year. Due to this high number of variables and observations, it becomes difficult to apply some statistical methods without firstly using others as precursors. Classical methods have also been found to be inefficient as they do not have the ability to handle large datasets and mixture of variables. The AIC and BIC automatically identified the three clusters of crimes. The findings may guide authorities when developing interventions tailored to better meet the needs of individual cluster of crimes. Existing plans may also be enhanced to the advantage of residents. More emphasise may be placed on crimes that pose a serious threat. The SAPS may use these findings when reporting on national crime statistics. For future studies, discriminant analysis can be applied to check the clusters' validity**.

Keywords: Data Reduction, Hierarchical Clustering, Information Criteria, Multivariate Analysis, SPSS TwoStep Clustering

JEL Code: B41

**North West University, South Africa, Private Bag X 2046, Mmabatho, RSA, 2735*
Tel: +2718 389 2515, +2782 591 2655
Fax: +2718 389 2597
*** The SAPS is acknowledged as a source of data used in this study.*

1 Introduction

In Social Science research, numerous variables are analysed in a bid to collect empirical evidence. These variables are a collection of metric and non-metric scales. Classical methods used for handling these complex data have several challenges such as those listed below:

- they increase the time needed to capture all constructs,
- they increase the cost of the investigation,
- they make the analysis of the data complex and at times impossible, and hence,
- the large number of variables add another difficult conceptualization layer/level and interpretation level on the normally accepted and understood levels by a common human mind and
- lastly, this may render the whole investigation process difficult or worthless.

Given these challenges, variables with common characteristics may need to be grouped into different clusters. This reduces the complexities and inconsistencies in the data. The display of the results also becomes plausible and sound. Furthermore, the analysis of the results becomes easy and logical. The main focus of this study is to identify and group the 29 common crime variables reported at the 1119 police stations across the nine provinces in South Africa. Manly (2001), recommended that in order to measure crime in a satisfactory manner, different categories of crime need to be classified into groups of similar or comparable offences. He further notes that while classification systems can solve these problems, different jurisdictions perform this classification in different ways.

Considering this recommendation, the current study classifies different crimes according to their similarities or the seriousness perceived or their quantities. Given the number of variables and observations, a modern SPSS TwoStep clustering procedure is explored to achieve the objectives without compromising the analysis. The use of this method allows us to take a different perspective on the data with no preconceived notions regarding profiles, similarities, or performance measures. Consequently, we should through this method be able to reduce the number of variables by collecting them into fewer dissimilar clusters.

The country as a whole may benefit from the study as the findings will be communicated to the South African Police Services (SAPS) authorities and people may be made aware of the seriousness of certain crimes. This may save the SAPS a lot of money and time as more focus will be on a particular cluster of crime. The clusters formed may be reviewed, evaluated and discussed by responsible personnel in the department to better understand the behaviours that link those variables within a cluster, and differentiate them from those in other clusters. Finally, the findings of this study may add to existing literature on the use of this modern clustering method to large data sets especially when both the continuous and categorical variables are concerned. Researchers of crime in the country find it easy to do crime analysis using clusters of crime instead of individual variables.

Several studies have employed clustering procedures in different fields such as those by (Thanassoulis, 1996; Yin, et al., 2007; Leonard & Droege, 2008; Rege, et al., 2008; Kim, et al., 2009; Po, et al., 2009).

2 Theoretical perspectives

This section is devoted to discussing cluster analysis in general and a TwoStep procedure in particular. The next section discusses these methods as they will be used in this study.

2.1 Cluster analysis

Cluster analysis according to Hair et al. (2010) and Fraley and Raftery (1998), is defined as a group of multivariate techniques whose primary purpose is to group objects based on the characteristics they possess. This technique divides a large group of observations into smaller groups based on their similarities or dissimilarities exhibiting a high internal (within a cluster) and external (between clusters) heterogeneity. Lattin et al., (2003) defines this technique as general element in stopping rules which measures the diversity of all the observations across all clusters. When performing cluster analysis, the observations are grouped by taking distances and similarities into consideration (Rencher and Christensen, 2012). As variables are being clustered, the analysis becomes more descriptive than predictive as the main concern is relationships in the data set. Therefore, no condition for linearity of the relationships among variants is assumed (Atlas et al., 2013). Cluster analysis is not dynamic but rather static method used for describing current situation. This method is therefore not convenient in estimation analysis.

Clustering methods are applied when we the intention is to group together naturally in various categories (Schiopu, 2010). The clusters should represent categories of items with many features in common; for example, crimes used as variables in this study. The application of data mining techniques, such as neural networks and decision trees are recommended before clustering the data if the problem is complex. These recommended classical methods are effective and accurate when small data sets are used. They have also been found not to scale up to the very large datasets. The only time these methods will be effective is when these large data are first reduced into smaller datasets. They use the hierarchical or partitioning algorithms. The hierarchical algorithms are known for forming the clusters successively, on the basis of clusters established before.

The advantage of partitioning algorithms is that they determine all the clusters concurrently. In addition, they build different panels and then evaluate them relative to certain criteria (Hair et al., 2010 and Schiopu, 2010). Unfortunately these methods do not offer an effective option or criteria for automatically determining the cluster number. Özdamar (1999) cautions that the available criteria associated with classical methods do not offer precise solutions in obtaining ideal cluster number. Instead they may be used as provision for guide on determining the number of clusters. Based on the nature of the data used and the problems associated with classical clustering methods, this study adopts a modern SPSS TwoStep cluster algorithm to help achieve the objectives. The next section gives a brief description of this procedure.

3 Data and methodology

This section describes the data and methodological procedure used for data analysis.

3.1 Data used in the study

The variables used in this study are the 29 serious crime ratios per 100 000 of the population across the 1119 police stations in the 9 provinces in South Africa. The data covers the period of financial year 2009/2010 totalling 1119 cases during this period. The source of this data is the national office of the SAPS website (www.saps.gov.za). Appendix A gives the definitions of the variables used in this study.

3.2 Preliminary analyses

Prior to data analysis, data is prepared by checking inconsistencies and ensuring that it is suitable for the method used. Firstly, the sample used is checked for adequacy using the Kaiser-Meyer-Olkin test. The variables are also checked for reliability using the Cronbach’s alpha. These measures are discussed below.

3.2.1 Data adequacy and reliability

The Kaiser-Meyer-Olkin (KMO) is a measure used to check the adequacy of a sample. Small values of the KMO are a perfect indication that the connections between sets of variables cannot be explained by other variables. The KMO is described by the following equation:

$$\text{KMO} = \frac{\underbrace{\sum\sum r_{ij}^2}_{i\neq j}}{\underbrace{\sum\sum r_{ij}^2}_{i\neq j} + \underbrace{\sum\sum a_{ij}^2}_{i\neq j}} \quad (1)$$

Where r_{ij} represent Pearson correlation and a_{ij} is partial correlation between items i and j. The KMO value must be greater than 0.5 but less than or equal to 1 for the sample to proof adequate. According to Field (2005), a value closer to 1 indicates that patterns of correlations are relatively compact and so cluster analysis should yield distinct and reliable clusters. Kaiser (1974) suggested measures in the ranges: 0.8 or above, excellent; 0.7 moderate; 0.6 is mediocre; 0.5 is miserable and below 0.5 is unacceptable.

3.2.2 Commensurability

All clustering techniques require commensurable variables as highlighted by Fox (1982). It is imperative that intervally or ratio scaled variables are measure in identical scale units; otherwise the variables should be standardized by the range or z-transformed to have zero mean and unit standard deviation (Bacher, 2000). If variables of different measurement levels are used, the use of either a general distance measure of Gower's general similarity measure by Gower (1971) is recommended. Another recommendation by Bender et al., (2001) and Wishart (2003) is that the nominal and ordinal variables may be transformed to dummies and treated as quantitative variables. One of the advantages of SPSS TwoStep clustering over other methods is that it offers the possibility to handle continuous and categorical variables. This method can simultaneously handle quantitative and nominal variables with different scale units. However, the analyst can decide to define ordinal variables either as metric or nonmetric.

3.2.3 Reliability

In cluster analysis, the variables are expected to be homogeneous and measure similar construct, for instance, crime. Since crime is measured on a continuous scale, a certain degree of correlation between the variables is expected. Cronbach's alpha is used in this study as a measure of internal consistency and reliability of the data. It measures how well a set of items measure a single unidimensional latent construct. This measure is almost the same as the Pearson's correlation coefficient. Owing to the multiplicity of the variables measuring the clusters, the Cronbach's alpha is considered most suitable since it has the most utility of multi-item scales at the interval level of measurement (Cooper and Emory 1995). Its value ranges between 0 to 1 with values closer to 0 implying that the items do not measure the same construct and values closer to 1 measuring the same construct. Mathematically, Cronbach's α is defined as:

$$\alpha = \frac{kr}{1+(k+1)} \quad (2)$$

Where k is the average correlations between the variables and r is the number of variables. A commonly accepted rule of thumb for describing internal consistency using Cronbach's alpha according to Kline (1999) and Cronbach and Shavelson (2004) is as follows: $\alpha \geq 0.9$ is excellent, $0.8 \leq \alpha < 0.9$ is good, $0.7 \leq \alpha < 0.8$ is acceptable, $0.6 \leq \alpha < 0.7$ is questionable, $0.5 \leq \alpha < 0.6$ is poor and $\alpha < 0.5$ is unacceptable. The following section presents the data analysis algorithm used in the main analysis.

3.2 SPSS TwoStep cluster method

This modern clustering procedure was developed by Chiu et al., (2001). It has the capability to handle both metric and non-metric variables. It does so by extending the model-based distance measure used by Banfield and Raftery (1993) to situations with the mentioned variables. TwoStep clustering method is also known to utilising a two-step clustering approach similar to one used in (Zhang et al., 1996). Furthermore, it provides the capability to automatically compute the optimal number of clusters. This method is again recommended in situations when there is no introductory information. According to this method, SPSS cluster component automatically provides the proper number of clusters if the desired number of clusters is unknown. Described below are the two steps of this method.

3.2.1 Step 1: pre-cluster the data

In this step a sequential clustering approach as recommended by (Theodoridis and Koutroumbas, 1999) is used. This method scans the records one by one and then decides if the current record should merge with the previously formed clusters or start a new cluster based on the distance criterion. By doing this, a new data matrix with fewer cases for the next step is computed. Automatically, SPSS implements this procedure by constructing a modified cluster feature (CF) tree according to (Zhang et al., 1996). The CF-tree consists of levels of nodes, with each node containing a number of entries. A leaf entry represents a sub-cluster desired. According to this procedure, the non-leaf nodes and their entries guide a new record into a correct leaf node immediately.

3.2.2 Step 2: group the data into sub-clusters

In this step, the sub-clusters resulting from the first step are taken as input and grouped into the desired number of clusters. Since the number of sub-clusters is much less than the number of original records, the use of traditional clustering methods can be used effectively. Similar to agglomerative hierarchical techniques, the pre-clusters are merged using stepwise procedure. This procedure is repeated until all clusters are in collected in a unit cluster. In contrast to agglomerative hierarchical techniques, an underlying statistical model is used. The model assumes that the continuous variables x_j $(j = 1, 2, \dots, p)$ are within cluster i independent normal distributed with means μ_{ij} and variances σ_{ij}^2. The categorical variables a_j are within cluster i independent multinomial distributed with probabilities π_{ijl}, where (jl) is the index for the l-th category $(l = 1, 2, \dots, m_l)$ of variable a_j $(j = 1, 2, \dots, q)$. The Euclidean and a log-likelihood distance measures are available for this scenario. The latter can handle mixed type attributes and is defined as:

$$d(i, s) = \xi_i + \xi_s - \xi_{(i,s)} \tag{3}$$

ξ_i is interpreted as a kind of dispersion within the cluster. Similar to agglomerative hierarchical clustering, clusters with the smallest distance are merged in each stem. The log-likelihood function for the step with k clusters is computed as:

$$l_k = \sum_{v=1}^{k} \xi_v. \tag{4}$$

This function l_k is interpreted as dispersion within clusters not the exact log-likelihood function. In an instance where only nonmetric variables are used, l_k becomes the entropy within the k number of clusters.

Determining the number of clusters: SPSS utilises a two phase estimator to automatically determine the number of clusters. Firstly, estimator Akaike's Information Criterion (AIC) calculated as:

$$AIC_k = -2l_k + 2r_k, \tag{5}$$

and another estimator, the Bayesian Information Criterion (BIC) is computed as:

$$BIC_k = -2l_k + r_k \log n \tag{6}$$

Where r_k represents the number of independent parameters. The relative contribution of variables to form the clusters is computed for both types of variables (continuous and categorical).

For the continuous variables, the importance measure is based on:

$$t = \frac{\hat{\mu}_k - \hat{\mu}_{sk}}{\hat{\sigma}_{sk}} \sqrt{N_k}, \tag{7}$$

Where $\hat{\mu}_k$ is the estimator of k continuous variable mean, for entire dataset, and $\hat{\mu}_{sk}$ is the estimator of k continuous variable mean, for cluster j. This relative importance measure has a Student distribution with N_{k-1} degrees of freedom. The significance level is two-tailed. For the categorical variables, the importance measure is based on χ^2 test computed as:

$$\chi^2 = \sum_{l=1}^{L_k} \left(\frac{N_{skl}}{N_{kl}} - 1 \right)^2, \tag{8}$$

Which is distributed as a χ^2 with L_k degrees of freedom. The values of these criteria are generated simultaneously and have been reported to be good estimators of the maximum number of clusters according to Chiu et al., (2001). The maximum number of clusters is set equal to number of clusters where the ratio BIC_k/BIC_1 is in excess of c_1 [4] for the first time.

In the next phase, ratio change $R(k)$ in distance for k clusters is defined as:

$$R(k) = d_{k-1}/d_{k,} \tag{9}$$

Where d_{k-1} is the distance if k clusters are merged to k-1 clusters. The distance is $d_{k=}l_{k-1} - l_k$. The number of clusters is obtained for the solution where a big jump of the ratio change calculated as $R(k_1)/R(k_2)$. For the largest values of $R(k)(k = 1, 2, \dots, k_{max}; k_{max}$ is obtained from the first step. If the ratio change is larger than the threshold value c_2, the number of clusters is set to be equal to k_1, otherwise it set to be equal to the solution with $\max(k_1, k_2)$.

Assigning members to clusters: Allocation of members to the closest cluster is done deterministically according to the distance measure used. A disadvantage about this deterministic allocation is that it may result in biased estimates of the cluster profiles if the clusters overlap (Bacher, 2000).

Modification: This stage allows the analyst to define method of outlier treatment. Cluster analysis is sensitive to the inclusion of outliers and therefore outliers must be dealt with prior to obtaining the final cluster solution. According to Leonard and Droege (2008), in cluster analysis an outlier can describe a case that is either an extreme value within its own cluster or a value so extreme as not to belong to any cluster. A conventional level of significance, say, 5% may be specified as a value for the fraction of noise. If the number of cases is less than the defined fraction of the maximum cluster size then a pre-cluster is considered as a potential outlier cluster. Outliers may be ignored in the second step. Furthermore, missing values are replaced with the series mean. One of the disadvantages of the TwoStep method is that it does

[4] Default value based on simulation studies of the authors of SPSS TwoStep clustering

not allow missing values. The items that have missing values are not considered for analysis and this tend to reduce the sample size.

4 Empirical results

The analysis of data is done with the help of SPSS 22 for windows. The results of this analysis are presented below:

4.1 Preliminary results

Table 12. KMO and Bartlett's test

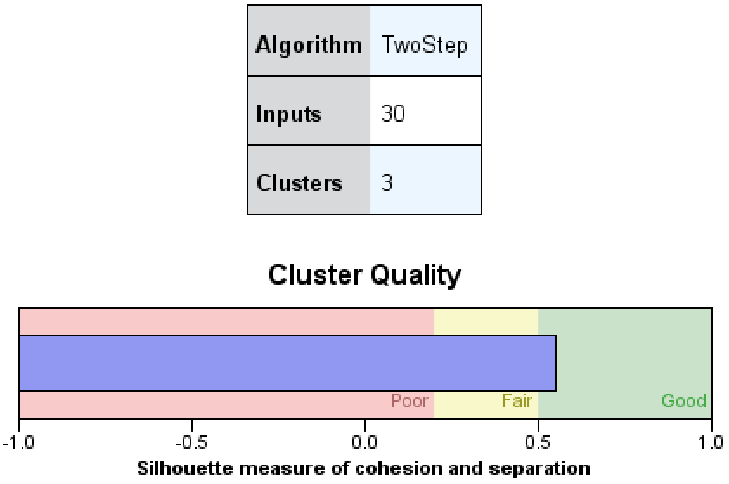
Kaiser-Meyer-Olkin Measure of Sampling Adequacy	0.949

As indicated in Table 1, the KMO test for measuring sampling display satisfactory results. KMO of 0.948 suggest that the degree of common variance between the 29 variables is marvellous entailing that if cluster analysis is conducted, the clusters extracted will account for a significant amount of variance.

Table 2. Reliability statistics

Cronbach's Alpha	N of Items
0.916	29

Figure 1. Model summary



Algorithm	TwoStep
Inputs	30
Clusters	3

Table 3. Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[5]	Ratio of BIC Changes[6]	Ratio of Distance Measures[7]
1	25136.425			
2	16091.877	-9044.548	1.000	2.498
3	12723.503	-3368.374	.372	5.138
4	12407.139	-316.364	.035	1.093
5	12153.772	-253.368	.028	1.149
6	11987.668	-166.104	.018	1.699
7	12063.135	75.468	-.008	1.197
8	12195.462	132.327	-.015	1.253
9	12386.119	190.657	-.021	1.033
10	12584.096	197.976	-.022	1.066
11	12795.938	211.843	-.023	1.039
12	13015.554	219.616	-.024	1.373
13	13289.895	274.341	-.030	1.059
14	13572.379	282.484	-.031	1.021
15	13857.705	285.326	-.032	1.037

[5] The changes are from the previous number of clusters in the table

[6] The ratios of changes are relative to the change for the two cluster solution

[7] The ratios of distance measures are based on the current number of clusters against the previous number of clusters

Cronbach's alpha has been computed for the 29 variables. Table 2 above displays some of the results obtained. The results shows overall alpha as 0.916, which is excellent as described by Cronbach and Shavelson (2004). This is an indication of strong internal consistency among the 29 crime variables. Essentially this means that criminals who passionately tended to commit say, murder, also tended to commit other crimes. Similarly, criminals who unintentionally committed murder tended not to commit other crimes.

4.2 A TwoStep cluster results

This section discusses the results based on TwoStep clustering Algorithm. Presented first is the Auto-Clustering statistics summarised in Figure 1 and Table 3 through Table 5. These results are used to assess the optimal number of clusters in the analysis.

The output in Figure 1 confirms that is good to represent the 29 variables in three clusters.

Table 3 reveals that the lowest BIC coefficient (11987.668) is for six clusters. However, this is not in accordance with SPSS algorithm (Figure 1) which reveals the optimal number of clusters as three. Also shown in Table 4 is the largest ratio of BIC changes (0.372) and corresponding distances (5.138) for three clusters. The AIC in Appendix B also concur with the BIC on the optimal number of cluster as three. As a result, this confirms information as displayed in Figure 1 that 29 serious crimes in South Africa can best be separated into three clusters. The proportion of cluster distribution according to the three determined clusters is shown in Table 4.

Table 4. Cluster distribution

		% of Combined	% of Total
Cluster	1	61.5%	61.5%
	2	27.4%	27.4%
	3	11.1%	11.1%
	Combined	100.0%	100.0%
Total			100.0%

The output shows cluster 1 comprising 61.5% of variables and the third and smallest cluster with about 11.1% variables.

Table 5 reveals the contribution of each variable within clusters. It is clear that malicious damage to property (100%) and robbery with aggravating circumstances (97%) are the domineering crimes in South Africa. Common assault, sexual crimes and others also pose a major threat to residents contributing between 60% and 70% to crimes. Common robbery contributes about 60% to crimes in the country. However, stock theft, public violence, truck hijacking and drug-related crimes are not a matter of concern. These variables contribute less than 20% to crime. Appendix C clearly shows a visual distribution of these variables.

Table 5. Variable importance

Nodes	Variable	Importance
v15	Stock theft	0.0119
v27	Public violence	0.0991
v23	Truck hijacking	0.1103
v17	Drug-related crime	0.1934
v19	All theft not mentioned elsewhere	0.3146
v29	Neglect and ill treatment of children	0.3305
v28	Crimen injuria	0.3396
v18	Driving under the influence of alcohol or drugs	0.3425
v30	Kidnapping	0.3937
v2	Murder	0.3988
v20	Commercial crime	0.4192
v9	Arson	0.4343
v21	Shoplifting	0.4439
v16	Illegal possession of firearms and ammunition	0.4611
v26	Culpable homicide	0.4633
v22	Carjacking	0.4891
v25	Robbery at residential premises	0.5032
v14	Theft out of or from motor vehicle	0.5083
v5	Assault with the intent to inflict grievous bodily harm	0.5238
v13	Theft of motor vehicle and motorcycle	0.5279
v11	Burglary at business premises	0.5472
v4	Attempted murder	0.5705
v8	Common robbery	0.6018
v3	Sexual crimes	0.6114
v12	Burglary at residential premises	0.6209
v24	Robbery at business premises	0.6480
v6	Common assault	0.6904
v7	Robbery with aggravating circumstances	0.9703
v10	Malicious damage to property	1

5 Conclusions

Clustering methods are more effective in fields which use large datasets. These methods can be used to find hidden patterns in the data. Most data collected practically is a combination of both numerical and categorical attributes and therefore requires specialised clustering algorithms such as TwoStep clustering. Classical clustering methods do not handle mixture of variables very well. This study adopted a TwoStep algorithm due to the large set of data analysed. The method does not require the use of hierarchical and non-hierarchical clustering at the same time. It has the ability to determine the optimal number of clusters automatically.

Preliminary analysis of results confirmed that the data used meet the requirements for cluster analysis. Using this method on 29 crimes reported at the 1119 police stations in South Africa, three profiles were automatically identified with both the AIC and BIC. The most important profile contains 13 variables with the weights of between 100% and 50%. These may be regarded as the most domineering crimes in South Africa. The third profile contains 4 least hostile variables contributing not more than 20% to crimes. The remainder out of 29 crimes is a composition of cluster number 2. These profiles are clearly visible in Figure 2.

This information may be used by the SAPS authorities. They may refer to these findings when amending policies or coming up with strategies to combat crime. The results clearly show the most to least threatening crimes in the country. More focus may be given to the most hostile crimes in the short run and plans to fight crimes in the third profile may be for long term. Individual researchers, governmental and private sectors responsible for the implementation of the findings obtained from crime analysis can also use the findings of this study. Multivariate techniques such as multiple regression, discriminant and multivariate analysis of variance can be used as follow-up methods using the three profiles as variables to their analysis.

References

1. Altas, D., Kubas, A. and Sezen, J. (2013), "Analysis of environmental sensitivity in Thrace region through TwoStep cluster", *Trakia Journal of Science,* Vol. 11 No. 3, pp. 318-329.
2. Bacher, J. (2000), "A probabilistic clustering model for variables of mixed type", *Quality and Quantity,* Vol. 34, pp. 223-235.
3. Banfield, J.D. and Raftery, A.E. (1993), "Model-based Gaussian and non-Gaussian clustering", *Biometrics,* Vol. 49, pp. 803-821.
4. Bender, S., Brand, R. and Bacher, J. (2001), "Re-identifying register data by survey data: An empirical study", *Statistical Journal of the UN Economic Commission for Europe,* Vol. 18 No. 4, pp. 373-381.
5. Chiu, T., Fang, D., Chen, J., Wang, Y. and Jeris, C. (2001), "A Robust and scalable clustering algorithm for mixed type attributes in large database environment", In proceedings of the 7th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 263-268.
6. Cooper, D.R. and Emory, C.W. (1995), *Research methods,* Homewood, IL. Richard D Irwin Inc.
7. Cronbach, L.J. and Shavelson, R.J. (2004), "My current thought on coefficient alpha and successor procedures", *Educational and Psychological Measure,* Vol. 64 No. 3, pp. 391-418.
8. Field, A. (2005), "Factor analysis using SPSS" [online] Available from URL: http://wwwsussexacuk/users/andyf/factorpdf.
9. Fraley, C. (1998), "Algorithms for model-based Gaussian hierarchical clustering", *SIAM Journal of Scientific Computing*, Vol. 20, pp. 270-281.
10. Fraley, C. and Raftery, A.E. (1998), "How many clusters? Which clustering method? Answers via model-based cluster analysis", *Computer Journal,* Vol. 4, pp. 578-588.
11. Gower, J.C. (1971), "A general coefficient of similarity and some of its properties", *Biometrics,* Vol. 27, pp. 857-872.
12. Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E. (2010), M*ultivariate data analysis: A global perspective* (7th ed), Upper Saddle River, Pearson Prentice Hall.
13. Kaiser, H.F. (1974), "An index of Factorial Simplicity", *Psychometrika,* Vol. 3, pp. 934-36.
14. Kim, J., Yang, J. and O'lafsson, S. (2009), "An optimization approach to partitional data clustering", *Journal of the Operations Research Society,* Vol. 60 No. 8, pp. 1069-1084.
15. Kline, P. (1999), *The handbook of psychological testing* (2nd ed)., London, Routledge.
16. Lattin, J., Carroll, J.D. and Green, P.E. (2003), *Analyzing multivariate data*, Canada, Brooks/Cole publishing.
17. Leonard, S.T. and Droege, M. (2008), "The uses and benefits of cluster analysis in pharmacy research", *Research in Social and Administrative Pharmacy,* Vol. 4, pp. 1-11.
18. Manly, B.F.J. (2001), *Multivariate Statistical Methods: A Primer,* Chapman and Hall/CRC.
19. Özdamar, K. (1999), *Statistical data analysis with package programs-II*, Eskişehir, Kaan Pub.
20. Po, R.W., Guh, Y.Y. and Yang, M.S. (2009), "A new clustering approach using data envelopment analysis", *European Journal of Operations Research*, Vol. 199, pp. 276-284.
21. Rege, M., Dong, M. and Fotouhi, F. (2008), "Bipartite isoperimetric graph partitioning for data co-clustering", *Data Mining and Knowledge Discovery,* Vol. 16 No. 3, pp. 276-312.
22. Rencher, A.C. and Christensen, W.F. (2022), *Methods of Multivariate Analysis,* (3rded) Wiley-Interscience, ISBN 9780470178696.
23. Schiopu, D. (2010), "Applying TwoStep cluster analysis for identifying bank customers' profile", *Seria StiinÑe Economice,* Vol. 12 No. 3, pp. 66-75.
24. Suhr, D. (2009), "Principal component versus exploratory factor analysis", SUGI 30 Proceedings Retrieved from http://wwwsascom/proceedings/sugi30/203-30pdf.
25. Thanassoulis, E. (1996), "A data envelopment analysis approach to clustering operating units for resource allocation purposes", *Omega,* Vol. 24, pp. 463-476.

26. Theodoridis, S. and Koutroumbas, K. (1999), *Pattern recognition,* New York, Academic Press.

27. Wishart, D. (2003), "k-Means clustering with outlier detection, mixed variables and missing values in M Schwaiger and Opitz (Eds), Exploratory data analysis in empirical research", Proceedings of the 25th Annual Conference of the Gesellschaft f¨ur Klassifikation, University of Munich, March 14-16, 2001, Studies in classification, data analysis, and knowledge organization, 216-226, Berlin, Springer.

28. Yin, X., Han, J. and Yu, P.S. (2007), Crossclus: User-guided multi-relational clustering, *Data Mining and Knowledge Discovery*, Vol. 15 No. 3, pp. 321-348.

29. Zhang, T., Ramakrishnon, R. and Livny, M. (1996), "BIRCH: An efficient data clustering method for very large databases", Proceedings of the ACM SIGMOD conference on management of data, pp. 103-114, Montreal, Canada.

Appendix A. Variables used

The following is a list of variables used in the analysis. They are all measured on the metric scale and are therefore continuous.

Table 13. Statistical variables

1	Murder-the unlawful and intentional killing of a human being by another
2	Rape-unlawful compelling of a person through duress to have sexual intercourse
3	Attempted murder-an undertaking to do an act that entails more than mere preparation but does not result in the successful completion of the act
4	Assault with the intent to inflict grievous bodily harm-unlawfully touching people without their consent with an intention to harm them
5	Common assault-intentionally or recklessly causing another person to apprehend immediate infliction of unlawful force
6	Robbery with aggravating circumstances-robbery in which a firearm or other dangerous weapon is wielded and bodily harm is threatened
7	Assaults with intend to rob-unlawfully causing harm to people with an intention to take their personal belongings
8	Arson-intentionally and maliciously setting fire to property
9	Malicious damage to property-unlawfully and intentionally damaging property
10	Burglary at business premises-intrusion or the trespassing into someone else's business premises
11	Burglary at residential premises-intrusion or the trespassing into someone else's house or home
12	Theft of motor vehicle and motorcycle-taking of a vehicle without the owner's authorisation
13	Theft out of or from motor vehicle-unlawful appropriation of objects from the motor vehicle
14	Stock theft-unlawful and intentional taking of someone's horse, cow, donkey' etc
15	Illegal possession of firearms and ammunition-unlawful ownership of unregistered dangerous weapons
16	Drug-related crime-offence committed as a result of drug or alcohol use
17	Driving under the influence of alcohol or drugs-the act of driving a motor vehicle with blood levels of alcohol in excess of a legal limit
18	All theft not mentioned elsewhere-theft of any kind that is not listed here in
19	Larceny-the unlawful taking of personal property with intent to deprive the rightful owner of it permanently
20	Shoplifting-theft of goods from a retail establishment
21	Carjacking-take control of someone's car by force
22	Truck hijacking-force the owner or driver of a truck to give you total control of it
23	Robbery at business premises-unlawfully taking the property or goods at someone's business by use of violence or intimidation
24	Robbery at residential premises-unlawfully taking the property or goods at someone's house or home by use of violence or intimidation
25	Culpable homicide- negligently killing of another person
26	Public violence-violent disturbance of the public peace by a group of people assembled for a common purpose
27	*Crimen injuria*-act of unlawfully, intentionally and seriously impairing the dignity of another
28	Neglect and ill treatment of children-intentionally ignoring and forcing unnecessary treatment on children
29	Kidnapping-unlawfully seizing people by force against their will

Source: Stats SA

It must be noted that the definitions given to these variables are context specific and must be appropriated with SAPS in mind. Some other countries might have different meanings to these concepts or phrases.

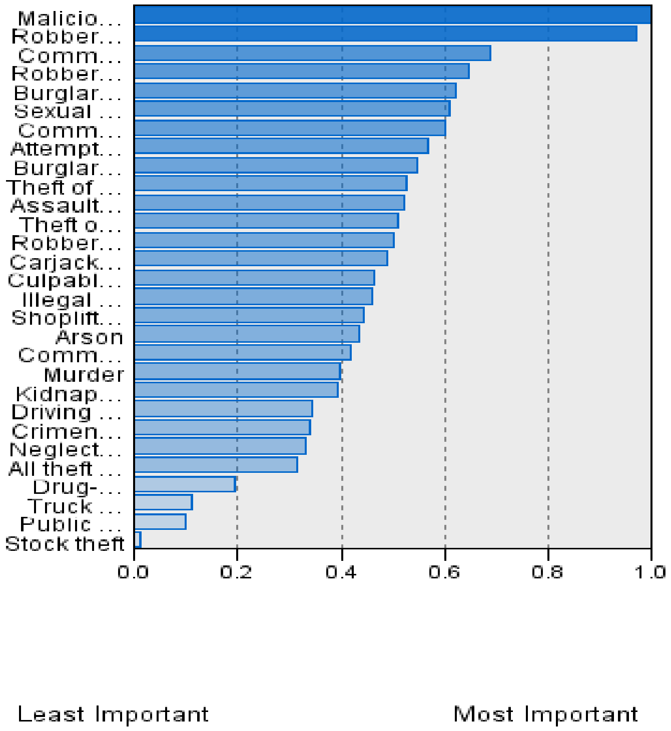
Appendix B. Auto clustering AIC

Table 3. Auto clustering AIC

Number of Clusters	Akaike's Information Criterion (AIC)	AIC Change	Ratio of AIC Changes	Ratio of Distance Measures
1	40542.003			
2	33269.238	-7272.764	1.000	2.753
3	32125.541	-1143.698	0.157	2.580
4	33122.489	996.948	-0.137	1.361
5	34478.943	1356.454	-0.187	1.122
6	35943.443	1464.500	-0.201	1.075
7	37469.839	1526.396	-0.210	1.361
8	39215.127	1745.287	-0.240	1.337
9	41113.435	1898.308	-0.261	1.078
10	43044.493	1931.058	-0.266	1.191

Appendix C. Visual display of variable importance

Figure 2. Variable importance


Malicio...
Robber...
Comm...
Robber...
Burglar...
Sexual ...
Comm...
Attempt...
Burglar...
Theft of ...
Assault...
Theft o...
Robber...
Carjack...
Culpabl...
Illegal ...
Shoplift...
Arson
Comm...
Murder
Kidnap...
Driving ...
Crimen...
Neglect...
All theft ...
Drug-...
Truck ...
Public ...
Stock theft
Least Important
Most Important