STUDENT SESSION

SOFTWARE IMPLEMENTATION AND COMPARATIVE ANALYSIS OF ALGORITHMS FOR CLUSTERING DATA ON CRIMINAL ACTIVITIES

Milica Petrić*,
[0009-0004-2039-6513]

Vladica Stojanović
[0000-0002-3819-4387]

Department of Informatics and Computer Sciences, University of Criminal Investigation and Police Studies, Belgrade, Serbia

Abstract:

This study investigates the application of clustering algorithms in crime data analysis, with a special emphasis on their software implementation and interpretation within security-oriented systems. A comparative analysis of four widely used clustering algorithms (K-Means, Agglomerative Hierarchical Clustering, Gaussian Mixture Models, and DBSCAN) is conducted to determine the most suitable method and the optimal number of clusters. The selected procedures are applied to bivariate time series of real-world criminal activities, with a focus on clustering performance, interpretability, and identification of significant patterns. The resulting clusters are interpreted as different regime states of the observed system, providing insight into temporal variations in criminal activities. The results obtained thus indicate that clustering methods can serve not only as descriptive tools, but also as a valuable analytical component in security systems and decision support frameworks.

Keywords:

Clustering, Criminal Activities, Software Implementation, Time Series, Unsupervised Learning.

INTRODUCTION

Contemporary security analytics increasingly relies, in addition to traditional forensic approaches, on information technology techniques, with the aim of discovering hidden patterns, identifying risks and supporting decision-making processes. In this context, the analysis of criminal activities makes extensive use of machine learning and data mining (DM) methods, which enable the processing of large and complex datasets [1]. Data clustering as an unsupervised learning method is of particular importance here, as it allows the analysis of unlabelled data, which is often the case in real security systems [2]. By grouping similar objects into homogeneous wholes, clustering allows the discovery of latent structures in the data, which may indicate different forms and intensities of criminal activities. Therefore, the application of clustering in crime analysis involves identifying behavioural patterns, detecting atypical events, and separating characteristic incidents [3]. This is particularly important in time series analysis, where the resulting clusters can be interpreted as different "regime" states of the system, i.e., periods of different intensity of criminal activity [4].

Correspondence:
Milica Petrić

e-mail:
milica.petric2001@gmail.com

DOI: 10.15308/Sinteza-2026-577-584

Nevertheless, the selection of an adequate clustering algorithm and the interpretation of the results represent a significant challenge, as different approaches (partitional, hierarchical, density and probabilistic) start from different assumptions about the data structure and can lead to different results [5]. Based on the above, the goal of this research is twofold: (i) a comparative analysis of the most commonly used clustering algorithms (K-Means, agglomerative-hierarchical algorithm, Gaussian mixture models, and DBSCAN) on real crime data, and (ii) the interpretation of the resulting clusters as different regime states in bivariate time series. Special emphasis is placed on the software implementation of various clustering techniques and algorithms, the selection of the optimal clustering method and number of clusters, as well as the connection of clustering methods with the analysis of crime dynamics. In this way, clustering is not viewed solely as a data grouping technique, but as an analytical tool for identifying structures of importance for the prevention, control, and prediction of criminal activities.

2. MATERIALS AND METHODS

This section first describes the dataset used in the study, which represents the dynamics of certain related criminal activities. Subsequently, the applied clustering techniques, as well as the procedure used to determine the most appropriate clustering method and the optimal number of clusters, are described.

2.1. DATASET

The dataset used in this study consists of records of the number of criminal acts committed on the territory of the Republic of Serbia, collected based on the official report of the Ministry of Internal Affairs. The data were observed on a daily basis, during the period from January 1, 2015 to December 31, 2024. In this way, time series of length T=3653 are obtained, which for the purposes of analysis are constructed into two bivariate time series (labelled as Series 1 and Series 2), each representing a pair of related criminal activities. The first, Series 1, includes crimes related to serious bodily harm (KD_121) and minor bodily harm (KD_122), while the second, Series 2, refers to property crimes with elements of violence, in particular aggravated theft (KD_205) and robbery (KD_206).

The classification of criminal offenses was carried out based on the official Criminal Code of the Republic of Serbia, and the choice of the above-mentioned paired variables is motivated by the fact that they are usually viewed together, under similar socio-economic and security aspects. This also allows for modelling of observed data as bivariate time series, which enables joint analysis of their dynamics and interdependence. To ensure comparability and computational efficiency, a subset of the monthly data is considered, consisting of n=120 consecutive observations for each time series. Before applying the clustering procedure, the data are first statistically processed, so that the descriptive statistical analysis shown in Table 1 provides initial insight into their structure and variability.

Table 1. Summary statistics of observed time series of criminal activity (KD_121, KD_122, KD_205, KD_206)

Statistics	Series 1		Series 2	
	KD_121	KD_122	KD_205	KD_206
Mean	78.38	97.94	7.467	104.4
Median	78.00	99.00	7.000	77.00
Mode	74.00	101.0	7.000	67.00
Standard Deviation	15.42	22.69	3.889	73.19
Variance	237.8	514.6	15.12	5357
Kurtosis	2.956	2.790	4.296	4.627
Skewness	-0.310	0.352	0.870	1.434
Range	75.00	114.0	23.00	323.0
Minimum	33.00	57.00	0.000	25.00
Maximum	108.0	171.0	23.00	348.0
Cross-correlation	0.660		0.565	

More specifically, for bodily injuries (KD_121 and KD_122), it is noticeable that both variables show relatively stable behaviour, with somewhat higher to moderate dispersion around the mean values. There is somewhat greater variability in the case of minor bodily injuries (KD_122), compared to severe bodily injuries (KD_121), suggesting that less severe incidents occur with greater temporal fluctuations. Moreover, skewness values close to zero indicate approximately symmetric distributions, while kurtosis suggests a somewhat flatter distribution compared to the standard normal case (when Kurtosis =3).

On the other hand, in the case of property crimes, the variability is much more pronounced. In particular, the crime of robbery (KD_206) has a significantly higher standard deviation and range, indicating the presence of extreme values and occasional spikes in criminal activity. This is also supported by positive skewness and higher values of kurtosis, suggesting a right-skewed distribution with heavier tails. Such statistical properties indicate that the second bivariate series (KD_205 and KD_206) is more heterogeneous and prone to extreme fluctuations, which may also affect the performance of clustering algorithms.

Finally, the cross-correlation between the components of the observed time series is equal to 0.660 for minor and serious bodily injuries (Series 1), while for robbery crimes it is 0.565 (Series 2). This indicates a significant, but not extremely high correlation that would, for instance, enable some other (e.g., regression) analysis of observed series. Moreover, this confirms the interconnectedness of the observed criminal phenomena, which is suitable for the application of clustering techniques.

2.2. CLUSTERING ALGORITHMS

As mentioned above, four widely used clustering algorithms are considered in this study. They represent different paradigms for unsupervised learning, with different approaches (partition-based, hierarchical, model-based, and density-based approaches). A more detailed description of these algorithms is given as follows (see, e.g., [6] for some more detail):

1. K-Means is used as a representative partitioning algorithm that assigns observations to a predefined number of clusters by minimizing the variance within clusters. Due to its computational efficiency, it is particularly suitable for large data sets, although it assumes clusters of approximately spherical shape.
2. Agglomerative-hierarchical (AH) clustering is a method whose primary goal is to transform the distances between data into a series of nested partitions. In this way, it does not produce a single solution, but an entire clustering hierarchy. First, each data point is treated as a separate cluster, and then in each subsequent iteration, the two closest clusters are "merged" until the desired number of clusters remains.
3. Gaussian mixture models (GMM) use a stochastic approach, where clusters are obtained based on probabilistic mixtures of Gaussian distributions. In this way, unlike other clustering methods, GMM allows for a more flexible representation of more complex data structures.
4. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a method based on data densities, which identifies clusters as regions of high density separated by sparse areas. It is particularly effective in detecting clusters of arbitrary shape and identifying noise, without the need for a predefined number of clusters.

It is worth noting that all the above clustering algorithms are implemented in software environment using the "sklearn.cluster" module in the 64-bit "IDLE Python 3.14" framework [7]. This allows for a comparative analysis of different clustering strategies, in terms of their performance and interpretability, as well as their efficiency in segmenting criminal activity data.

2.3. OPTIMAL CLUSTERING METHOD AND NUMBER OF CLUSTERS

Choosing the optimal clustering method and the appropriate number of clusters is a key step in further analysis, as different algorithms and parameter choices can lead to significantly different results. In this study, the selection is based on the Silhouette Score (SS), a widely used metric for evaluating clustering quality, first introduced in [8]. For each observation (x_i), the SS is defined as:

$$\mathrm{S(i)} = \frac{\mathrm{b(i)} - \mathrm{a(i)}}{\max\{\mathrm{a(i)},\ \mathrm{b(i)}\}} = \begin{cases} 1 - \frac{\mathrm{a(i)}}{\mathrm{b(i)}}, & \mathrm{a(i)} < \mathrm{b(i)} \\ 0, & \mathrm{a(i)} = \mathrm{b(i)} \\ \frac{\mathrm{b(i)}}{\mathrm{a(i)}} - 1, & \mathrm{a(i)} > \mathrm{b(i)}. \end{cases}$$

Equation 1. Definition of the SS values

In Equation 1, the term $a(i)$ represents the average distance between the i-th observation and all other points within the same cluster (intra-cluster distance), while $b(i)$ denotes the minimum average distance

between the *i*-th observation and points belonging to the nearest neighbouring cluster (inter-cluster distance). Thus, SS takes values in the interval [-1,1], where values close to 1 indicate that the observation matches well with its own cluster and poorly with neighbouring clusters, values close to 0 indicate that the observation is located near the boundary between clusters, while negative values indicate potential misclassification.

The overall clustering quality is then estimated by averaging the values of $S(i)$, defined by Equation 1, over all observations $x_1,\ldots,x_n$. Hence, higher average SS values indicate better defined and more separated clusters and vice versa. In this study, SS values are calculated for each clustering algorithm, as well as for different numbers of clusters, wherever applicable. In this way, the optimal clustering configuration is chosen as the one that maximizes the average SS values, thus ensuring a consistent and objective comparison of clustering methods, and providing a basis for selecting the most suitable model for further analysis. As already mentioned, the selection procedure is implemented in the Python programming environment using the "scikit-learn" library, in particular the "sklearn.cluster" module for clustering algorithms, "sklearn.mixture" module for GMM method, as well as the "sklearn.metrics" module for SS calculations and cluster validation. A software implementation used to evaluate multiple clustering methods, calculate SS values, and determine the optimal clustering configuration is illustrated in Listing 1.

```
import numpy as np
from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN
from sklearn.mixture import GaussianMixture
from sklearn.metrics import silhouette_score

def evaluate_clustering(X, k_range = range(2,10)):
    ss_kmeans, ss_ah, ss_gm = [], [], []
    ss_db = None

    # K-Means
    for k in K_range:
        labels = KMeans(n_clusters=k, random_state=42).fit_predict(X)
        ss_kmeans.append(silhouette_score(X, labels))
        best_k_kmeans = k_range[np.argmax(ss_kmeans)]
        best_ss_kmeans = max(ss_kmeans)

    # Agglomerative Hierarchical Clustering
    for k in K_range:
        labels = AgglomerativeClustering(n_clusters=k).fit_predict(X)
        ss_ah.append(silhouette_score(X, labels))
        best_k_ah = k_range[np.argmax(ss_ah)]
        best_ss_ah = max(ss_ah)

    # Gaussian Mixture Model
    for k in K_range:
        labels = GaussianMixture(n_components=k, random_state=42).fit(X).predict(X)
        ss_gmm.append(silhouette_score(X, labels))
        best_k_gmm = k_range[np.argmax(ss_gmm)]
        best_ss_gmm = max(ss_gmm)

    # DBSCAN
    labels = DBSCAN(eps=0.5, min_samples=5).fit_predict(X)
    if len(set(labels)) > 1:
        ss_db = silhouette_score(X, labels)

    scores = [("kmeans", best_k_kmeans, best_ss_kmeans),
        ("ah", best_k_ah, best_ss_ah),
        ("gmm", best_k_gmm, best_ss_gmm)]

    if ss_db is not None:
       scores.append(("dbscan", None, ss_db))

    best_model = max(scores, key=lambda x: x[2])

     return best_model
```

Listing 1. Python implementation of SS-based selection of clustering methods and number of clusters

3. RESULTS AND DISCUSSION

This section presents the results of the clustering analysis and provides their interpretation in the context of the crime data. First, a comparative evaluation of the applied clustering algorithms based on the SS values and the corresponding optimal number of clusters is conducted. Then, the resulting clustering structures are analysed from a dynamic perspective, with particular emphasis on their interpretation as distinct criminal activity regimes.

3.1. COMPARATIVE ANALYSIS OF CLUSTERING RESULTS

The performance of the considered clustering algorithms is evaluated using the Silhouette Score, together with the corresponding optimal number of clusters. The results for both datasets are summarized in Table 2, while the graphical representation of the Silhouette Score values is given in Figure 1, below.

Note that for the dataset of bodily harm crimes (KD_121 and KD_122), the K-Means algorithm achieves the highest SS value (0.468), indicating the most coherent clustering structure.

On the other hand, the DBSCAN and GMM algorithms provide comparable results (0.451 and 0.443, respectively), suggesting that they can also adequately capture the underlying structure of the data. In contrast, the AH method yields a lower result (0.425). Further, for the dataset related to property crimes (KD_205 and KD_206), both K-Means and AH clustering achieve identical and significantly higher SS value (0.692), indicating a well-defined cluster structure. The GMM algorithm identifies a larger number of clusters (k=4), but with lower clustering quality (0.598), while DBSCAN method yields the lowest performance (0.580). These results indicate that clustering performance strongly depends on the data structure. While K-Means consistently provides robust results in both datasets, hierarchical clustering performs particularly well in the presence of more heterogeneous data, which is in line with some previously known results [9]. Finally, the other two clustering methods appear less suitable for the time series analysed. See Figure 1.

The clustering structure obtained using the K-Means algorithm is illustrated in Figure 2 for both data series. In the left plot of Figure 2, corresponding to bodily injury offenses, two well-separated clusters can be observed,

Table 2. Optimal values for clustering crime data using various clustering algorithms

Algorithm	Series 1		Series 2	
	Optimal number of clusters (k)	Silhouette Score (SS)	Optimal number of clusters (k)	Silhouette Score (SS)
K-Means	2	0.468	3	0.692
AH	2	0.425	3	0.692
GMM	2	0.443	4	0.598
DBSCAN	-	0.451	-	0.580

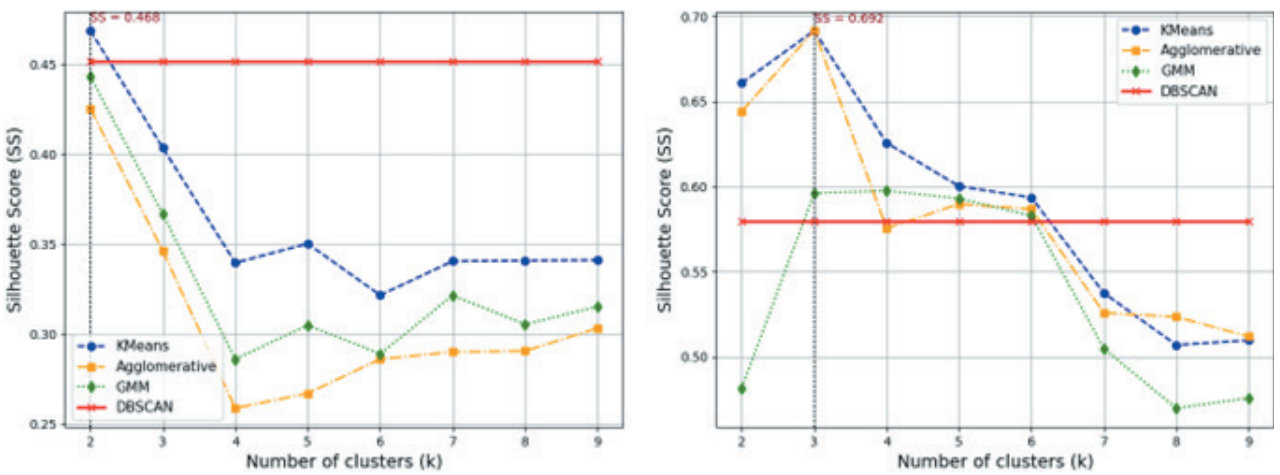

Figure 1. Dependence of SS values on number of clusters and clustering methods

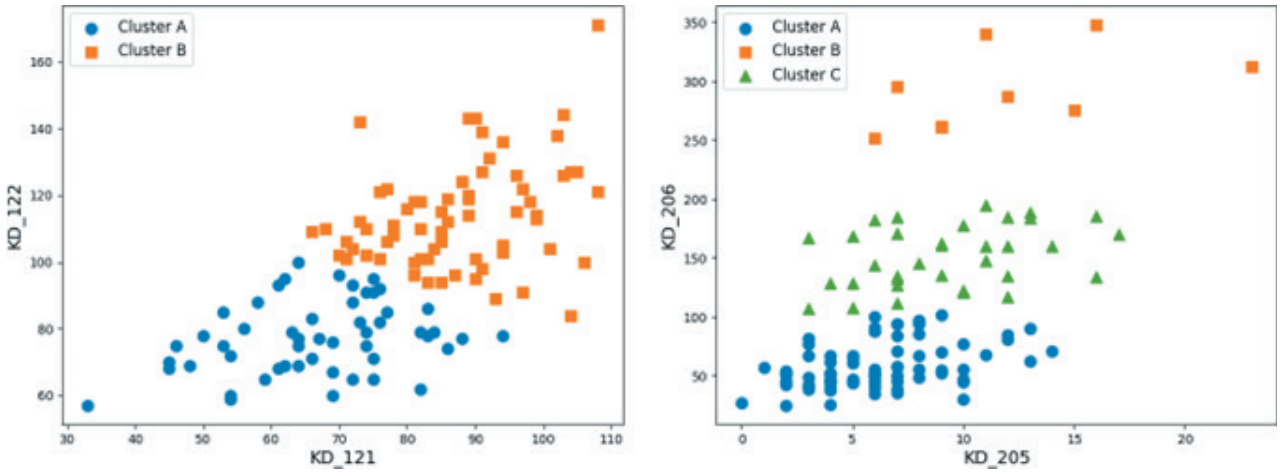

Figure 2. K-means clustering with the optimal number of clusters

indicating a stable segmentation between lower and higher levels of criminal activity. This confirms the suitability of K-Means for capturing the underlying structure of this dataset. Otherwise, in the right plot of Figure 2, related to property crimes, three clusters are identified, reflecting a more complex structure. In addition to the basic cluster, corresponding to low intensity of criminal offenses, the presence of medium and high intensity clusters indicates a layered organization of criminal activities. This is particularly pronounced for the crime of robbery (KD_206) and is consistent with the higher variability observed in this series.

According to the clustering results obtained in this way, the following interpretation can be given:

- For Series 1 (the bodily injury data set), the two resulting clusters clearly separate different levels of intensity of criminal activity. The first, Cluster A, represents periods of lower intensity, characterized by a reduced number of both serious and minor injuries. The second one (Cluster B) encompasses periods of increased violence with significantly higher values of both variables, which will be analysed in more detail below.
- For Series 2 (property crimes), three distinct clusters are identified, reflecting a specific structural pattern of criminal activity. The first, Cluster A, corresponds to relatively low values of both variables, while the second (Cluster B) is characterized by moderate levels of aggravated theft (KD_205) but consistently high values of robbery (KD_206). In this way, a well-defined, compact region is formed, although relatively close to the previous Cluster A. Conversely, Cluster C encompasses extreme events, with significantly elevated values, especially in the number of robberies. It indicates the presence of a specific type of criminal activity pattern, where robbery incidents occur more intensively even without a proportional increase in the associated crimes.

3.2. DYNAMIC ASPECTS OF CLUSTERING RESULTS

In the following, using the "K-Means Cluster" tool in the IBM SPSS Statistics software environment [10], cluster labels are associated with each element of the dataset, and the resulting descriptive statistical analysis is presented in Table 3. For the bodily injury dataset, the two clusters clearly correspond to different intensity levels of criminal activity. Accordingly, similar conclusions can be drawn as before. For bodily injury crimes, the first cluster is characterized by lower mean values of both variables, representing periods of reduced violence, while the second cluster includes higher mean values, indicating periods of increased intensity of violent incidents. For the data set related to property crimes, the first cluster corresponds to basic activity with relatively low values of both types of crimes, while the second is characterized by moderate values of aggravated theft and consistently elevated values of robbery. Finally, the third cluster represents extreme events, with significantly elevated values, especially in cases of robbery, although compared to the others, this cluster has a significantly smaller amount of data.

Below, a dynamic analysis of the time evolution of the above series is performed, where clearly defined periods of dominance of different clusters are observed. For the bodily injury dataset, shown in Figure 3, two dominant regimes can be identified. More precisely, the period from approximately 2015 to 2018 is predominantly characterized by Cluster B with high intensity of both criminal acts. This is followed by a transitional phase between 2019 and 2021, during which both clusters (A and B) appear alternately, indicating increased variability but a reduced number of crimes. In the later period, from 2022 to 2024, the low-intensity Cluster A becomes more dominant, indicating a stabilization of the system and a decrease in the overall level of crime. A significantly more complex temporal structure is observed for property crime, as shown in Figure 4. In the initial phase (approximately 2015–2016), Cluster C dominates with the highest intensity of criminal activity, reflecting unstable and elevated dynamic levels. During the period 2017–2019, the middle Cluster B becomes more pronounced, indicating a transitional regime with moderate but still elevated levels of criminal activity. Finally, the period 2020-2024 is characterized by Cluster A, i.e., both time series then attain a stable regime with lower and less variable levels of criminal activity.

Table 3. Descriptive statistics of clusters obtained using the K-means method

Statistics	KD_121		KD_122		KD_205			KD_206		
	Cluster A	Cluster B	Cluster A	Cluster B	Cluster A	Cluster B	Cluster C	Cluster A	Cluster B	Cluster C
Count	52	68	52	68	76	34	10	76	34	10
Mean	66.75	87.26	77.36	113.7	6.118	9.235	11.70	58.93	151.5	289.7
St. Dev.	12.63	10.77	10.67	15.73	3.024	3.718	5.10	19.49	26.23	34.13
Min.	33	66	57	84	0	3	6	25	107	252
Max.	94	108	100	171	14	17	23	102	195	348

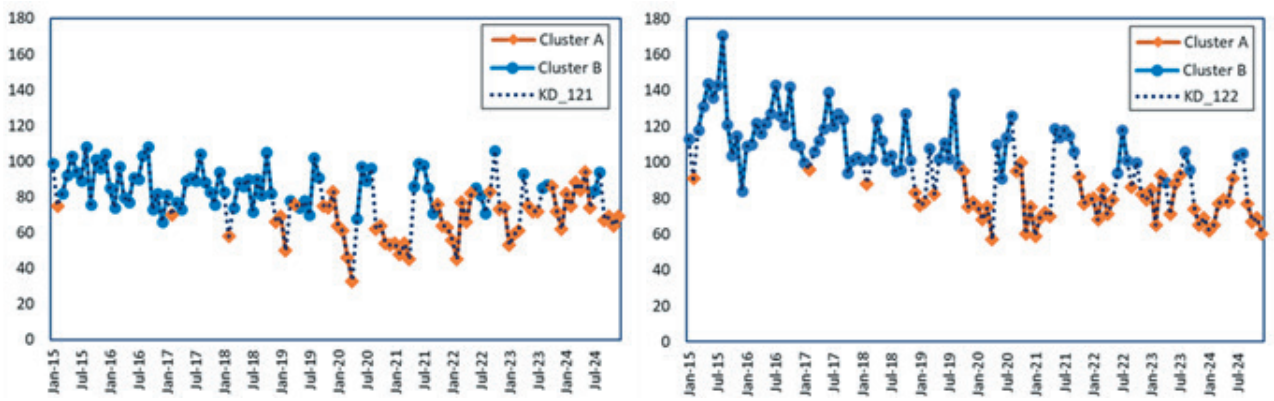

Figure 3. Cluster structure of the temporal dynamics of the number of criminal offenses of bodily harm

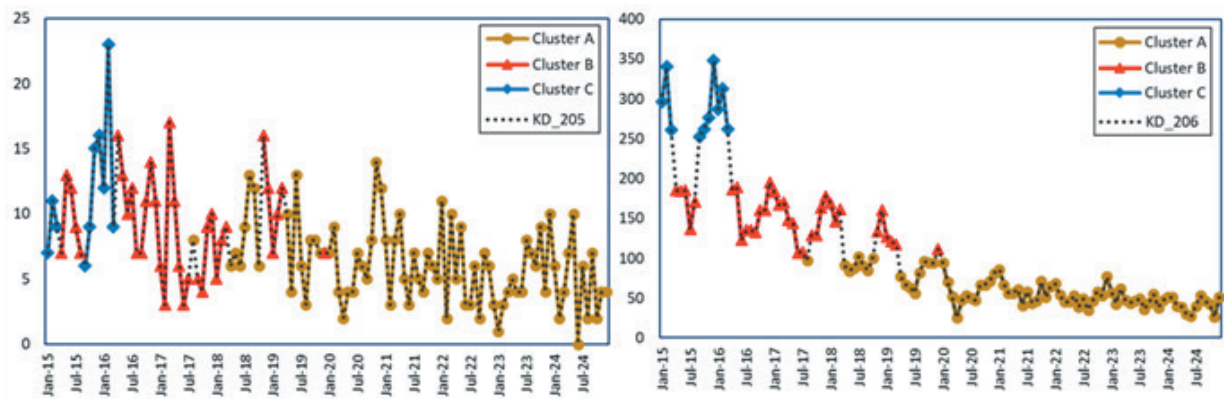

Figure 4. Cluster structure of the temporal dynamics of the number of property crimes

Overall, such dynamics confirm that the observed criminal activities develop through structurally distinct and temporally stable regimes, and not (merely) through random fluctuations.

4. CONCLUSION

This study examined the application of clustering methods in the analysis of crime dynamics, with an emphasis on their software implementation and practical usage. A comparative analysis of four commonly used clustering algorithms was conducted in terms of their efficiency, and then conducted on real bivariate time series. The results thus obtained indicate that clustering performance strongly depends on the structural properties of the data, with the K-Means algorithm providing the most robust results in both observed datasets. In addition to the methodological comparison, this analysis also confirmed the temporal identification of clusters, which correspond to different regimes and levels of intensity of criminal activities. From a practical aspect, these results demonstrate that clustering can serve as an effective analytical tool in security systems, enabling the identification of high-risk periods and data-driven decision-making. To this end, some future research could focus on integrating clustering with stochastic or regime-switching models, as well as extending analysis to high-dimensional data in real time.

5. ACKNOWLEDGEMENTS

The authors sincerely thank the Ministry of Internal Affairs of the Republic of Serbia, whose officially provided the dataset presented in this study.

REFERENCES

[1] J. Marin, G. Guerreros and D. Calderon, "Using Big Data Analytics to Identify Trends and Group Crimes through Clustering," *Int. J. Comput.*, vol. 23, no. 3, pp. 396-406, 2024. doi: 10.47839/ijc.23.3.3658

[2] N. Jabeen and P. Agarwal, "Data Mining in Crime Analysis," in *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, Jaipur, India, 2020. doi: 10.1007/978-981-15-6707-0_10

[3] T. Li, Y. Zhang, J. Zhao and H. Zhang, "Using Data-Mining to Solve Criminal Cases," *Open Access Lib. J.,* vol. 10, e9685, 2023. doi: 10.4236/oalib.1109685

[4] S. Stojičić, V. S. Stojanović, M. Jovanović, D. Joksimović and R. Radovanović, "Bivariate Generalized Split-BREAK Process with Application in Modeling Crime Dynamics," *Mathematics*, vol. 14, no. 5, 754, 2026. doi: 10.3390/math14050754

[5] N. Mitrović, V. S. Stojanović, M. Jovanović and D. Mladjan, "Forensic and Cause-and-effect Analysis of Fire Safety in the Republic of Serbia: An Approach Based on Data Mining," *Fire*, vol. 9, no. 4, p. No. 302, 2025. doi.org/10.3390/fire8080302

[6] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. on Neural Networks,* vol. 16, no. 3, p. 645–678, 2005.doi: 10.1109/TNN.2005.845141

[7] S.-L. Documentation, "Scikit-Learn: Machine Learning in Python," [Online]. Available: https://scikit-learn.org/stable/modules/clustering.html#clustering. [Accessed 07 April 2026].

[8] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *J. Comput. Appl. Math.*, vol. 20, p. 53–65, 1987. doi: 10.1016/0377-0427(87)90125-7

[9] H. Abdalla, "A Brief Comparison of K-means and Agglomerative Hierarchical Clustering Algorithms on Small Datasets," in *International Conference on Wireless Communications*, Berlin, Germany, 2021. doi: 10.1007/978-981-19-2456-9_64

[10] U. Baizyldayeva, R. Uskenbayeva and S. Amanzholova, "Decision Making Procedure: Applications of IBM SPSS Cluster Analysis and Decision Tree," *World App. Sci. J.*, vol. 21, no. 8, p. 1207–1212, 2013. doi: 10.5829/idosi.wasj.2013.21.8.2913