



МЕЖДУНАРОДНА НАУЧНА КОНФЕРЕНЦИЯ „МЕХАТРОНИКА, ЕКО И ЕНЕРГОСПЕСТЯВАЩИ СИСТЕМИ И ТЕХНОЛОГИИ“ - ПЛОВДИВ' 2026

ПРИЛАГАНЕ НА АЛГОРИТЪМА K-MEANS ЗА КЛЪСТЕРИРАНЕ И АНАЛИЗ НА ПРОСТРАНСТВЕНО-ВРЕМЕВИ ТРАНСПОРТНИ ДАННИ

APPLICATION OF THE K-MEANS ALGORITHM FOR CLUSTERING AND ANALYSIS OF SPATIO-TEMPORAL TRANSPORT DATA

Teodora Mecheva

*Technical University of Sofia, branch Plovdiv; teodora.mecheva@tu-plovdiv.bg
Center of competence "Smart mechatronic, eco-and energy-saving systems and technologies"*

Abstract

Clustering temporal traffic data can improve road safety and efficiency in Intelligent Transportation Systems. The aim of the study is to analyse spatio-temporal transport data in order to extract traffic characteristics on the section of St. Petersburg Boulevard in Plovdiv. The present study is part of a larger group of studies that apply conventional methods for clustering data. The data are initially grouped into 24-hour time series for each direction, missing values are filled, and then K-means clustering of speed and number of vehicles passed is applied. The number of clusters is determined by the Silhouette score, and the results are visualized via MDS transformation. The results show that traffic flow in the direction of the Ring Road exhibits greater variability - the traffic volume fluctuates within a wider range throughout the day compared to the traffic heading toward the city Center.

Keywords: K-Means; clustering; traffic patterns; intelligent transport.

INTRODUCTION

In modern urban environments, the efficient management of transportation systems has become one of the most critical challenges for urban planning and sustainable development. The proliferation of digital infrastructure—ranging from GPS-enabled devices to traffic flow sensors—generates vast amounts of data daily. This "Big Data" in transport offers unprecedented opportunities to understand mobility patterns, reduce congestion, and optimize public infrastructure [1-3].

Extracting actionable insights from raw, high-dimensional transport datasets requires robust data mining and machine learning techniques. Among these, cluster analysis stands out as a powerful tool for discovering hidden structures within data without prior labeling. This study focuses on the

application of the K-Means algorithm, a cornerstone of unsupervised learning, due to its efficiency and scalability in processing large-scale spatial and temporal datasets [4-5].

The primary objective of this research is to categorize transportation data into meaningful groups to identify key mobility trends, such as peak travel intervals, and anomalous patterns in the network. The paper details the experimental workflow, including data preprocessing, constructing time series, handling missing values, the selection of the optimal number of clusters using the Silhouette score, and a interpretation of the experimental results of clustering.

METHODS

This section presents the methods used for data processing and analysis. The main stages

of data preparation, time series construction, missing value handling, similarity measurement, clustering, and visualization of results are discussed. In the experimental stage of this study, a variant in which an additional step - data normalization was also investigated. The final result of clustering in this case was that the data formed a dense, homogeneous cluster, which indicates that the shape of the traffic load and the speed is similar. The current variant was chosen in order to be able to see the values of the daily profiles on a real scale. Figure 1 presents a chart of the pipeline. The ovals represent the stages at which data is outputed.

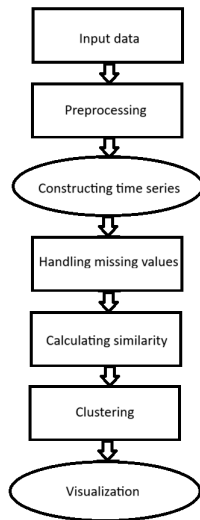


Fig. 1. The pipeline of the used methodology.

The following subsections of this section discuss the stages of the methodology, with the clustering step including an additional stage - selecting the number of clusters.

2.1. Preprocessing and Input Data.

In the experiment is used data for 2022, 2023 and 2024 from the road equipment on St. Petersburg Blvd. of the Technical University of Sofia, Plovdiv Branch in Plovdiv. It was necessary to use such a long period because there are many days for which there are no records. The data contain timestamp, direction and information about the number of vehicles passing and the

average speed. The data are in the following format:

```
date Direction v_count v_vel
1/20/2023 13:00 Ring Road 426 43.93
1/20/2023 13:00 Center 312 48.46
```

2.2. Constructing time series.

The time series are constructed by forming a vector of 24 points for each day, corresponding to an hour of the day. The rows represent a day, and the columns a combination of direction and time. Separate matrices are formed for the two directions - Center and Ring Road, as well as for speed and number of vehicles.

2.3. Handling missing values

Days for which data is not available are not processed. Missing hours are filled by 2D linear interpolation. The value is calculated as a weighted average of the closest observed values - c_1 , c_2 and c_3 , with the weights λ depending on the geometric arrangement of the points.

$$c_{d,h} = \lambda_1 c_1 + \lambda_2 c_2 + \lambda_3 c_3 \quad (1)$$

If interpolation is not possible, the nearest neighbor method is used. The remaining missing values are filled by the median per column [6].

2.4. Calculating similarity between time series.

The similarity metric used in this study is Euclidian distance, defined as:

$$D(x,y) = \sqrt{\sum_{t=1}^T (x_t - y_t)^2} \quad (2)$$

[7].

2.5. Clustering.

K-Means is an iterative clustering algorithm that partitions observations into k clusters by minimizing the intra-cluster

variance in Euclidean space. K-Means uses computed centroids that represent the average of all points in the corresponding cluster. The algorithm is based on minimizing the following objective function:

$$J = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2, \quad (3)$$

where C_i is the set of observations in cluster i , μ_i is the centroid of the cluster, $\|x - \mu_i\|^2$ is the squared Euclidean distance between an observation and the centroid. The procedure begins with an initialization of the k centroids, after which two main steps are repeated:

- *Assignment step*: each observation is associated with the closest centroid according to the Euclidean distance.
- *Update step*: centroids are recalculated as the arithmetic mean of all points in the corresponding cluster [8].

2.6. Determining the optimal number of clusters

To select the optimal number of clusters k , the Silhouette metric is used. It evaluates the quality of clustering by the ratio between the intra-cluster and inter-cluster distances:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4)$$

where $a(i)$ is mean distance to the cluster and $b(i)$ is mean distance to the nearest cluster. The procedure involves testing values of k in the range from 2 to 5 and calculating the Silhouette coefficient for each value; selecting k at which the coefficient is maximum. The advantage of Silhouette coefficient is that it accesses the quality of clustering without the need for predefined labels, taking into account the compactness of clusters and their separability, which allows for an objective assessment of the results [9].

2.7. Visualization and interpretation

The results are visualized using the Multidimensional Scaling (MDS) method, which projects the day profiles into a three-dimensional space while preserving the relative distances between objects. The method provides an intuitive representation of clusters, facilitates the detection of structures and outliers, and aids to the interpretation of results [10].

RESULTSSSSS

The aim of current study is to identify typical traffic patterns by applying K-Means clustering on daily traffic profiles for different days and directions.

3.1. Silhouette score.

Figures 2 and 3 show the Silhouette score for the number of vehicles and speed towards the Center. As can be seen from the graphs, the optimal number of clusters in both cases is 2.

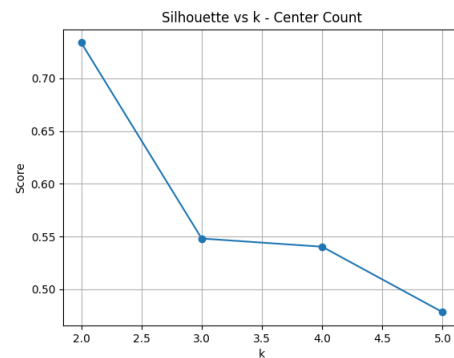


Fig. 2. Silhouette score for number of vehicles for direction Center.

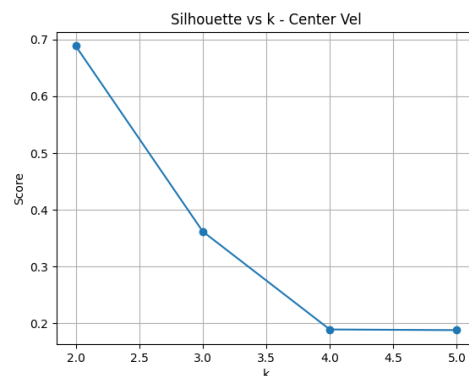


Fig. 3. Silhouette score for velocity for direction Center.

Figures 4 and 5 show the Silhouette score for the number of vehicles and the speed towards the Ring Road. As can be seen from the graph, the optimal number of clusters is 3 for the number of vehicles and 2 for the velocity.

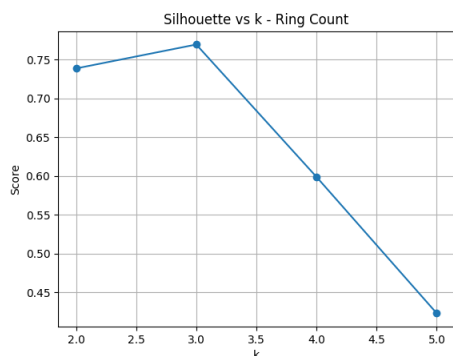


Fig. 4. Silhouette score for number of vehicles for direction Ring Road.

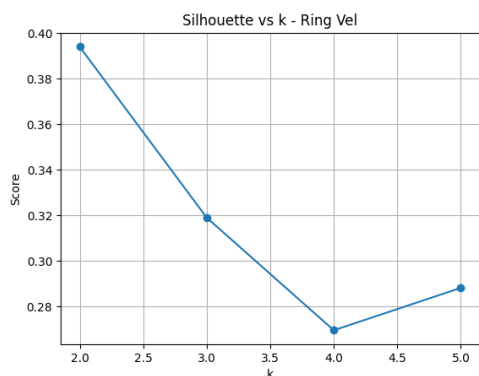


Fig. 5. Silhouette score for velocity for direction Ring Road.

Based on the graphics can be concluded that traffic in direction Center is more homogeneous, while in direction Ring Road - more diverse.

3.2. 3D graphics of MDS.

K-Means clustering results of Euclidian distance are visualized using 3D MDS projection of daily traffic time series. Each point represents a daily observation, colored according to cluster membership. In the plots, individual clusters stand out clearly, with each color representing a different cluster.

Figure 6 shows 3D MDS graphic for number of vehicle for direction Center.

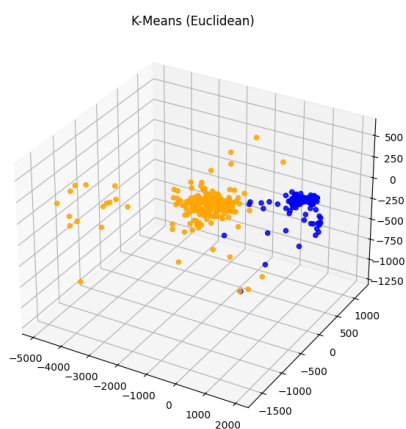


Fig. 6. 3D MDS graphic for number of vehicles for direction Center.

On the graph of number of vehicles for direction center the two clusters are clearly separated. C0 (blue) is smaller and probably covers days with more stable profiles throughout the day, while C1 (orange) is larger and represents days with higher variability in the number of cars.

Figure 5 shows 3D MDS cluster profiles of the average speed for the direction center. Each profile reflects the average speed dynamics within a 24-hour cycle for a group of days with similar behavior.

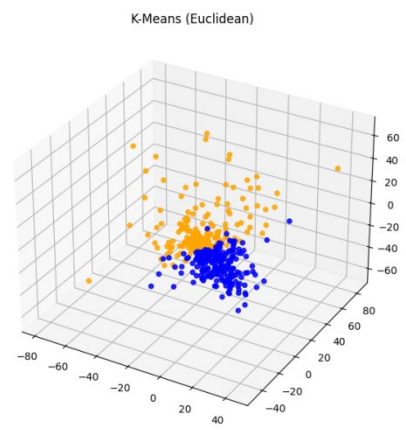


Fig. 7. 3D MDS graphic velocity for direction Center.

In the velocity plot for the Center direction of k-means, the data appears to be distributed in a compact cluster with evenly spaced points around it. K-means splits this cluster into 2 through the middle.

Figure 8 shows 3D MDS graphic of number of vehicles in direction Ring Road.

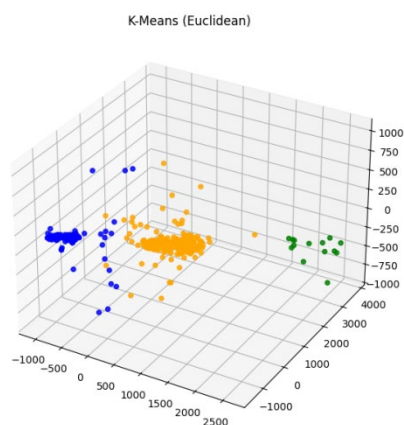


Fig. 8. 3D MDS graphic for number of vehicles for direction Ring Road.

The vehicles count graph for direction Ring Road clearly distinguishes 3 clusters: C0 (blue) – a compact group on the left of the graph. These are probably days with more stable and regular traffic, with less variation in the number of cars during the day. C1 (orange) – more scattered points in the central part of the graph. This indicates days with high variability in the number of cars – for example, peak days or days with accidents/events; C2 (green) – a separate, small cluster on the right. These are anomalous or rare cases, for example, holidays, special events etc.

Figure 9 shows 3D MDS graphic of velocity in direction Ring Road.

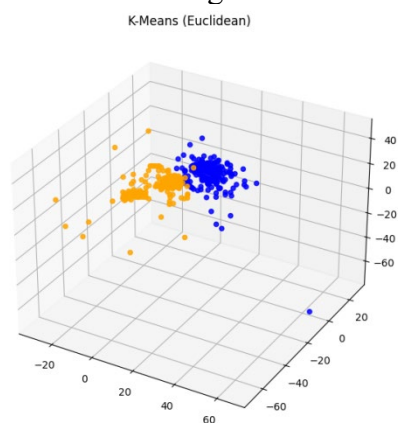


Fig. 9. 3D MDS graphic for velocity for direction Ring Road.

The speed graph of the K-Means algorithm presents 2 equally large clusters, with the orange one being more scattered, likely including holidays and days with incidents, while the blue one is more compact, reflecting days with regular traffic.

CONCLUSIONS

The article presents a method for analyzing daily traffic profiles using a combination of Euclidian distance and K-Means clustering. The method allows detection of typical traffic patterns.

The results obtained from the clustering provide information about the nature of the daily traffic profiles in the studied section - the two lanes of St. Petersburg Blvd. in Plovdiv.

Additionally, experiments with normalization show that it leads to a more compact cluster, at the cost of losing information about the real values. Therefore, an approach without normalization was chosen, which allows for better interpretation of the results in the context of real traffic.

The results show a difference in the homogeneity of traffic in both directions. In the direction towards the Center, homogeneous traffic is observed, which is confirmed by the number of selected clusters - 2, both for the number of vehicles and for speed. In the direction of the Ring Road, the clusters for the number of vehicles are 3, and the clusters for speed - 2. This indicates greater variations in traffic, probably influenced by factors such as outflow from the city, events or seasonal effects.

MDS visualizations further confirm these observations. Towards the Center, the clusters are clearly demarcated and compact, indicating stability of traffic patterns. Along the Ring Road, the clusters are more dispersed, reflecting higher dynamics and the presence of atypical days.

Despite the results, the study has some limitations. First, the presence of missing data necessitates the use of interpolation, which



may introduce inaccuracies. Furthermore, the analysis is based on data from a single location, which limits the possibility of direct generalization to other urban environments.

Future work will focus on comparing the results obtained with results from other conventional clustering methods, such as K-Means, Hierarchical Clustering, and Density based clustering. Another possibility for future studies is to enrich the studied data with meteorological data.

Acknowledgments

This research was funded by the European Regional Development Fund within the OP “Research, Innovation and Digitalization Programme for Intelligent Transformation 2021-2027”, Project CoC “Smart Mechatronics, Eco- and Energy Saving Systems and Technologies”, No. BG16RFPR002-1.014-0005.

REFERENCES

- [1]Zhang J, Wang FY, Wang K, Lin WH, Xu X, Chen C. Data-Driven Intelligent Transportation Systems: A Survey, IEEE Transactions on Intelligent Transportation Systems. 2011, 12, no. 4, pp. 1624-1639, DOI: 10.1109/TITS.2011.2158001.
- [2]Chowdhury M, Dey K, Apon A. Data analytics for intelligent transportation systems. 2nd ed.; Elsevier. 2024.
- [3]Attari MYN, Ejlaly B, Heidarpour H, Ala A. Application of data mining techniques for the investigation of factors affecting transportation enterprises. IEEE Transactions on Intelligent Transportation Systems, 2022, 23, 7, pp. 9184-9199, DOI: 10.1109/TITS.2021.3092348.
- [4]Cebecauer M, Jenelius E, Gundlegård D, Burghout W. Revealing Representative Day-types in Transport Networks Using Traffic Data Clustering. Journal of Intelligent Transportation Systems, 2024, 28, 5, pp. 695-718, DOI: 10.1080/15472450.2023.2205020.
- [5]Zhou F, Yao J, Yin H. Identifying the Passenger Transport Corridors in an Urban Rail Transit Network Based on OD Clustering. Sustainability, 2025, 17, 9127. DOI: 10.3390/su17209127
- [6]Interpolation in 2D. Available online: <https://users.tricity.wsu.edu/~hudson/Teaching/EE221/Text/16%20Interpolation%20in%20D.pdf> (accessed on 13.05.2026).
- [7]Euclidian distance. Available online: <https://www.sciencedirect.com/topics/engineering/euclidean-distance> (accessed on 13.05.2026).
- [8]Sinaga KP, Yang MS. Unsupervised K-Means Clustering Algorithm, IEEE Access, 2020, 8, pp. 80716-80727, DOI: 10.1109/ACCESS.2020.2988796.
- [9]What is Silhouette Score. Available online: <https://www.geeksforgeeks.org/machine-learning/what-is-silhouette-score/> (accessed on 13.05.2026).
- [10]Davison ML, Sireci SG. Multidimensional Scaling, Handbook of Applied Multivariate Statistics and Mathematical Modeling, Academic Press, 2000, p.p. 323-352, DOI: <https://doi.org/10.1016/B978-012691360-6/50013-6>.