K-means clustering method in organizing passenger transportation in a smart city

Yurii Matseliukh[†], Vasyl Lytvyn[†] and Myroslava Bublyk[*,†]

Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine

Abstract

An analysis of a heterogeneous data set on the duration of electric transport races in an average-sized city was conducted. The possibilities of using the K-means clustering method in organizing passenger transportation in a smart city were studied, including the analysis of passenger flows by passenger types, identification of transport hotspots, identification of inefficient routes or their sections, and construction of dynamic models for predicting changes in flows, as well as the features of its application for optimizing the operation of the transport system were determined. Data analysis revealed sections of routes with different intensity of transport flows, depending on their location in urban areas, seasonality, events in the city, changes in transport flows due to detours, repair work, etc. An algorithm for selecting a clustering method was proposed based on clustering quality assessment metrics, including the elbow method, the silhouette method, and the Calinski-Harabasz index. It is recommended to use clustering to create routes with reduced waiting times, fewer transfers, and compliance with passenger needs.

Keywords

passenger transportation; smart city; clustering analysis; K-means method; systems analysis

1. Introduction

The problem of organizing passenger transportation in a smart city is closely related to the search for effective methods and tools that ensure optimal dynamic interaction with vehicles when organizing passenger transportation. When organizing passenger transportation, it is important to consider various data, such as traffic routes, waiting times, duration of schedule execution, traffic changes, vehicle load, weather conditions, environmental efficiency, individual needs of passengers, etc. The collected large data sets require proper storage, appropriate analysis, effective grouping, and high-quality clustering.

The relevance of the problem is declared at the global level in many documents (agreements, communiqués, declarations, etc.). International organizations such as the UN and the EU actively support initiatives to reduce greenhouse gas emissions in urban agglomerations. Today, the problem of organizing low-carbon passenger transportation is extremely relevant due to global warming and the growing need for environmentally friendly vehicles. A smart city that uses modern technologies and management systems can become an effective tool for achieving these goals.

Solving this problem is impossible without using the methods and tools of the theory of systems analysis and information systems tools. From the point of view of systems analysis and information systems, the problem of organizing low-carbon passenger transportation is an integral part of complex transport flow management systems. A systems approach allows us to consider all components of the system (vehicles, infrastructure, passengers) as interconnected elements that influence each other. Information systems play a key role in collecting, processing, and analyzing

ISW-2025: Intelligent Systems Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

indeed.post@gmail.com (Y. Matseliukh); vasyl.v.lytvyn@lpnu.ua (V. Lytvyn); my.bublyk@gmail.com (M. Bublyk)

0000-0002-1721-7703 (Y. Matseliukh); 0000-0002-9676-0180 (V. Lytvyn); 0000-0003-2403-0784 (M. Bublyk)

data to optimize routes and waiting times for transport, as well as for monitoring environmental indicators.

The problem has broad practical significance, which grows every year and each time with the introduction of modern systems of dynamic interaction of passengers with vehicles in a smart city, among which we highlight the following: reducing carbon emissions, improving the quality of life, saving resources and forming a smart urban infrastructure. The collected data sets require proper and high-quality analysis, which cannot be carried out without the use of effective clustering algorithms during route optimization, the introduction of environmentally friendly vehicles on routes, as well as during the effective organization of transportation aimed at reducing waiting times and vehicle load, improving passenger comfort, reducing greenhouse gas emissions, reducing transportation costs, optimizing the use of transport infrastructure and energy-efficient technologies, etc. In general, solving this problem makes a significant contribution to the development of smart cities that use modern technologies to improve the quality of life of residents. Therefore, finding ways to apply clustering methods, including the K-means method, in the organization of passenger transportation in a smart city is an important component of the general problem of developing methods and means of dynamic interaction of passengers with vehicles and is of significant practical importance for the development of low-carbon passenger transportation in public transport of large, medium and small cities. The object of research is the process of clustering data sets on the organization of passenger transportation in a smart city. The subject of research is the principles of optimizing passenger transportation in public transport and improving the implementation of their transportation schedules.

2. Well-known studies clustering methods in organizing passenger transportation in a smart city

Having conducted a detailed review of cluster analysis methods and tools [1, 2] used for modelling [3], route optimization [4], big data analysis [5] and their clustering in order to develop adaptive algorithms for organizing a transport network in a smart city [6, 7], we see that special attention is paid to decision-making models for optimizing passenger flows, taking into account modern approaches based on both bottom-up (agglomerative) [8, 9] and top-down (divisive) [10] clustering methods, as well as distributive [11], fuzzy clustering [12], DBSCAN [13] and self-organizing maps [14].

Modern research in the field of systems analysis [15-17] confirms that the integration of clustering methods into the transport management system contributes to the creation of adaptive, efficient and low-carbon transportation models [18, 19]. The development of the fundamental foundations of clustering methods in the organization of passenger transportation, the development of modern information and communication systems for passenger transportation by public transport are the subjects of research by well-known scientists both in Ukraine and abroad. Among the researchers whose contribution contributed to the development of theoretical foundations and practical experience in the application of cluster data analysis in the organization of passenger transportation based on the concept of a smart city, it is appropriate to note such representatives as: Bezdek J. [20], Bublyk M. [21, 22], Esther M. [23], Jane A. [24] Kohonen T. [25], Koshtura D. [26], Lytvyn V. [27], Lov A. [28], Nat N. [29], Sun L. [30], Tibshirani R. [31].

Among the main clustering methods used for big data analysis in the field of organizing passenger transportation in a smart city [32-36], it is necessary to consider the methods of hierarchical clustering, partitioning clustering, density-based clustering, grids, artificial neural networks. They provide an opportunity not only to understand the individual needs of passengers and patterns of service consumption but also contribute to the optimization of resources to meet these needs. To identify which clustering methods are used for organizing passenger flows, a comparison of the main clustering methods in the field of organizing passenger transportation in a smart city was carried out [1-36].

In the case of hierarchical clustering, agglomerative or, as they are otherwise called, bottom-up methods assume that each element in the data set is a separate cluster. The process of merging the two closest clusters into one occurs according to certain rules (according to a specified metric) until only one cluster is formed. Bottom-up methods are used to organize passenger flows when it is necessary to determine the structure of routes, and the merging occurs according to similar route sections or territories. In divisive or, otherwise, top-down methods, on the contrary, start with one cluster that includes all the data and divide it into smaller clusters. They are used to organize passenger flows when it is necessary to allocate separate routes or their separate sections (races, zones) with different passenger flow intensity to optimize services in specific areas.

In the case of divisive clustering, the K-means method divides the data into clusters by finding the midpoint in each cluster, repeating this process until a stable distribution is achieved. This method necessarily performs an exact data distribution, where each object belongs to only one class. However, it is poorly adapted to data with a complex distribution or with qualitative characteristics. The K-means method is useful for zoning the transport network, for example, when optimizing routes by demand zones. The K-medoid method identifies the centres of clusters that have the greatest possible separation from the total passenger flow. A distinctive feature of the K-medoid method is its greater resistance to noise. It is used to optimize the operating schedule of vehicles, determine the optimal location of stops on routes, for example, metro stations, which can cover the greatest number of passenger needs. The fuzzy C-means clustering method helps to create more flexible and adaptive passenger flow management strategies, which is important in modern urban environments with the dynamic nature of demand for transport services. The fuzzy C-means clustering method differs from the traditional K-means method in that it allows elements to belong to several clusters simultaneously with different degrees of membership. This means that each element has a certain probability of belonging to each of the k clusters. This method uses a membership function that determines the degree of membership of an element to each of the possible clusters. The goal of the optimal distribution is to minimize a function that considers both the distance to the centres of the clusters and the degrees of membership. It gives the best results in complex systems with high ambiguity and overlap between data. It is used to create more flexible and adaptive urban transport zones, where passengers belong to several zones at the same time, considering the unpredictability of demand, for example, changes in passenger flow during the day or under different weather conditions. It reveals patterns invisible to other methods, which can be important in developing optimization strategies. The use of C-means provides a more accurate picture of the segmentation of transport users, reduces the risk of excess or insufficient volumes of services on certain sections of the transport network. It is also used to predict mixed needs and their impact on the distribution of clusters, thereby improving strategic decision-making in transport organization.

In density-based classification, the most well-known method is DBSCAN, which is most often used to analyse passenger flows with various densities on different routes and to identify clusters considering time dynamics. In grid-based classification, the STING method divides the data into smaller groups for analysis to create models of time intervals and densities. Among the artificial neural networks used for clustering, the most common is the self-organizing map method, which allows creating dynamic maps of passenger flows considering time intervals and densities on different routes but requires a significant amount of data for training the neural network, pre-normalization of the data and division into smaller groups for analysis.

Therefore, each clustering method has its own advantages and can be applied in different scenarios for the effective organization and optimization of passenger transportation in a smart city. The choice of method depends on the specifics of the task, data characteristics and goals set when analyzing transport and passenger flows. Bottom-up clustering methods are used when it is necessary to determine the structure of routes, combine similar sections of routes or territories adjacent to the route. Top-down methods are used when it is necessary to identify individual routes or their individual sections (races, zones) with different intensity of passenger flows in specific areas. The K-means distributive clustering method is useful for zoning the transport network when optimizing routes by the duration of the races, by demand zones, etc. The K-medoid distributive clustering method, due to its greater resistance to noise, is used to optimize the operating schedule of vehicles, determine the optimal location of stops on routes to meet passenger needs. Fuzzy C-means clustering method – when developing more flexible and adaptive passenger flow management strategies in modern urban environments with a dynamic nature of demand for transport services. The DBSCAN method is most often used to analyze passenger flows with various densities on different routes and for clustering considering time dynamics. The self-organizing map method – for creating dynamic maps of passenger flows considering time intervals and densities on different routes.

A smart city generates a huge amount of data from various sources, such as GPS systems, mobile applications, sensors, social networks and video surveillance. Processing such large volumes of data is critically important for making informed decisions in the field of passenger transportation. Passenger flows are constantly changing depending on the time of day, day of the week, weather, social events, etc. Clustering methods allow you to identify key groups or patterns in such flows to better understand their nature. Clustering methods are the basis of many modern artificial intelligence algorithms. The use of intelligent transport systems (ITS) [37-41] requires complex analysis and modeling algorithms to identify optimal routes and manage the transport network. Effective data clustering allows you to optimize routes, reduce downtime and total emissions, which is important in the context of combating climate change. Such grouping is extremely important for the development of adaptive route optimization algorithms, as it allows you to effectively allocate transport resources, reduce waiting times and minimize emissions of carbon compounds. As noted by Bublyk M. [42], the concept of smart specialization for the transformation of the Ukrainian economy includes not only the optimization of the economic activities of transport companies, but also the transition to a green economy, where a significant role is played by reducing CO2 emissions through the introduction of innovative solutions in the transport industry. The basis of innovative models for reducing emissions into the atmosphere is the concept of technosoliton, developed by Bublyk M. [43, 44], where the damage and losses in highly polluting sectors of the economy, which have remained transport for many years, were assessed. This concept is of particular importance in the development of strategies for organizing passenger transportation, since route optimization using the K-means method allows not only to improve the quality of service, but also to contribute to the reduction of emissions into the atmosphere, which is crucial for achieving sustainable development goals [45-48].

Summarizing the above analysis of recent studies on the problem of applying clustering methods, including the K-means method, in the organization of passenger transportation in a smart city, today the still previously unsolved part of the general problem is methods for determining patterns of passenger flows, optimizing transport routes and increasing network efficiency in real time. to improve passenger comfort, reduce greenhouse gas emissions, reduce transportation costs, optimize the use of transport infrastructure and energy-efficient technologies, etc. Insufficient attention has also been paid to finding effective ways to apply clustering algorithms during route optimization, when implementing environmentally friendly routes, as well as during effective transportation organization aimed at reducing passenger waiting time, in general, or vehicle congestion, in particular. This indicates the need for scientific research in this direction, namely, to study the possibilities of using the K-means clustering method in organizing passenger transportation in a smart city and to determine the features of their application for optimizing the organization of passenger transportation by public transport, which is the purpose of this work.

The article solves the following tasks: studying the features of clustering methods and their metrics in organizing passenger transportation in a smart city; analyzing a large-scale heterogeneous data set on the duration of electric transport trips within an average-sized city; developing a simple and most effective algorithm for choosing a clustering method based on metrics for assessing the quality of clustering data collected from the infrastructure of passenger transportation by public transport in smart cities.

3. Materials and methods

Among the methods used for data analysis, comparison and grouping was the cluster analysis method, namely the K-means method. The key feature of the application of data clustering methods is the choice of distance metric, which among many other different indicators should be chosen based on its relevance to a specific example. In our case, this is a study of a dataset of low-carbon vehicle traffic on a single route in an average-sized city, so the Euclidean distance was used, which is described by the formula (1):

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{1}$$

where $x = (x_1, x_2, \dots, x_n)$ – characteristic vector of point *x*, which contains *n* components; $y = (y_1, y_2, \dots, y_n)$ – characteristic vector of point *y*, which contains *n* components; *i* — index for each of the attributes (attribute number).

The choice of the number of clusters directly affects the quality of clustering, so it is important to choose the optimal number of clusters for a given data set. In our case, it was the elbow method, which involves analyzing the dependence of the sum of squared distances *SSE(k)* between points and the centers of their clusters on the value *k*. The sum of squared distances is calculated by formula (2).

$$SSE(k) = \sum_{i=1}^{k} \sum_{x_j \in C_i} |\, x_j - \mu_i|^2, \tag{2}$$

where *SSE(k)* is the sum of squared distances; *k* is the number of clusters; x_j is the point in the data set belonging to cluster C_j ($x_j \in C_j$); μ_i is the center of the *i*-th cluster.

The K-means algorithm is one of the most common clustering methods used to partition a data set into k clusters. It works iteratively, minimizing the sum of the squares of the distances of points to the cluster centers. The K-means algorithm consists of 4 stages: initialization, assigning points to clusters; updating the cluster centers and checking the stopping criterion.

Initialization consists of selecting *k* initial cluster centers $\mu_1, \mu_2, \dots, \mu_K$, randomly or using special strategies, and assigning points to clusters. Each data element xi is assigned to the nearest center μ_K of the cluster according to the criterion of the smallest distance, for the calculation of which formula (3) is used:

$$c_i = \arg\min_k ||\, x_i - \mu_k||^2, \tag{3}$$

where c_i is the cluster to which point x_i is assigned; $\|x_i - \mu_k\|^2$ is the square of the Euclidean distance.

The cluster centres are updated each time a new point is added to the cluster. Each point allocated to the cluster to which it is closest according to the criterion of the smallest distance is considered sequentially. After all points are assigned to clusters, the new centre μ_k of each cluster is calculated as the average value of all points belonging to it (4):

$$\mu_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i, \tag{4}$$

where S_k is the set of points belonging to the *k*-th cluster.

The centroid is sequentially recalculated each time a new point is added to the cluster, i.e. when the division of points into clusters changes, then the coordinates of the centroids change to new ones. To be sure that each point has been optimally assigned to the correct cluster, the distance of each cluster point to the centre of its own cluster and to the centre of the nearest opposite cluster is compared according to formula (5):

$$C_{opt} = \sum_{k=1}^{K} \sum_{x_i \in S_k} \operatorname{argmin}_k |\, x_i - \mu_k|^2. \tag{5}$$

The iterative transfer of points continues with each new division into clusters until the last division is recognized as the result of clustering.

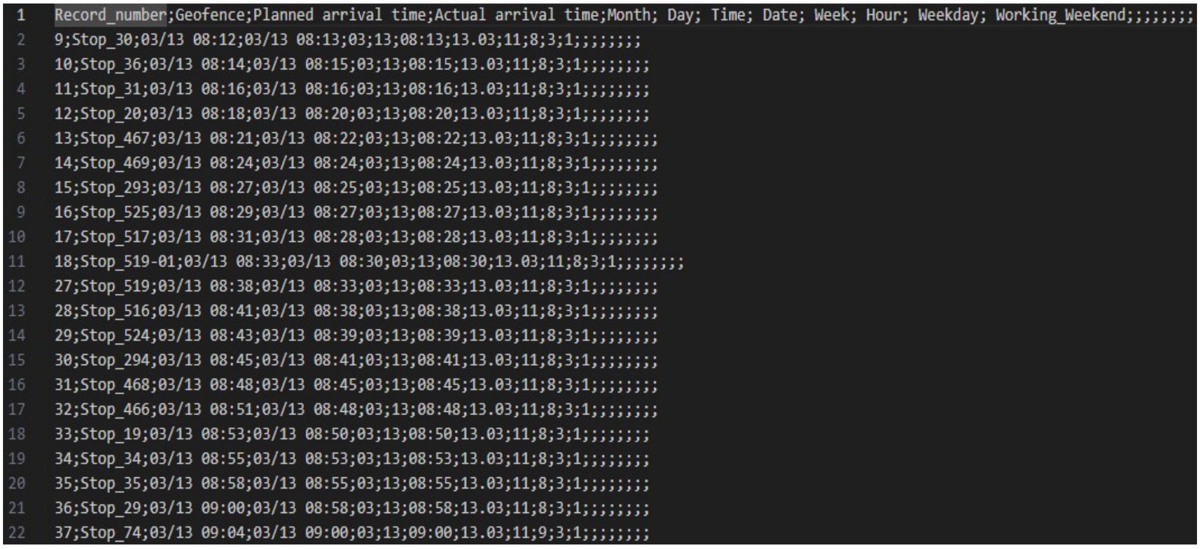
Checking the stopping criterion indicates that the algorithm is stopped. The clustering algorithm is terminated if the cluster centres stop changing or the changes are insignificant. Otherwise, we return to step 2. It is quite possible that the K-means algorithm will not find a final solution. In this case, it is advisable to stop the algorithm after the algorithm reaches the previously selected maximum iteration value. Thus, the K-means algorithm iteratively improves the distribution of points between clusters by reducing the value of the loss function.

4. Results

4.1. An analysis of a heterogeneous data set on the duration of electric transport races in an average-sized city

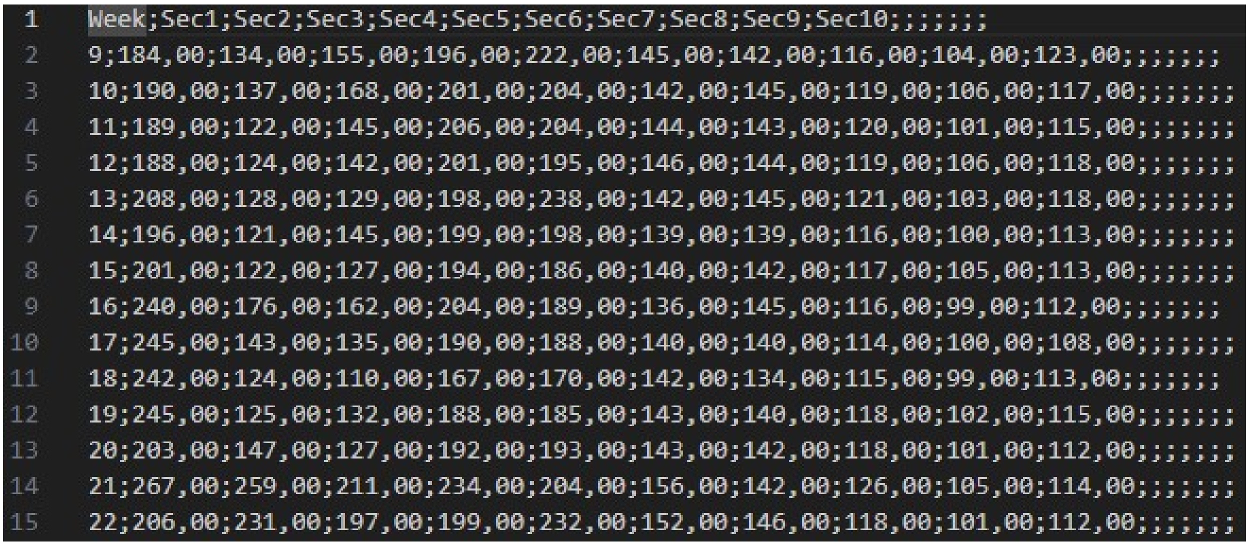
In our case, a dataset on the duration of low-carbon public transport trips within an average-sized city was used for research. Here, we analyze in real time during the study period the duration of each trip by each vehicle on the route within the same route. The data structure has the following form: Record number; Geozone; Planned arrival time; Actual arrival time; Month; Day; Time; Date; Week; Hour; Day of the week; Working / non-working. The total volume is 890999 records. After cleaning the data from empty cells, separating incomplete, additional, erroneous and information falling out of the general time frame of the duration of operation of vehicles on the route, 716960 records remained in the dataset, where the appearance of the first 21 records is shown in Fig. 1.

As a result of the analysis of the collected data, the duration of each leg of the journey was aggregated within each working hour by vehicles within each day for each week during the study period. Since the duration of the journey within one hour by all vehicles on the route within one route for each week during the year is also characterized by a complex, heterogeneous and large-scale structure, therefore it requires appropriate processing before starting the cluster analysis. At the last stage, after cleaning and grouping the data, a matrix of passenger transport schedules for each of the 10 legs was obtained with the average values of the duration of the leg for each week during the study period. As an example, Fig. 2 shows the duration of the leg on average per day during each week of the study period.



```
Record_number;Geofence;Planned arrival time;Actual arrival time;Month; Day; Time; Date; Week; Hour; Weekday; Working_Weekend;;;;;;;
9;Stop_30;03/13 08:12;03/13 08:13;03;13;08:13;13.03;11;8;3;1;;;;;;;;
10;Stop_36;03/13 08:14;03/13 08:15;03;13;08:15;13.03;11;8;3;1;;;;;;;;
11;Stop_31;03/13 08:16;03/13 08:16;03;13;08:16;13.03;11;8;3;1;;;;;;;;
12;Stop_20;03/13 08:18;03/13 08:20;03;13;08:20;13.03;11;8;3;1;;;;;;;;
13;Stop_467;03/13 08:21;03/13 08:22;03;13;08:22;13.03;11;8;3;1;;;;;;;;
14;Stop_469;03/13 08:24;03/13 08:24;03;13;08:24;13.03;11;8;3;1;;;;;;;;
15;Stop_293;03/13 08:27;03/13 08:25;03;13;08:25;13.03;11;8;3;1;;;;;;;;
16;Stop_525;03/13 08:29;03/13 08:27;03;13;08:27;13.03;11;8;3;1;;;;;;;;
17;Stop_517;03/13 08:31;03/13 08:28;03;13;08:28;13.03;11;8;3;1;;;;;;;;
18;Stop_519-01;03/13 08:33;03/13 08:30;03;13;08:30;13.03;11;8;3;1;;;;;;;;
27;Stop_519;03/13 08:38;03/13 08:33;03;13;08:33;13.03;11;8;3;1;;;;;;;;
28;Stop_516;03/13 08:41;03/13 08:38;03;13;08:38;13.03;11;8;3;1;;;;;;;;
29;Stop_524;03/13 08:43;03/13 08:39;03;13;08:39;13.03;11;8;3;1;;;;;;;;
30;Stop_294;03/13 08:45;03/13 08:41;03;13;08:41;13.03;11;8;3;1;;;;;;;;
31;Stop_468;03/13 08:48;03/13 08:45;03;13;08:45;13.03;11;8;3;1;;;;;;;;
32;Stop_466;03/13 08:51;03/13 08:48;03;13;08:48;13.03;11;8;3;1;;;;;;;;
33;Stop_19;03/13 08:53;03/13 08:50;03;13;08:50;13.03;11;8;3;1;;;;;;;;
34;Stop_34;03/13 08:55;03/13 08:53;03;13;08:53;13.03;11;8;3;1;;;;;;;;
35;Stop_35;03/13 08:58;03/13 08:55;03;13;08:55;13.03;11;8;3;1;;;;;;;;
36;Stop_29;03/13 09:00;03/13 08:58;03;13;08:58;13.03;11;8;3;1;;;;;;;;
37;Stop_74;03/13 09:04;03/13 09:00;03;13;09:00;13.03;11;9;3;1;;;;;;;;
```

Figure 1: View of the dataset on the duration of each journey by each vehicle in real time during the study period (authors' calculation based on collected data).



```
Week;Sec1;Sec2;Sec3;Sec4;Sec5;Sec6;Sec7;Sec8;Sec9;Sec10;;;;;;;
9;184,00;134,00;155,00;196,00;222,00;145,00;142,00;116,00;104,00;123,00;;;;;;;
10;190,00;137,00;168,00;201,00;204,00;142,00;145,00;119,00;106,00;117,00;;;;;;;
11;189,00;122,00;145,00;206,00;204,00;144,00;143,00;120,00;101,00;115,00;;;;;;;
12;188,00;124,00;142,00;201,00;195,00;146,00;144,00;119,00;106,00;118,00;;;;;;;
13;208,00;128,00;129,00;198,00;238,00;142,00;145,00;121,00;103,00;118,00;;;;;;;
14;196,00;121,00;145,00;199,00;198,00;139,00;139,00;116,00;100,00;113,00;;;;;;;
15;201,00;122,00;127,00;194,00;186,00;140,00;142,00;117,00;105,00;113,00;;;;;;;
16;240,00;176,00;162,00;204,00;189,00;136,00;145,00;116,00;99,00;112,00;;;;;;;
17;245,00;143,00;135,00;190,00;188,00;140,00;140,00;114,00;100,00;108,00;;;;;;;
18;242,00;124,00;110,00;167,00;170,00;142,00;134,00;115,00;99,00;113,00;;;;;;;
19;245,00;125,00;132,00;188,00;185,00;143,00;140,00;118,00;102,00;115,00;;;;;;;
20;203,00;147,00;127,00;192,00;193,00;143,00;142,00;118,00;101,00;112,00;;;;;;;
21;267,00;259,00;211,00;234,00;204,00;156,00;142,00;126,00;105,00;114,00;;;;;;;
22;206,00;231,00;197,00;199,00;232,00;152,00;146,00;118,00;101,00;112,00;;;;;;;
```

Figure 2: Average daily duration of each journey by vehicle for each week (authors' calculation based on collected data)

Using the Python pyplot tools from the matplotlib library, we visualize the average daily duration of the Sec1 race by vehicles for each week, constructing the graph shown in Fig. 3:

```
from matplotlib import pyplot as plt
df['Sec1'].plot(kind='line', figsize=(8, 4), title='Sec1')
plt.gca().spines[['top', 'right']].set_visible(False)
```

Using the same tools (pyplot from the matplotlib library) Python, we visualize the average daily duration of the race by vehicles for each week for the remaining 9 races, where Fig. 4 shows the graph for race Sec2.

```
from matplotlib import pyplot as plt
df['Sec2'].plot(kind='line', figsize=(8, 4), title='Sec2')
plt.gca().spines[['top', 'right']].set_visible(False)
```

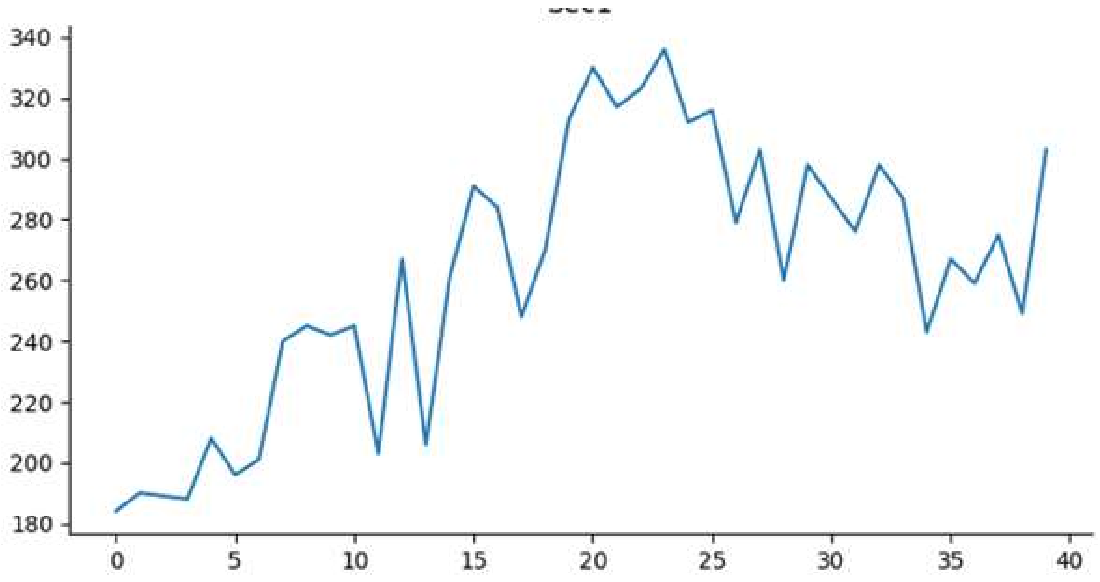


Figure 3: Average daily duration of Sec1 leg by vehicles for each week (authors' calculation based on collected data)

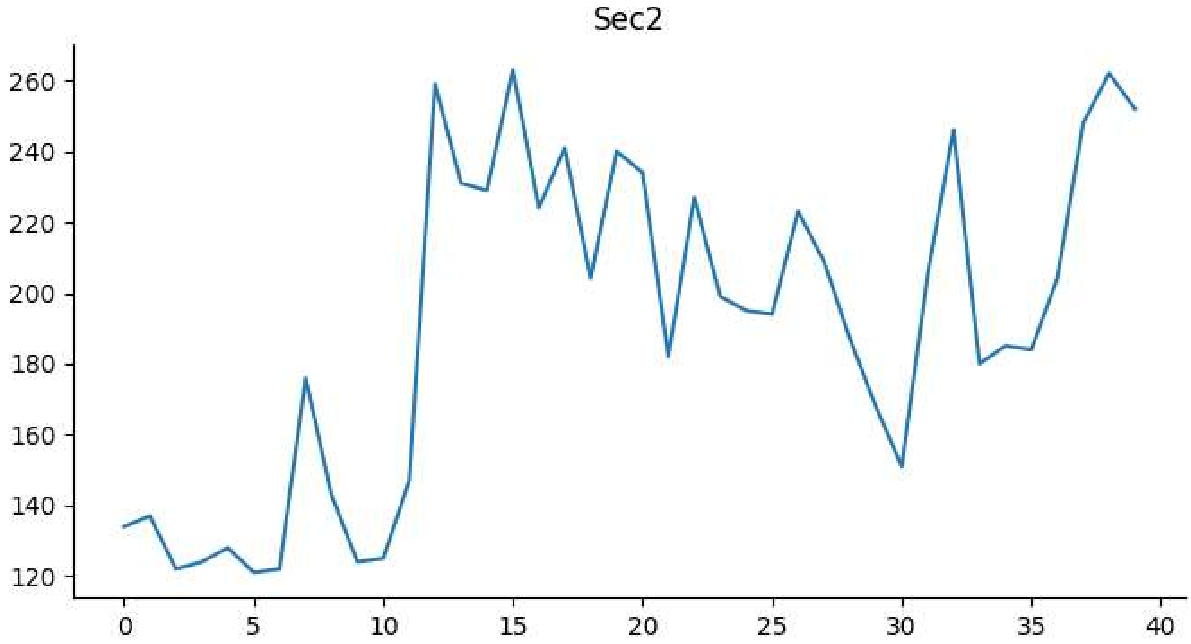


Figure 4: Average daily duration of Sec2 race by vehicles for each week (authors' calculation based on collected data)

Analyzing the structure of the dataset using Python tools, the frequency characteristics of the dataset for each of the races, where Fig. 5 shows the result for race Sec1.

```
from matplotlib import pyplot as plt
df['Sec1'].plot(kind='hist', bins=20, title='Sec1')
plt.gca().spines[['top', 'right',]].set_visible(False)
```

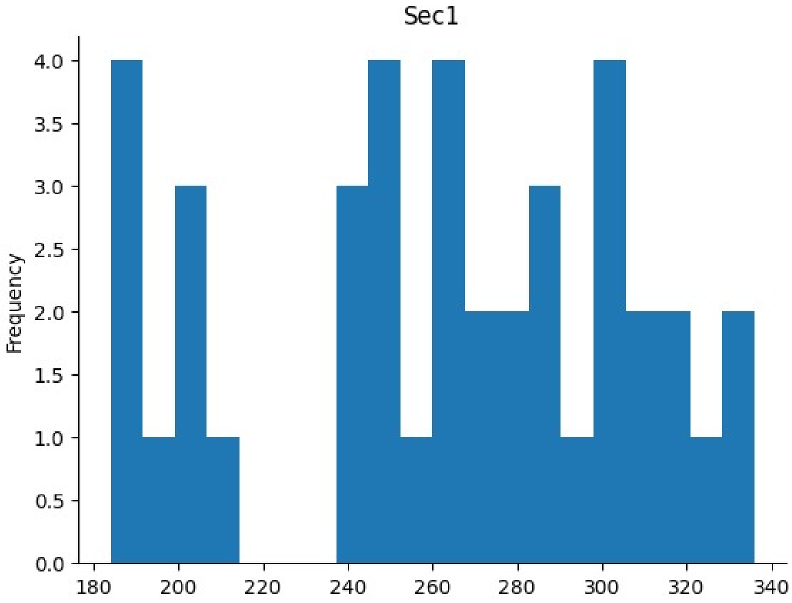


Figure 5: Frequency characteristics of the average daily duration of the Sec1 run by vehicles

Using matplotlib, a graph was generated with weeks on the x-axis and race time in seconds on the y-axis for each of the races from Sec1 to Sec10 (Fig. 6). Data was read from the CSV file using pandas, converting the decimal point data to floating point numbers.

Fig. 6 shows a graph of the average daily duration of each race by vehicles for each week during the study period, obtained using Python tools:

```
import pandas as pd
import matplotlib.pyplot as plt
import io
import seaborn as sns
import numpy as np
for col in df.columns[1:11]:  # Columns 'Sec1' to 'Sec10'
    df[col] = df[col].str.replace(',', '.').astype(float)
plt.figure(figsize=(12, 6))
for col in df.columns[1:11]:
    plt.plot(df['Week'], df[col], marker='o', label=col)
plt.xlabel('Week')
plt.ylabel('Race Time (seconds)')
plt.title('Race Time vs. Week for All Sections')
plt.legend(loc='upper right')
plt.grid(True)
plt.tight_layout()
plt.show()Do not insert line breaks in your title.
```

The graph (Fig. 6) shows the dependence of the race time in seconds on the week number for each of the sections (Sec1–Sec10). Each section is represented by a line of a different color with markers. The x-axis is the week number, and the y-axis is the race time in seconds. The plot has a grid for better readability, and a legend in the upper right corner identifies each section. Sec1 has the highest overall race time. Sec9 and Sec10 have the lowest and most stable race times. This Line Plot of All Sections shows us the trend of the race time for each section over all the weeks studied (Fig. 6).

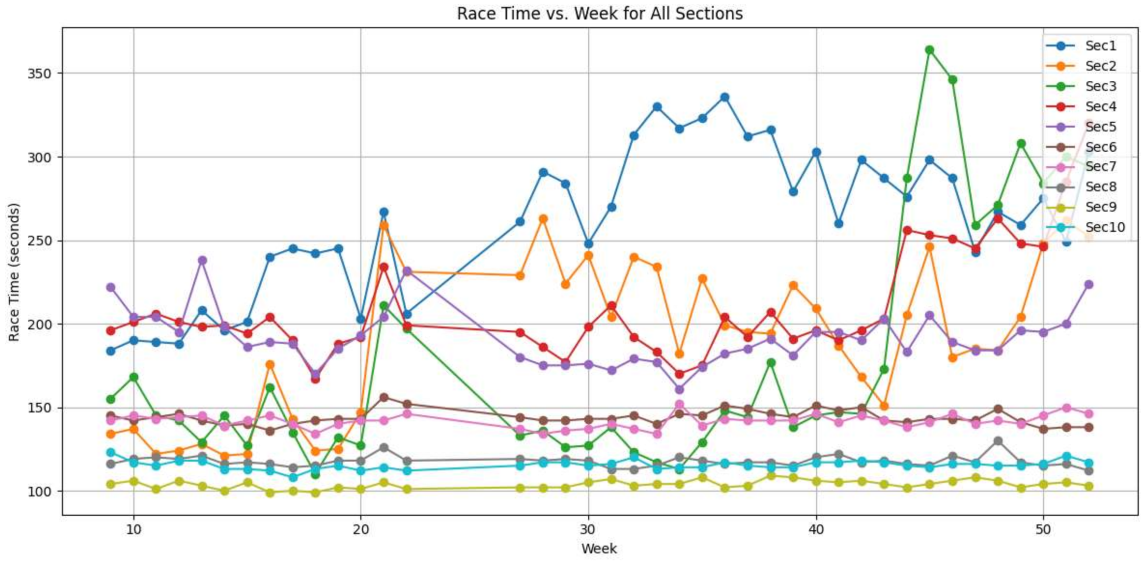


Figure 6: Line plot of average daily duration of each vehicle trip for each week during the study period for all sections

We see that the average daily duration of the races on average over the year is the highest for the Sec1 race (00:04:22), and the lowest for the Sec9 race (00:01:44), which indicates the dependence of passenger transportation in an average-sized city on traffic and the type of race, because the Sec1 race is a race in the city center with a high probability of congestion, and the Sec9 race is a race on an isolated line specifically for this public transport. The averaged average duration of the race for each week for the entire route indicates the presence of several hypotheses: hypothesis 1 about the existence of seasonal dependences of the amount of transport on the roads, as well as hypothesis 2 about the influence of weather changes on the duration of the races.

The research also used a Box Plot of All Sections, which shows us a statistical summary of the distribution of race times for each section of the route, highlighting the median, quartiles, and outliers (Fig. 7).

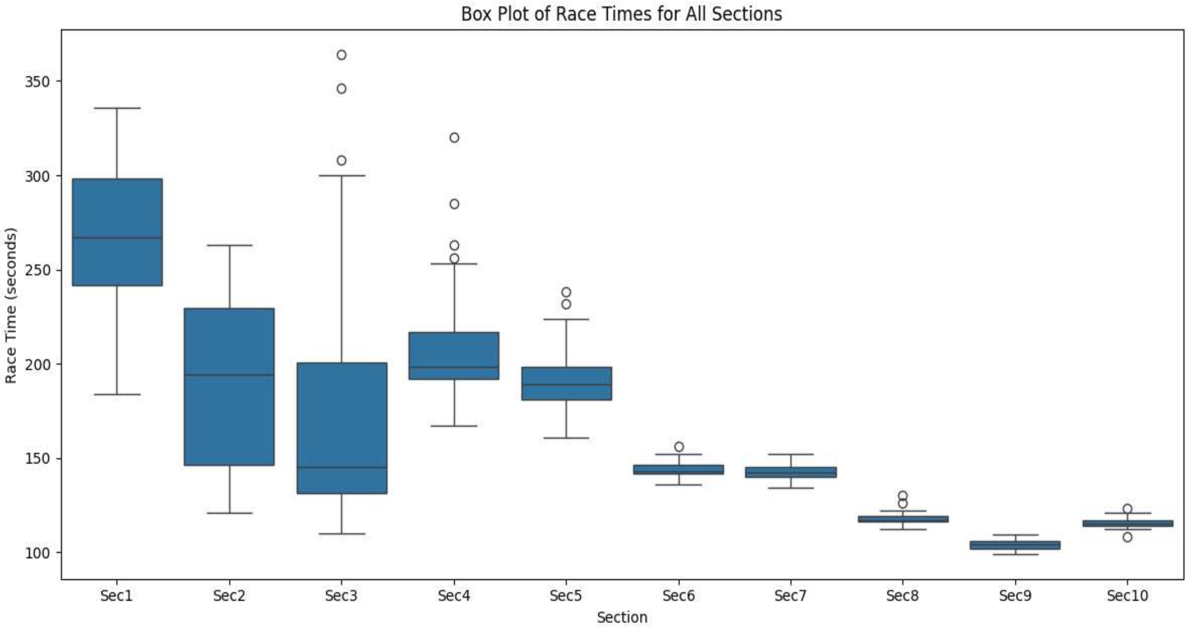


Figure 7: Box plot of average daily duration of each vehicle trip for each week during the study period for all sections

```
plt.figure(figsize=(12, 6))
df_melted = df.melt(id_vars=['Week'], value_vars=df.columns[1:11], var_name='Section', value_name='Race Time')
sns.boxplot(x='Section', y='Race Time', data=df_melted)
plt.xlabel('Section')
plt.ylabel('Race Time (seconds)')
plt.title('Box Plot of Race Times for All Sections')
plt.tight_layout()
plt.show()
```

Fig. 8 shows that the average race execution time generally increases with increasing week number, which indicates the existence of seasonality in this studied dataset.

```
df['Average Time'] = df.iloc[:, 1:11].mean(axis=1)
plt.figure(figsize=(10, 6))
plt.scatter(df['Week'], df['Average Time'])
plt.xlabel('Week')
plt.ylabel('Average Race Time (seconds)')
plt.title('Scatter Plot of Week vs. Average Race Time')
plt.grid(True)
plt.tight_layout()
plt.show()
```

The Histogram of Race Times provides a general idea of the frequency distribution of times for all races in all sections of the route (Fig. 9). Typically, race times are between 100 and 150 seconds, but there are races that exceed the time by 2-3 times.

```
plt.figure(figsize=(10, 6))
plt.hist(df.iloc[:, 1:11].values.flatten(), bins=20, color='skyblue')
plt.xlabel('Race Time (seconds)')
plt.ylabel('Frequency')
plt.title('Histogram of Race Times')
plt.tight_layout()
plt.show()
```

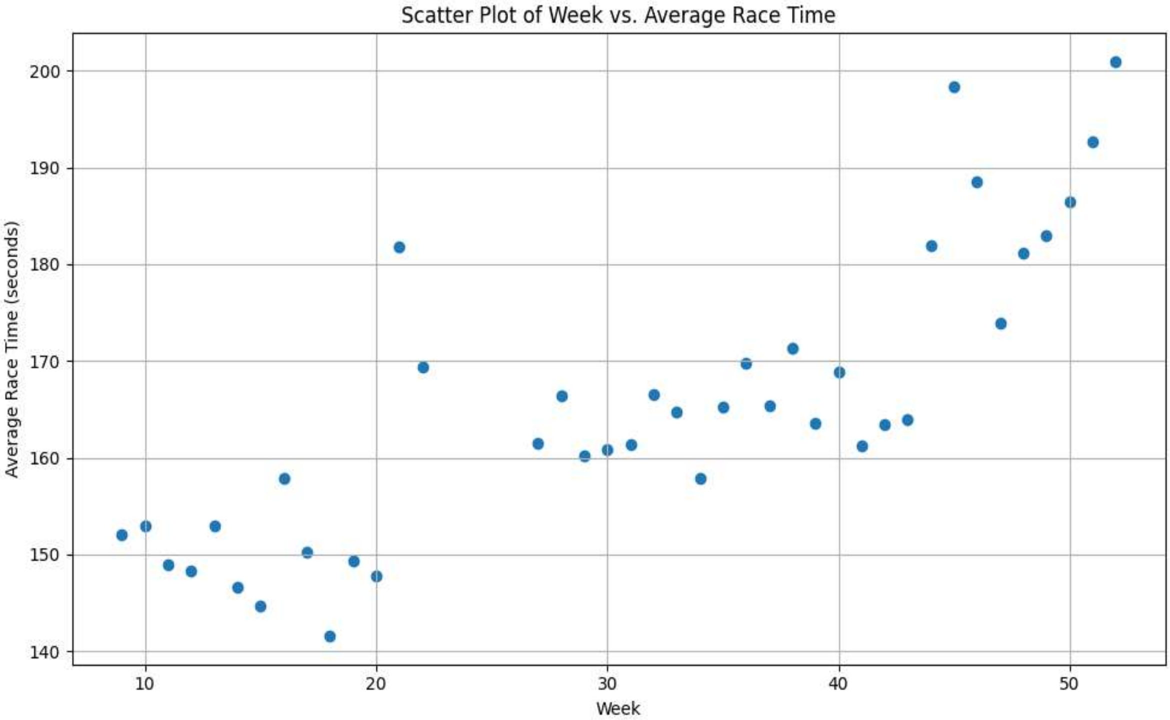


Figure 8: Scatter plot of week vs. average time average daily duration of each trip by vehicles for each week during the study period for all sections

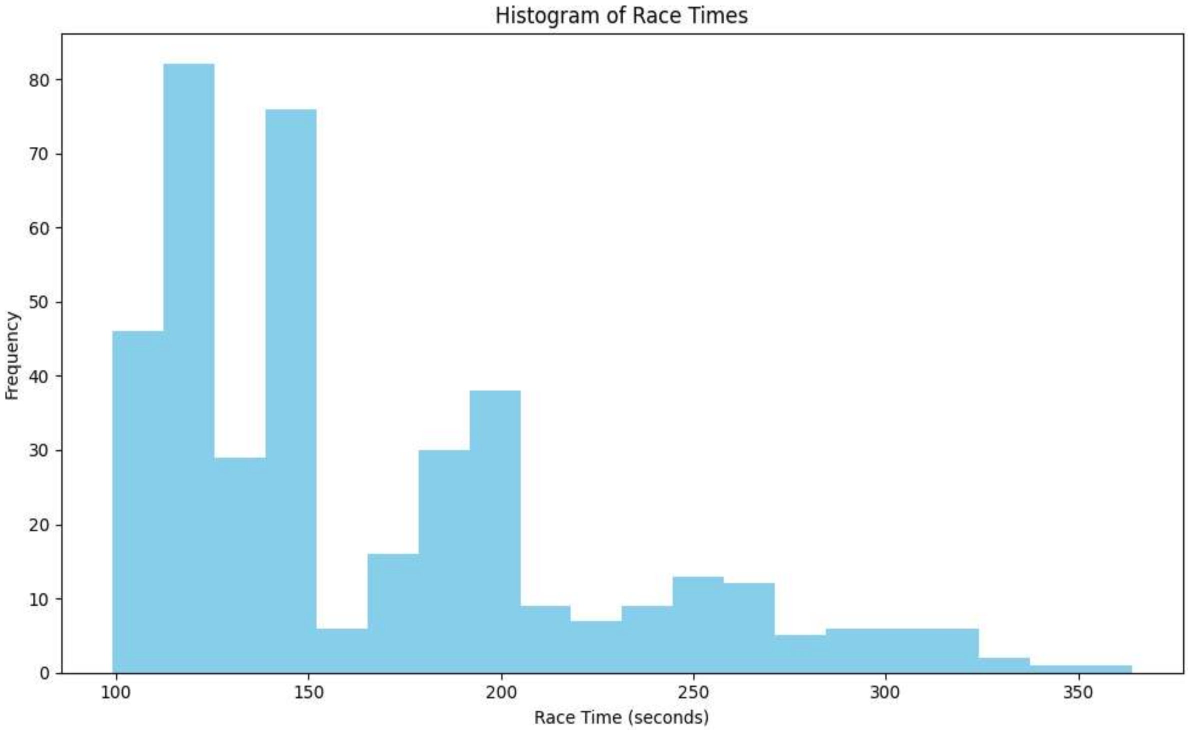
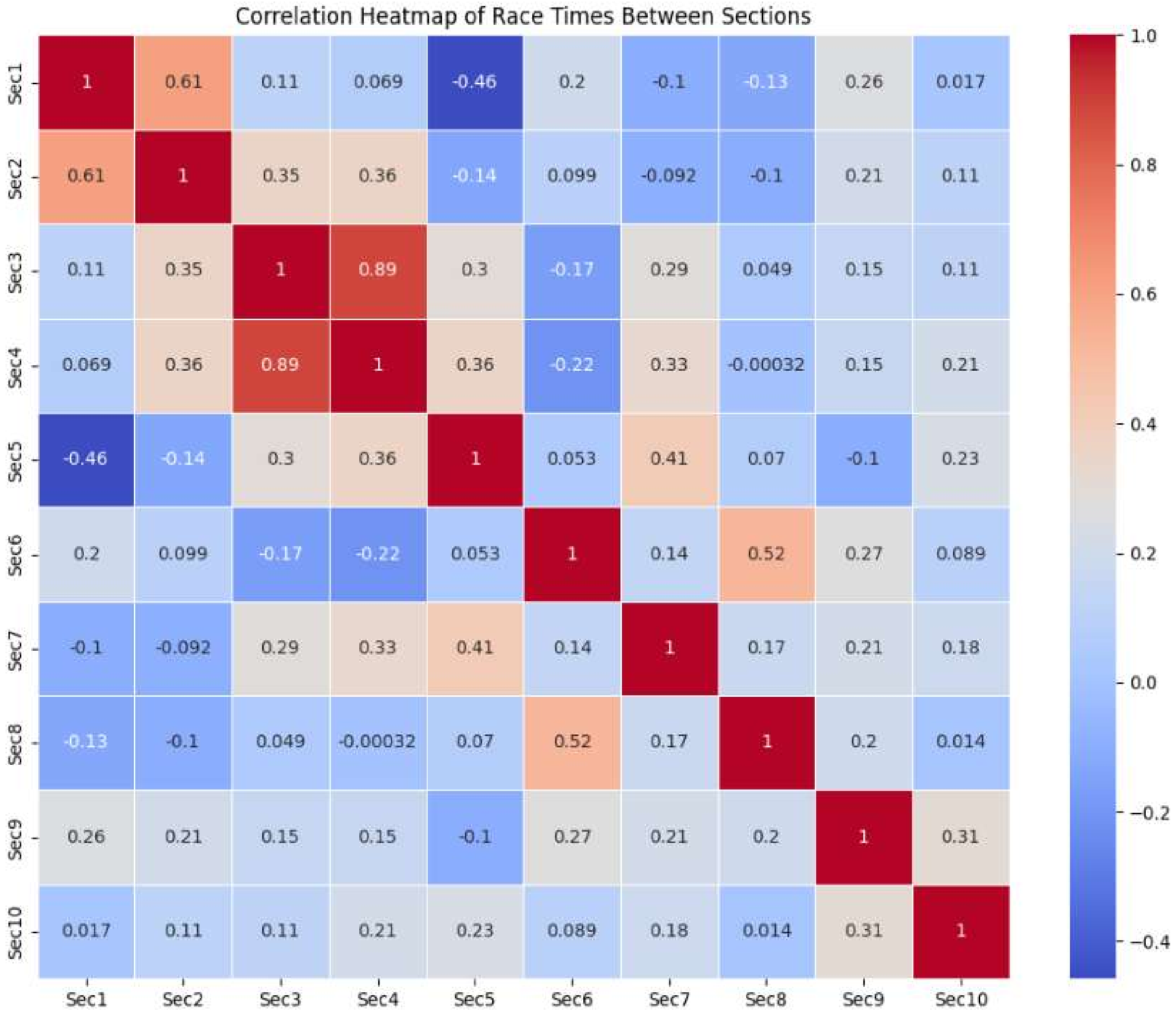


Figure 9: Histogram of race times for average daily duration of each race by vehicles for each week during the study period for all sections

The Correlation Heatmap shows the correlation between race times on different sections of the route for each section (Fig. 10).



	Sec1	Sec2	Sec3	Sec4	Sec5	Sec6	Sec7	Sec8	Sec9	Sec10
Sec1	1	0.61	0.11	0.069	-0.46	0.2	-0.1	-0.13	0.26	0.017
Sec2	0.61	1	0.35	0.36	-0.14	0.099	-0.092	-0.1	0.21	0.11
Sec3	0.11	0.35	1	0.89	0.3	-0.17	0.29	0.049	0.15	0.11
Sec4	0.069	0.36	0.89	1	0.36	-0.22	0.33	-0.00032	0.15	0.21
Sec5	-0.46	-0.14	0.3	0.36	1	0.053	0.41	0.07	-0.1	0.23
Sec6	0.2	0.099	-0.17	-0.22	0.053	1	0.14	0.52	0.27	0.089
Sec7	-0.1	-0.092	0.29	0.33	0.41	0.14	1	0.17	0.21	0.18
Sec8	-0.13	-0.1	0.049	-0.00032	0.07	0.52	0.17	1	0.2	0.014
Sec9	0.26	0.21	0.15	0.15	-0.1	0.27	0.21	0.2	1	0.31
Sec10	0.017	0.11	0.11	0.21	0.23	0.089	0.18	0.014	0.31	1

Figure 10: Correlation heatmap for average daily duration of each vehicle trip for each week during the study period for all sections

```
plt.figure(figsize=(10, 8))
corr_matrix = df.iloc[:, 1:11].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=.5)
plt.title('Correlation Heatmap of Race Times Between Sections')
plt.tight_layout()
plt.show()
```

From Fig. 10 it can be seen that the Sec3 and Sec4 sections have a strong correlation, which indicates an unresolved problem of a transport node with high traffic intensity between these sections, which causes delays on the route.

Summarizing this analysis of passenger transportation in an average-sized city, it was found that the average daily duration of each leg for each week increases with the beginning of the autumn-winter period and reaches its maximum in the 52nd week of the year (00:03:21), lower than average values of the average daily duration of each leg are observed in the spring period, with the minimum value (00:02:22) falling on the 18th week of the year (end of April - beginning of May).

4.2. A cluster analysis of the data set on the duration of electric transport races in an average-sized city

Cluster analysis of such a large-scale heterogeneous data set on the duration of electric transport trips within a medium-sized city with a developed public transport network was carried out using the K-means clustering method due to its feature of necessarily exact distribution of data between clusters.

It should be noted that there are several options for selecting the optimal value of the number of clusters k, among which the elbow method, the silhouette method and the Calinski-Harabasz index are most often used. The elbow method considers subjectively understandable graphs of the nature of the change in the scatter of points (Wtotal → max) from the largest value for all points in one cluster to the smallest value (Wtotal → 0) with an increase in the number of groups k (k→ n).

The silhouette method measures how similar the points in one cluster are compared to other clusters. The value of the silhouette index is in the range [−1,1], where larger values indicate better clustering quality. This method assesses how well the points are located inside their clusters compared to other clusters. A larger value of the silhouette coefficient indicates better clustering quality. The Calinski-Harabasz index, also known as the dispersion ratio criterion, involves determining the ratio of the intercluster separation to the intracluster dispersion, normalized by their number of degrees of freedom. The highest value of the Calinski-Harabasz index indicates that the clusters are defined most clearly. Although this metric is best suited for calculating the value of the number of clusters, it has the same drawback as the silhouette coefficient - it overestimates the estimate for convex cluster shapes and underestimates the estimate for complex cluster shapes. In order to find the optimal number of clusters k for the data set with the average daily durations of each of the races during the week on the route in an average-sized city (Fig. 12), the elbow, silhouette and Calinski-Harabasz methods were used. The results of estimating the coefficient of total variation of points within the cluster relative to the cluster center SSE by the elbow method are shown in Fig. 11. The optimal value of the number of clusters is k=5 with the value of SSE=70896.042 (Fig. 11). The results of the estimation of the silhouette coefficient Si by the silhouette method are shown in Fig. 12. In our case, the maximum value of the silhouette coefficient Si =0.507 occurs at k=2, which is considered the optimal value of the number of clusters for clustering (Fig. 12). Fig. 13 shows the results of the estimation of the Calinski-Harabasz index and the corresponding values of the number of clusters. In our case, the maximum value of the Calinski-Harabasz index S =56.186 occurs at k=3 (Fig. 13), which indicates the optimal value of the number of clusters for data clustering.

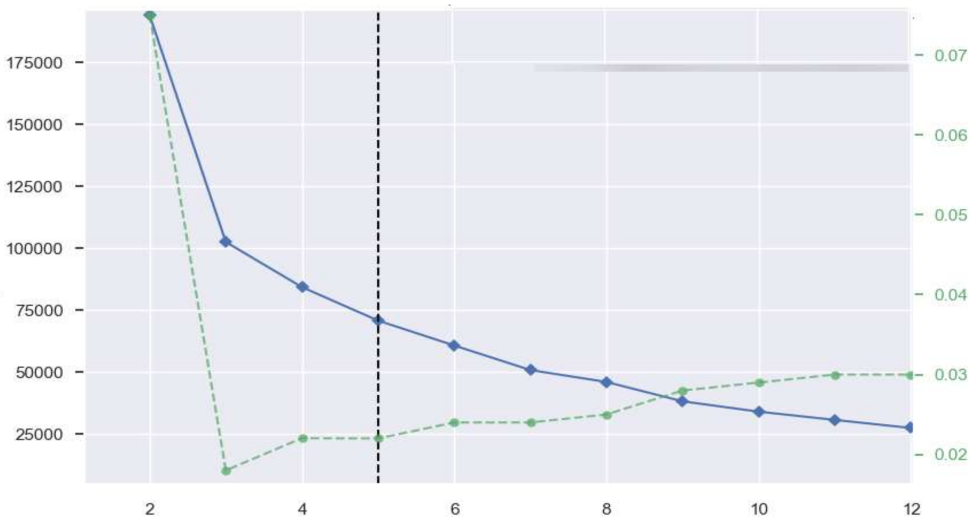


Figure 11: The dependence of the SSE (Sum of Squared Errors) value on the number of clusters k, calculated using the elbow method, where the SSE(k) estimation graph is marked in blue - the left axis, and the expected learning time is marked in green - the right axis.

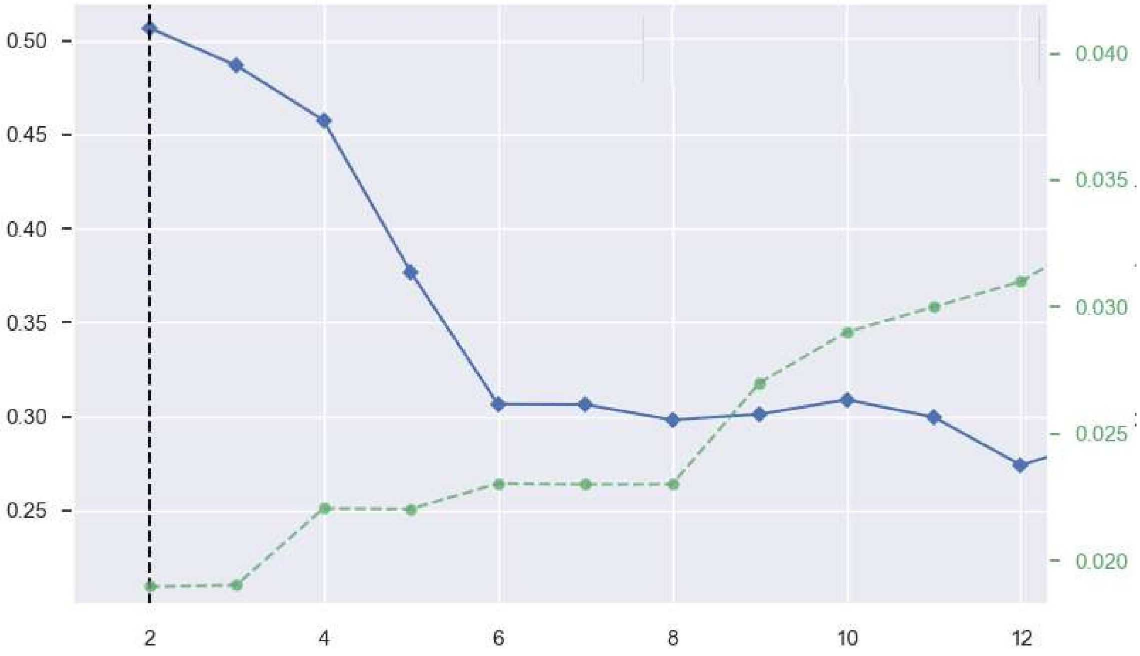


Figure 12: Dependence of the silhouette coefficient on the number of clusters k, where the blue graph indicates the Si(k) estimate – left axis, and the green graph indicates the expected learning time – right axis.

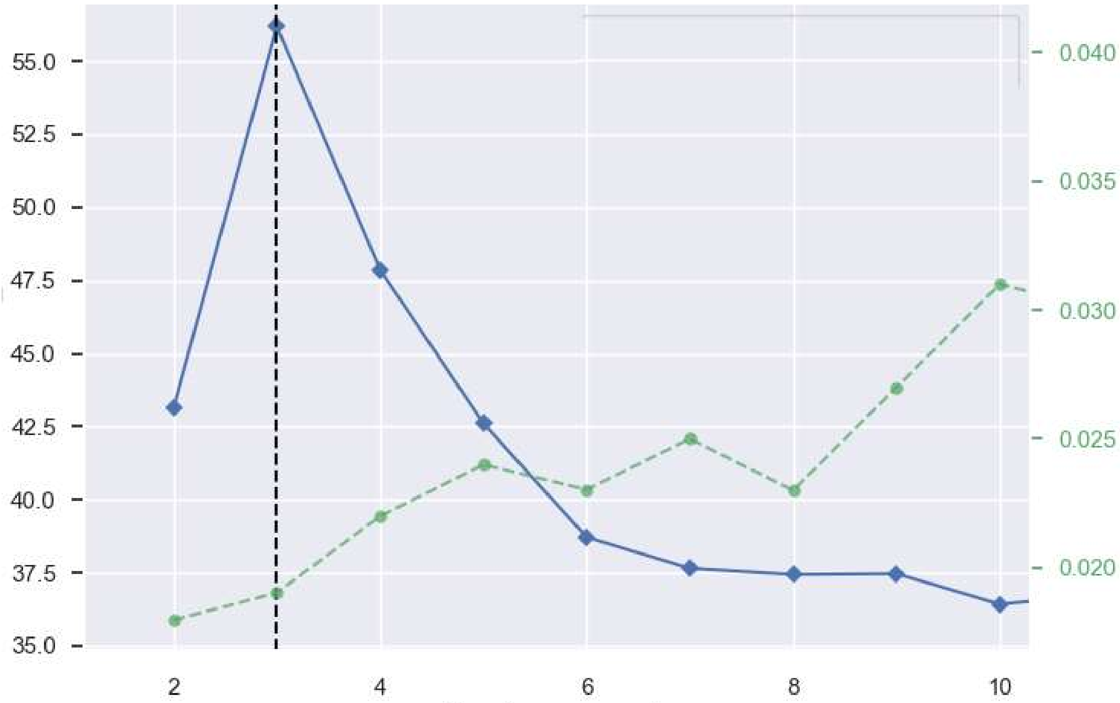
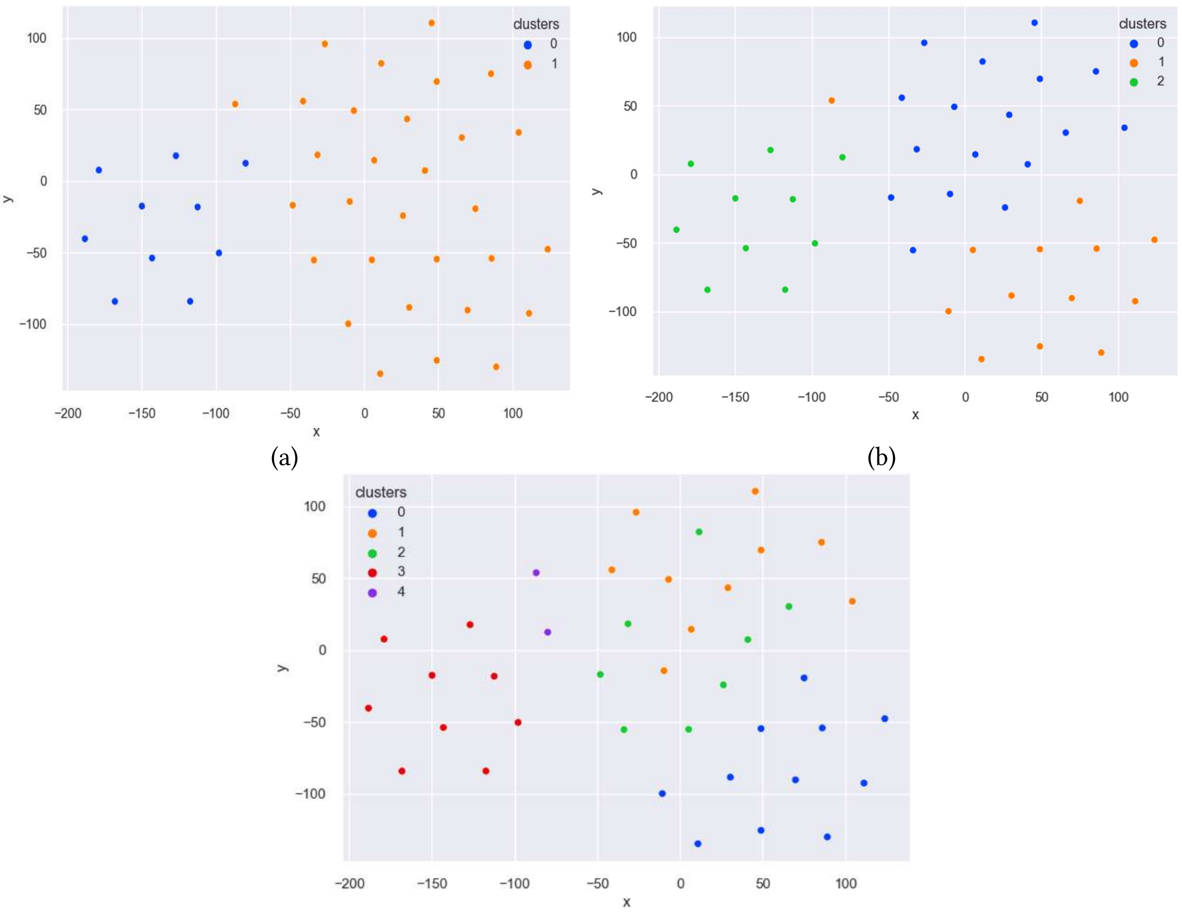


Figure 13: Dependence of the Calinski-Harabasz index on the number of clusters k, where the blue graph of the S(k) estimate is marked – the left axis, and the green graph is the expected learning time – the right axis.

When clustering the average daily duration of the races for each week using the K-means method, the results of calculating the number of clusters k using the elbow, silhouette and Calinski-Harabasz methods were taken into account, respectively k=5, k=2 and k=3 (Fig. 11 – Fig. 13). Fig. 14 shows the distribution of data (average daily values of the duration of each race for each week during the year) into clusters, obtained for k=2 (a); k=3 (b) and k=5 (c).



(a)

(b)

(c)

Figure 14: Results of clustering of average daily values of passenger transportation schedules for each week during the year on each section (leg), namely: clustering of the data set for k=2 (a); clustering of the data set for k=3 (b); clustering of the data set for k=5 (c).

5. Discussion

Let's conduct a detailed analysis of the distribution of data into clusters. When divided into two clusters, where the value of k=2 was obtained by the silhouette method, we have clusters with numbers 0 and 1 (according to Fig. 14 (a). The first cluster under number 0 forms the data of execution of passenger transportation schedules on each leg for weeks 9-20 and 22-43 with average daily values close to the average or less than it (Fig 1 - Fig 2, Table 1). The second cluster under number 1 forms the data of execution of passenger transportation schedules on each leg for weeks 21 and 44-52 with average daily values significantly higher than the average Fig 1 - Fig 2, Table 1) for at least two legs. This cluster is also characterized by the presence of weeks (21, 45, 46, 50-52) with a significant excess (by 1.5-2 times) of the average daily values of execution of passenger transportation schedules on three or more legs. Most of such significant exceedances occur in the autumn-winter period of the year, which is due to difficult weather conditions.

Table 1
Division into clusters using the K-means method of the average daily duration of each journey by vehicles for each week

week	9	10	11	12	13	14	15	16	17	18	19	20	21	22	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
k=2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
k=3	1	1	1	1	1	1	1	1	1	1	1	1	0	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	0	0	0	0	0	0	0	0
k=5	0	0	0	0	0	0	0	2	0	0	0	0	4	4	1	1	1	1	1	1	1	1	1	2	2	2	1	2	2	2	2	3	3	3	3	3	3	3	3	3

When divided into three clusters, where the value of k=3 was obtained by the Calinski-Harabasz method, we have clusters with numbers 0; 1 and 2, which are displayed in Fig. 14 (b). The first cluster under number 0 forms the data of passenger transportation schedules for weeks 21 and 44 - 52 with average daily values significantly higher than the annual average on each leg (Fig 1, Fig 2, Table 1)) mainly for three or more legs. This cluster is also characterized by the presence of weeks (21, 44-46, 48, 50-52) with a significant excess of average daily values of passenger transportation schedules on three or more legs, which is due to difficult weather conditions in the autumn-winter period. This indicates the dependence of the duration of the legs on seasonality. The second cluster under number 1 forms the data of execution of passenger transportation schedules only for weeks 9-20 and 22 with average daily values less than the annual average on each leg (Fig 1, Fig 2, Table 1). The third cluster under number 2 forms the data of execution of passenger transportation schedules for weeks 27 - 43 with average daily values close to the annual average on each leg, and insignificant excesses of the annual average are observed on no more than two legs during the week (Fig 1, Fig 2, Table 1).

When divided into five clusters (k=5), the value of which was obtained by the elbow method (Fig. 11), we have clusters with numbers 0; 1; 2; 3 and 4, shown in Fig. 14 (c). The first cluster under number 0 forms the data on the execution of passenger transportation schedules on each leg for weeks 9-15 and 17-20 with average daily values less than the annual average (Fig 1, Fig 2, Table 1). The second cluster under number 1 forms the data on the execution of passenger transportation schedules on each leg for weeks 27-35 and 39 with average daily values higher than the annual average for no more than two legs (Fig 1, Fig 2, Table 1). The third cluster under number 2 forms the data on the execution of passenger transportation schedules for weeks 16, 36 - 38 and 40-43 with average daily values close to the annual average for almost every leg (Fig 1, Fig 2, Table 1), with the excess of the annual average being observed for no more than one leg. Exceedances are observed only for the city center run, which indicates the dependence of the run duration on their location in specific urban areas. The fourth cluster under number 3 is formed by the data on the execution of passenger transportation schedules for weeks 44–52 with average daily values significantly higher than the annual average mainly on three or more runs (Fig 1, Fig 2, Table 1), which is due to the presence of seasonality in the studied dependence of the run duration. This cluster is also characterized by the presence of weeks 45 and 52 with a significant excess of the average daily values of the execution of passenger transportation schedules on five runs, which may indicate both a high impact of traffic together with seasonality. The fifth cluster under number 4 forms the data of passenger transportation schedules execution only for weeks 21 and 22 in the summer period with average daily values significantly higher than the annual average on three or more routes (Fig 1, Fig 2, Table 1). This cluster indicates only the high impact of traffic on the average daily values of passenger transportation schedules execution on routes in the city center. It should be noted that no excesses of the average annual values of schedule execution were observed for routes 6-10, which are on an isolated line allocated only for this type of electric transport. This indicates the optimal way to solve the problems with passenger transportation by public transport, but it is complex, because it requires significant investments in the city's infrastructure and is long in implementation.

Thus, the cluster analysis of a large-scale heterogeneous data set on the duration of electric transport trips within an average-sized city revealed individual sections (trips, zones) of the route with different intensities, which are highly influenced by traffic, their location in specific urban areas (city center, residential area, etc.), as well as seasonality. Despite the subjectivity of determining the optimal value of the number of clusters using the elbow method, we see that dividing the average daily duration of trips for each week into clusters gave the best results for k=5, where the value of the estimate of the intra-cluster total variation of points within the cluster relative to the cluster center SSE=70896.042 (Fig. 11). It should also be noted that at k=5 the values of the silhouette coefficient Si =0.378 and the Calinski-Harabasz index S =42.5086 are not significantly less than the maximum values of the silhouette coefficient (Fig. 12) and the Calinski-Harabasz index (Fig. 13), respectively.

Thus, it can be stated that the proposed algorithm for selecting a clustering method based on internal metrics for assessing the quality of clustering data collected from the infrastructure of passenger transportation by public transport in a medium-sized city is quite simple and effective. The clustering metrics included the elbow method, the silhouette method and the Calinski-Harabasz index, which allow for a quick and easy selection of the optimal value of the number of clusters, as well as taking into account the features of the data. The elbow method allows us to take into account the intra-cluster general variation of points within a cluster relative to the cluster center, the silhouette method measures how similar the points in one cluster are compared to other clusters, and the highest value of the Calinski-Harabasz index indicates that the clusters are defined most clearly.

Thus, the K-means clustering method revealed the races with a high excess of the average daily values of the duration of the races compared to the average annual ones also indicate an increase in the waiting time of passengers at stops, which affects the number of passengers transported and the quality of the services provided. This indicates the need to make informed decisions in the field of passenger transportation by public transport in the city in order to optimize it.

We recommend using this K-means clustering method when analyzing the average daily duration of each trip by vehicle for each week during the studied period to make informed decisions in the field of passenger transportation by public transport in a smart city, namely for optimizing routes, adapting the transport network itself, forecasting and planning demand for transport services, implementing personalized services, as well as integrating different types of transport to create a single effective multimodal transport system.

Thus, when analyzing passenger flows using the K-means clustering method, the identified areas of high demand will allow creating optimal transport routes that meet the real needs of passengers at a specific point in time, reducing waiting times and the number of transfers to the minimum possible. This K-means clustering method is also useful when analyzing changes in passenger needs and will facilitate the adaptation of public transport routes to changes in demand, for example, adding new stops, changing vehicle schedules and their schedules. This will also allow city government leaders to better plan infrastructure projects and investments in the modernization of the transport system in order to integrate different modes of transport (electric transport, regular buses, metro, if available) to create a single efficient transport system. In a smart city, personalization of services is also important, where mobile applications for public transport play an important role, which, when providing personalized recommendations to passengers on choosing the optimal route or travel time, will use the results of clustering the duration of the race schedules in real time. The main problems that should be solved using big data clustering are the allocation of passenger clusters by type (workers, students, tourists, etc.), identification of hot spots (areas with the highest demand for transport at a certain time), identification of inefficient routes or low load on individual sections of the transport network, analysis of the dependence of passenger flows on external factors (weather, events in the city, social trends), as well as building dynamic models for predicting changes in flows.

Therefore, the obtained results of cluster analysis of the average daily duration of each journey by vehicles for each week during the studied period have practical value in optimizing routes, adapting the transport network itself, forecasting and planning demand for transport services, implementing personalized services, as well as integrating different types of transport to create a single effective multimodal transport system. It was recommended to use clustering to optimize routes, namely to create optimal transport routes that have reduced waiting times and fewer transfers, and also meet the real needs of passengers at the time they specify.

6. Conclusions and prospects for further development

In order to study the possibilities of applying clustering methods in organizing passenger transportation in a smart city, a study was conducted to study the features of their application to improve the organization of passenger transportation by public transport. This made it possible to establish that the choice of a clustering method depends on the specifics of the task, data characteristics and goals set when analyzing transport and passenger flows. Thus, bottom-up clustering methods are used when it is necessary to determine the structure of routes, to combine similar sections of routes or territories adjacent to the route. Top-down methods are used when it is necessary to identify individual routes or their individual sections (races, zones) with different passenger flow intensity for further optimization of services in specific zones. The K-means distributive clustering method is useful for zoning the transport network, for example, when optimizing routes by the duration of the races, by demand zones, etc. The K-medoid distribution clustering method is more robust to noise, so it is used to optimize the operating schedule of vehicles, determine the optimal location of stops on routes to best meet passenger needs. The C-means fuzzy clustering method is used to develop more flexible and adaptive passenger flow management strategies, which is important in modern urban environments with the dynamic nature of demand for transport services. The DBSCAN method, which classifies elements based on density, is most often used to analyze passenger flows with different densities on different routes and for clustering taking into account time dynamics. The self-organizing map method is used for clustering to create dynamic maps of passenger flows taking into account time intervals and densities on different routes.

As a result of the cluster analysis of passenger transportation in an average-sized city with a developed public transport network, it was found that the collected data on the duration of each journey by vehicles within each day for each week during the studied period have a complex, heterogeneous and large-scale structure, therefore they require appropriate processing before starting the analysis. The cluster analysis of such a large-scale heterogeneous data set on the duration of electric transport journeys within an average-sized city was carried out using the K-means clustering method, since this method, by reducing the value of the loss function, necessarily implements an accurate data distribution, where each object belongs to only one class. A simple and most effective algorithm for choosing a clustering method is proposed based on internal metrics for assessing the quality of clustering of data collected from the infrastructure of passenger transportation by public transport in an average-sized city. The clustering metrics included the elbow method, the silhouette method and the Calinski-Harabasz index, which allow for a quick and simple selection of the optimal value of the number of clusters. The elbow method allows us to establish the intra-cluster general variation of points within a cluster relative to the cluster center, the silhouette method measures how similar the points in one cluster are compared to other clusters, and the highest value of the Calinski-Harabasz index indicates that the clusters are defined most clearly.

The obtained results of the cluster analysis of the average daily duration of each trip by vehicles for each week during the studied period have practical value in optimizing routes, adapting the transport network itself, forecasting and planning the demand for transport services, implementing personalized services, as well as integrating different modes of transport to create a single effective multimodal transport system. It was recommended to use clustering for route optimization, namely for creating optimal transport routes that have reduced waiting times and fewer transfers, and also meet the real needs of passengers at the time they specify.

Therefore, the K-means clustering method when analyzing the average daily duration of each trip by vehicles for each week during the studied period is appropriate to use for optimizing the organization of passenger transportation by public transport in a smart city. The prospect of further research is the use of big data clustering to identify clusters of passengers by type (workers, students, tourists, etc.), identify hot spots (areas with the highest demand for transport at a certain time), identify inefficient routes or low load on individual sections of the transport network, analyze the dependence of passenger flows on external factors (weather, events in the city, social trends), as well as build dynamic models for predicting changes in flows.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] Saxena A., Prasad M., Gupta A., Bharill N., Patel O. P., Tiwari A., Er M. J., Ding W., Lin C. A review of clustering techniques and developments. Neurocomputing. 2017. No. 267. P. 664–681. DOI: 10.1016/j.neucom.2017.06.053

[2] Isoli N., Chaczykowski M. Net energy analysis and net carbon benefits of CO2 capture and transport infrastructure for energy applications and industrial clusters. Applied Energy. 2025. No. 382, 125227. DOI: 10.1016/j.apenergy.2024.125227

[3] A. Kowalska-Styczeń, M.Bublyk, V. Lytvyn, Green innovative economy remodeling based on economic complexity, Journal of Open Innovation: Technology, Market, and Complexity 9(3) (2023) 100091. doi: 10.1016/j.joitmc.2023.100091

[4] L. Podlesna, M. Bublyk, I. Grybyk, Y. Matseliukh, Y. Burov, P. Kravets, O. Lozynska, I. Karpov, I. Peleshchak, R. Peleshchak, Optimization model of the buses number on the route based on queuing theory in a Smart City, CEUR Workshop Proceedings Vol-2631 (2020) 502-515. URL: https://ceur-ws.org/Vol-2631/paper37.pdf.

[5] Bianchini D., De Antonellis V., Garda M. A big data exploration approach to exploit in-vehicle data for smart road maintenance. *Future Generation Computer Systems.* 2023. No. 149. P. 701–716. DOI: 10.1016/j.future.2023.08.004

[6] A. Katrenko, I. Krislata, O. Veres, O. Oborska, T. Basyuk, A. Vasyliuk, I. Rishnyak, N. Demyanovskyi, O. Meh Development of traffic flows and smart parking system for smart city. CEUR Workshop Proceedings Vol-2604 (2020) 730–745. URL: https://ceur-ws.org/Vol-2604/paper50.pdf

[7] Y. Matseliukh, M. Bublyk, A. Bosak, M. Naychuk-Khrushch, The role of public transport network optimization in reducing carbon emissions, CEUR Workshop Proceedings Vol-3723 (2024) 340–364. URL: https://ceur-ws.org/Vol-3723/paper19.pdf

[8] Visan M., Negrea S. L., Mone F. Towards intelligent public transport systems in Smart Cities; Collaborative decisions to be made. *Procedia Computer Science.* 2021. No. 199. P. 1221–1228. DOI: 10.1016/j.procs.2022.01.155

[9] Ezugwu A. E., Ikotun A. M., Oyelade O. O., Abualigah L., Agushaka J. O., Eke C. I., Akinyelu A. A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence.* 2022. No. 110, 104743. DOI: 10.1016/j.engappai.2022.104743

[10] Chavent M., Lechevallier Y., Briant O. DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis.* 2007. No. 52(2). P. 687–701. DOI: 10.1016/j.csda.2007.03.013

[11] Celebi M. E., Kingravi H. A., Vela P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems With Applications.* 2012. No. 40(1). P. 200–210. DOI: 10.1016/j.eswa.2012.07.021

[12] A. Bakurova, V. Bilyi, A. Didenko,E. Tereschenko, Analytics module for the system for recording destruction due to russian aggression, in Monitoring of Geological Processes and Ecological Condition of the Environment 2023 (2023) 1–5. doi: 10.3997/2214-4609.2023520232

[13] Singh J., Singh D. A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects. *Advanced Engineering Informatics.* 2024. No. 62, 102799. DOI: 10.1016/j.aei.2024.102799

[14] Yan J., Liu J., Tseng F. An evaluation system based on the self-organizing system framework of smart cities: A case study of smart transportation systems in China. *Technological Forecasting and Social Change.* 2020. No. 153, 119371. DOI: 10.1016/j.techfore.2018.07.009

[15] M. Gvozd, , Ohinok, S., Ivaniuk, U., Protsak, K., , L. Chernobay, Independent factors simulation of the influence on the level of sustainable development in intellectual systems of management, CEUR Workshop Proceedings Vol-3426 (2023) 246–258. URL: https://ceur-ws.org/Vol-2870/paper118.pdf

[16] Prasetio E. A., Novizayanti D., Putri A. N. A. Cluster analysis of potential autonomous vehicle (AV) adopters in Indonesia's new capital. *Transportation Research Interdisciplinary Perspectives.* 2024. No. 29, 101318. DOI: 10.1016/j.trip.2024.101318

[17] Singh J., Singh D. A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects. *Advanced Engineering Informatics.* 2024. No. 62, 102799. DOI: 10.1016/j.aei.2024.102799

[18] Liu J., Li J., Chen Y., Lian S., Zeng J., Geng M., Zheng S., Dong Y., He Y., Huang P., Zhao Z., Yan X., Hu Q., Wang L., Yang D., Zhu Z., Sun Y., Shang W., Wang D., Chen X. Multi-scale urban passenger transportation CO_2 emission calculation platform for smart mobility management. *Applied Energy.* 2023. No. 331, 120407. DOI: 10.1016/j.apenergy.2022.120407

[19] O. Ptashchenko, L. Chernobay, S. Malykhina, I. Verezomska, S. Yaremchuk, Problems and prospects of application of strategies of personnel management of international companies in Ukrainian business practice, Financial and Credit Activity: Problems of Theory and Practice 1 (42) (2022) 406–411. doi: 10.55643/fcaptp.1.42.2022.3661

[20] Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. Bon, Springer, 1981, 245 p.

[21] M. Bublyk, Y. Matseliukh, Small-batteries utilization analysis based on mathematical statistical methods in challenges of circular economy, CEUR workshop proceedings Vol-2870 (2021) 1594-1603. URL: https://ceur-ws.org/Vol-2870/paper118.pdf

[22] M. Bublyk, V. Vysotska, Y. Matseliukh, V. Mayik, M. Nashkerska, Assessing losses of human capital due to man-made pollution caused by emergencies, CEUR Workshop Proceedings Vol-2805 (2020) 74-86. URL: https://ceur-ws.org/Vol-2805/paper6.pdf.

[23] EsterM., Kriegel H.-P., Sander J., Xu X. *A density-based algorithm for discovering clusters in large spatial databases with noise.* Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996. P. 226–231.

[24] Jain A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters.* 2010. No. 31(8). P. 651-666. DOI: 10.1016/j.patrec.2009.09.011

[25] Kohonen T. *Self-Organizing Maps* (3rd ed.). Bon, Springer, 2001. 554 p.

[26] M. Boikiv, T. Postranskyy, M. Afonin. Establishing patterns of change in the efficiency of regulated intersection operation considering the permitted movement directions. Eastern-European Journal of Enterprise Technologies 4(3(118) (2022) 17–26. doi: 10.15587/1729-4061.2022.262250

[27] T. Postranskyy, M. Afonin, M. Boikiv, R. Bura, Identifying patterns of change in traffic flows' parameters depending on the organization of public transport movement. Eastern-European Journal of Enterprise Technologies 5(3(131) (2024) 72–81. doi: 10.15587/1729-4061.2024.313636

[28] Law A. M. *Simulation Modeling and Analysis* (5th ed.). Bon, McGraw-Hill, 2015. 495 p.

[29] Nath N., Nitanai R., Manabe R., Murayama A. A global-scale review of smart city practice and research focusing on residential neighbourhoods. *Habitat International.* 2023. Vol. 142, P. 102963. DOI: 10.1016/j.habitatint.2023.102963

[30] Sun L., Zhao J., Zhang J., Zhang F., Ye, K., Xu C. Activity-based individual travel regularity exploring with entropy-space K-means clustering using smart card data. *Physica A: Statistical Mechanics and its Applications.* 2024. No. 636, 129522. DOI: 10.1016/j.physa.2024.129522

[31] Tibshirani R., Walther G., Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2001. No. 63(2). P. 411–423.

[32] Subudhi S., Panigrahi S. Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University - Computer and Information Sciences.* 2020. No. *32*(5). P. 568–575. DOI: 10.1016/j.jksuci.2017.09.010

[33] Duan Y., Liu C., Li S., Guo X., Yang C. An automatic affinity propagation clustering based on improved equilibrium optimizer and t-SNE for high-dimensional data. *Information Sciences*, 2023.No. 623. P. 434–454. DOI: 10.1016/j.ins.2022.12.057

[34] Bide P., Shedge R. Improved document clustering using K-means algorithm. *2015 IEEE Int. Conf. on Electrical, Computer and Communication Technologies (ICECCT).* New York: IEEE, 2015. 1048 p.

[35] Zhong C., Miao D., Wang R., Zhou X. DIVFRP: An automatic divisive hierarchical clustering method based on the furthest reference points. *Pattern Recognition Letters.* 2008. No. 29(16). P. 2067–2077. DOI: 10.1016/j.patrec.2008.07.002

[36] Elassy M., Al-Hattab M., Takruri M., Badawi S. Intelligent transportation systems for sustainable smart cities. *Transportation Engineering.* 2024. No. 16, 100252. DOI: 10.1016/j.treng.2024.100252

[37] Zhu W. A spatial decision-making model of smart transportation and urban planning based on coupling principle and Internet of Things. *Computers and Electrical Engineering.* 2022. No. 102, 108222. DOI: 10.1016/j.compeleceng.2022.108222

[38] Bushuev S., Inna L., Alla B., Alexander L., Khusainova M. Creating Urban Transportation Networks Grounded In the Principles of the Smart Port-City Paradigm. *Procedia Computer Science.* 2023. No. 231. P. 323–328. DOI: 10.1016/j.procs.2023.12.211

[39] Balbin P. P., Barker J. C., Leung C. K., Tran M., Wall R. P., Cuzzocrea A. Predictive analytics on open big data for supporting smart transportation services. *Procedia Computer Science.* 2019. No. 176. P. 3009–3018. DOI: 10.1016/j.procs.2020.09.202

[40] Khemakhem S., Krichen L. A comprehensive survey on an IoT-based smart public street lighting system application for smart cities. *Franklin Open.* 2024. No. 8, 100142. DOI: 10.1016/j.fraope.2024.100142

[41] Vidović K., Čolić P., Vojvodić S., Blavicki A. Methodology for public transport mode detection using telecom big data sets: Case study in Croatia. *Transportation Research Procedia.* 2021. No. 64. P. 76–83. DOI: 10.1016/j.trpro.2022.09.010

[42] N. Chukhray, N. Shakhovska, O. Mrykhina, M. Bublyk, L. Lisovska, Consumer aspects in assessing the suitability of technologies for the transfer, in: Computer sciences and information technologies, (CSIT), 2019, pp. 142–147, 8929879. doi: 10.1109/STC-CSIT.2019.8929879

[43] O. Pyroh, M. Prokopenko, L. Chernobay, R. Kovalenko, Y. Papizh, Y. Syta, Management of business processes and export-import activity of industrial enterprises in the digital economy, Estudios de Economia Aplicada, 39(5) (2021) 1-11. doi: 10.25115/eea.v39i5.5204.

[44] Y. Fornalchyk, I. Vikovych, Y. Royko, O. Hrytsun, Improvement of methods for assessing the effectiveness of dedicated lanes for public transport, Eastern-European Journal of Enterprise Technologies 1(3-109) (2021) 29–37. doi: 10.15587/1729-4061.2021.225397

[45] Y. Fornalchyk, I. Kernytskyy, O. Hrytsun, Y. Royko, Choice of the rational regimes of traffic light control for traffic and pedestrian flows, Scientific Review Engineering and Environmental Sciences 30(1) (2021) 38–50. doi: 10.22630/PNIKS.2021.30.1.4

[46] Y. Fornalchyk, E. Koda, I. Kernytskyy, O. Hrytsun, Y. Royko, R. Bura, P. Osiński, R. Barabash, R. Humenuyk, P. Polyansky, The impact of vehicle traffic volume on pedestrian behavior at unsignalized crosswalks. Roads and Bridges – Drogi i Mosty 22 (2023) 201-219. doi: 10.7409/rabdim.023.010

[47] A. Bakurova, H. Ropalo, E. Tereschenko, Analysis of the effectiveness of the successive concessions method to solve the problem of diversification, CEUR Workshop Proceedings Vol. 2917 (2021) 231–242. URL: https://ceur-ws.org/Vol-2917/paper21.pdf

[48] Dai Y., Hasanefendic S., Bossink B. A systematic literature review of the smart city transformation process: The role and interaction of stakeholders and technology. *Sustainable Cities and Society.* 2024. Vol. 101, P. 105112. DOI: 10.1016/j.scs.2023.105112

[49] Cai J., Luo J., Wang S., Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing.* 2018. No.300. P. 70–79. DOI: 10.1016/j.neucom.2017.11.077

[50] Chung S. Applications of smart technologies in logistics and transport: A review. *Transportation Research Part E: Logistics and Transportation Review.* 2021. Vol. 153, 102455. DOI: 10.1016/j.tre.2021.102455

[51] European Commission. *Sustainable Urban Mobility Package.* URL: https://ec.europa.eu/transport/themes/urban/ (last accessed: 01.02.2025).

[52] Kidmose B. A review of smart vehicles in smart cities: Dangers, impacts, and the threat landscape. *Vehicular Communications*. 2025. No. 51, 100871. DOI: 10.1016/j.vehcom.2024.100871

[53] Sood S. K. A scientometric analysis of quantum driven innovations in intelligent transportation systems. *Engineering Applications of Artificial Intelligence*. 2024. No. 138, 109258. DOI: 10.1016/j.engappai.2024.109258

[54] I. Jonek-Kowalska, Towards the Reduction of CO2 emissions. paths of pro-ecological transformation of energy mixes in European countries with an above-average share of coal in energy consumption. Resources Policy 77 (2022). doi: 10.1016/j.resourpol.2022.102701.

[55] R. Wolniak, I. Jonek-Kowalska, The level of the quality of life in the city and its monitoring, Innovation: The European Journal of Social Science Research 34(3) (2021) 376–398. URL: doi: 10.1080/13511610.2020.1828049.

[56] Z. Spicer, N. Goodman, D. Wolfe, How "smart" are smart cities? Resident attitudes towards smart city design, Cities 141 (2023) 104442. URL: doi: 10.1016/j.cities.2023.104442.

[57] Y. Dai, S. Hasanefendic, B. Bossink, A systematic literature review of the smart city transformation process: The role and interaction of stakeholders and technology, Sustainable Cities and Society 101 (2024) 105112. URL: doi: 10.1016/j.scs.2023.105112.

[58] A. Guenduez, I. Mergel, K. Schedler, S. Fuchs, C. Douillet, Institutional work in smart cities: Interviews with smart city managers, Urban Governance 2(1) (2024) 104-122. URL: doi: 10.1016/j.ugj.2024.01.003.

[59] I. Jonek-Kowalska, Housing Infrastructure as a Determinant of Quality of Life in Selected Polish Smart Cities, Smart Cities 5(3) (2022) 924–946. URL: doi: 10.3390/smartcities5030046.

[60] D. Taotao, D. John, Recent developments in bus rapid transit: a review of the literature, Transport Reviews 31 (1) (2011) 69–96. URL: doi: 10.1080/01441647.2010.492455