Periodica Polytechnica Transportation Engineering, 52(4), pp. 362–371, 2024

Applying Cluster Analysis for the Investigation of Travel Behavior and User Profiles

Matheus Moro Zamprogno[1], Domokos Esztergár-Kiss[1*]

[1] Department of Transport Technology and Economics, Faculty of Transportation Engineering and Vehicle Engineering, Budapest University of Technology and Economics, Műegyetem rakpart 3., H-1111 Budapest, Hungary
[*] Corresponding author, e-mail: esztergar-kiss.domokos@kjk.bme.hu

Received: 05 January 2024, Accepted: 30 July 2024, Published online: 02 August 2024

Abstract

Urbanization leads to a surge in demand for transportation and infrastructure improvements. In this context, understanding and optimizing travel behavior are crucial for effective transportation planning. This research investigates travel behavior patterns and user profiles in the realm of urban mobility. The study adopts an approach utilizing real-world data from an activity-based dataset collected through a survey. The methodological framework is characterized by a multi-step process which includes data preprocessing, cleaning, and aggregation, as well as principal component analysis and *k*-means cluster analysis with inertia evaluation for an optimal number of clusters. The cluster analysis unveils seven distinct clusters. Stability lovers are elderly people who prefer public transport, happiness seekers are attraction-driven car users, weekend shoppers, park goers, and sports practitioners rely on their cars for their activities, too. Furthermore, inflexible travelers value the service quality and "routine enthusiasts" stick to travel routines. Notably, bicycle usage prevails among stability lovers and routine enthusiasts, while shared transportation gets little attention in any of the clusters. By recognizing the adaptability of this methodology to specific city contexts, current research provides a way to understand travel behavior thus offering valuable insights for informed transportation policy planners.

Keywords

travel behavior, user profiles, clustering, preferences, activities

1 Introduction

Understanding travel behavior plays crucial role in urban and transportation planning and policymaking (Rashidi et al., 2015). Travelers aim to optimize their trips by doing the maximum number of activities and spending a minimum amount of time on using transport modes (Subbarao et al., 2020). These complicated patterns may require complex methods, such as machine learning algorithms.

Machine learning models have been already used to make predictions and classify data using statistical knowledge to perform operations in a supervised, unsupervised, or reinforced manner. The supervised method occurs when the input data are provided together with the output data. The unsupervised method is applied when the input data are provided alone to the model, and the reinforced manner takes place when the model learns through interaction. The supervised and unsupervised methods can be combined, where the mixed method is called semi-supervised (Koushik et al., 2020). Among the unsupervised methods, there is the *k*-means method, which provides data classification based on the distance of each individual to different centroids (Pedregosa et al., 2011). Therefore, this method groups individuals close to a common point and may determine similar attributes, which help identifying mobility patterns.

Moreover, data classification needs adequate data, which can be acquired by real-time sensing, research, observations, and surveys. In this research, a survey is applied to collect socio-demographic data and information on travel behavior to identify and classify the common traits. Therefore, this research work uses an unsupervised machine learning method called *k*-means cluster analysis to classify the travelers into groups and extract their shared attributes and travel behavior thus creating a profile description for each group together with an investigation of the choice of the number of clusters.

This study offers a valuable solution through the applied methodology. By utilizing a dataset, which comprises general activity behavior, the approach transcends

Cite this article as: Moro Zamprogno, M., Esztergár-Kiss, D. (2024) "Applying Cluster Analysis for the Investigation of Travel Behavior and User Profiles", Periodica Polytechnica Transportation Engineering, 52(4), pp. 362–371. https://doi.org/10.3311/PPtr.25663

the limitations of existing research works, which often relies solely on transit user data. Current study provides a more comprehensive understanding of travel behavior by encompassing such factors as activity frequency, spent time, transport mode, and socio-demographic profiles. The research work delves into not merely the travel behavior but into user profiles, as well. The adaptability of the methodology to specific localities and its potential to guide policy as well as land use decisions based on user profiles opens doors to innovative approaches in urban planning and transportation policy.

2 Literature review

Several authors (Jain and Tiwari, 2019; Koushik et al., 2020; Papadimitriou et al., 2017) have investigated activity-based data to understand people's travel behavior. Manoj and Verma (2015) aim to find similarities among non-workers in distinct household income groups and in socio-demographics, which influence their activity-travel choices. Therefore, activity-travel survey data are analyzed where travel indicators are studied, as well. Moreover, statistical models are presented to discover how sensitive the groups are to the behavior indicators. The work points out that low-income groups are more mobility-constrained, walk more, make more social stops, and do less recreational activities than members of the other income groups.

Socio-demographic data seems to be important in the exploration of mobility patterns. Hibino et al. (2020) use data from the Tokyo Person Trip Survey (Tokyo Metropolitan Area Transportation Planning Council, 2018) to identify differences in trip patterns based on gender, age, and household size. While using an activity-based model, the researchers aggregate the data and demonstrate them in graphs and tables, which show the trip characteristics of each group. Afterward, the considerations are summarized as results. The findings suggest that household composition is important in demand forecast given the differences between trip patterns and trends. Another important study by Subbarao et al. (2020) analyzes the travel behavior on weekdays and weekends and investigate the relationship of trip-chain and mode choice decisions on weekdays and weekends. Structural equation modeling is applied to understand the multidirectional relationship of 15-day activity-travel data of a survey. The results reveal that the mode choice is made before the decision on the trip chain and is more important than the trip chain on weekdays while they have equal importance during weekends. Moreover, the socio-demographic variables have significant impact on both mode and trip choices regardless of the day.

By aggregating a different path to the activity-based data, Victoriano et al. (2020) identify and characterize mobility strategy profiles from residents' seven-day activity, travel, expenditure, and social interaction diaries. An approach based on machine learning including self-organizing map and decision tree (DT) algorithms is used to identify mobility strategies characterized by travel behavior. The results show that the increased levels of private vehicle modal split tend to demonstrate higher levels of social interaction with individuals in the social circle than in case of public transport users.

In activity-based data, it is common that many different variables related to travel behavior and socio-demographics exist; thus, a dimensional reduction without losing important information is often required. Therefore, Jain and Tiwari (2019) use principal component analysis (PCA) to propose the application of a proxy variable of income on household data thus studying socioeconomic well-being scores (SEWS) and understanding the travel behavior of different income groups. The results show that the groups with low and middle SEWS rely more on walking and traveling short distances, which is different from the groups having high and very high SEWS.

Furthermore, PCA is used to understand emissions in the research conducted by Rashidi et al. (2015). The researchers present a framework using qualitative and quantitative approaches with PCA to estimate the emissions across the city by considering and comparing several pollutants and travel characteristics to investigate their relationship. Moreover, there is a clear indication that the urban variables examined by the scholars affect the emission rates.

To extract patterns from activity-based data, researchers combine the PCA with other methods. Xu and Lin (2016) model and optimize the selection process of transit projects by using multicriteria analysis and principal component analysis combined with a dynamic programming algorithm (PCA-DP). The mixed method covers the aspects of system performance evaluation from a payoff function generated by the PCA and used as input to the dynamic programming algorithm. Another combination of methods occurs with the introduction of logistic regression. Zhao et al. (2018) explore the attitudes of residents from Beijing, China, and it is shown how the cycling environment, travel behavior, urban distribution, and socio-demographic attributes influence them to choose

cycling or car purchasing. The researchers use PCA and multinomial logistic regression analysis to conclude that the cycling environment is related to future cycling behavior, short travel distances, low education and low income.

Furthermore, Gascon et al. (2020) examine connections between the built environment and public transport use in various cities of Europe by applying a PCA with logistic regression including several factors, such as personal characteristics, attitudes on transport, and frequency of transit use. The researchers conclude that a limitation on motorized vehicles in urban areas should be introduced, and investments in reliable and affordable transport services as well as improvements in residential built environments should be realized.

Matowicki et al. (2021) state that carsharing is used by travelers with different needs and the most relevant socio-demographic aspects influencing decision-making are the followings: family income, city, the number of available vehicles, attitudes, and perceptions. Moreover, the scholars use PCA and logistic regression to identify groups with various motivations aiming to uncover factors influencing users when picking carsharing as a transport mode.

Merging the PCA and clustering analysis is the topic of the study of Csáji et al. (2013) exploring the connections between the traits of human behavior. Data are extracted from an anonymized telecommunication dataset. By using the PCA and clustering methods, the researchers reduce the variable space and create groups that reveal most important geographical features and that people usually spend more time at only a few locations. Following the same direction, Papadimitriou et al. (2017) make an in-depth analysis of walking and crossing behavior influenced by human factors while highlighting primordial components of pedestrians' attitudes and identifying their profiles. PCA is applied on the travel characteristics, and through cluster analysis, two groups of individuals are identified. The results show that pedestrians can have positive and negative attitudes, and there are three principal components of pedestrian behavior: risk perception, risk-taking, and walking motivation.

Other researchers try to build profiles for the clusters resulted from the clustering methods. Egan et al. (2022) propose a typology for cycle parking by analyzing survey data in Ireland. The scholars use PCA to reduce variables, and cluster analysis is applied to make profiles, so that cycle parking classification is provided. Pronello and Camusso (2011) classify travelers based on their attitudes by using exploratory factor analysis and *k*-means cluster analysis to improve the comprehension of market segmentation and delineate transportation policies. The scholars classify the travelers as the followings: travel pleasure addicts, paying ecologists, time addicts, and timeservers. Practical market solutions, such as investing in public transport quality, making private transport uncomfortable, and promoting alternative modes are suggested as well.

Market segmentation is the topic of Li et al.'s (2013) study, who apply clustering methods. Attitudinal market segmentation is used to uncover some possible markets of the bicycle use by applying structural equation modeling and *k*-means cluster analysis on a survey that reveals travelers' commuting behavior. Moreover, the researchers divide the bicycle market into different segments and analyze the characteristics of each segment to promote cycling effectively. The *k*-means clustering provides a base for profiling each uncovered segment thus demonstrating that it is a relevant approach to understand the bicycle commuting activities. Furthermore, Rakha et al. (2014) investigate trip patterns within a chosen city and propose a methodology to generate urban trips from an activity-based travel survey by using PCA and *k*-means clustering to classify the profiles for mobility modeling while identifying activities people engage in and their travel inside the city. The researchers distinguish between the following four clusters: stay at home, worker, adventurer, and student.

These methods consider machine learning. Koushik et al. (2020) present an insightful work on this topic and explore the trends of travel behavior by reviewing the literature about analysis and modeling that applies machine learning techniques on activity-travel behavior. The study highlights the following methods: neural networks, support vector machines, decision tree, reinforcement learning, random forest, naive bayes classifier, recommender systems, PCA, and *k*-means cluster analysis. PCA and *k*-means seem not to be the most common techniques used in the field thus opening space for more investigation.

3 Method

Clustering is applied in various scientific fields including transportation. Hierarchical clustering methods order the specific items into clusters based on similar characteristics within the clusters (Saha et al., 2019). The clustering process can be agglomerative or divisive. In the former case items are collected into groups that form the clusters, in the latter case the items are separated into smaller clusters.

For the cluster analysis, an activity-based questionnaire was conducted in 2022 with the participation of 781 people from more European countries. The survey was distributed through direct e-mails, social media channels, and online

forums by using convenience sampling. The survey has six behavior-related questions and five socio-demographic-related questions where categorical and numerical answers are given. The questions assess the respondents' travel behavior related to leisure and general activities.

The dataset resulted from the survey is imported into a framework built in python to process the data and run the cluster analysis by using primarily Pandas and Scikit-learn libraries. In the data processing phase, the questions not answered are considered "not applicable". The participants with missing socio-demographic data are deleted since it is impossible to infer these data. Finally, due to the categorical variables, the dataset has to be transformed into a numerical format, which results in 375 components.

To reduce the number of variables and avoid the effect of outliers (Cheng, 2021), the data are merged. First, the "year of birth" variable is replaced by "age category" where each respondent is assigned to a category from 1 to 5, which represents age interval of each participant. The categories described in Table 1, have different interval distributions for the first and fifth categories to balance intervals.

The second merge intends to reduce the number of outliers in the data where some variables in the dataset with 375 components are logically combined according to the relative position and the number of individuals. This analysis is conducted by checking the count plot of each variable verifying what can be combined. As an example, in the variable "How often do you go to bar, pub, or café on average during a month?", the values "several times" and "almost every day" are merged. With the merged values, the dimensions of the data are reduced from 375 to 304.

To minimize the inertia of the analysis by reducing the dimensions of the data, the principal components of the dataset are identified by checking the total variance (43.72) in the dataset and extracting the number of variables responsible for 95% of this variance (40.62), which reveals 148 eigenvectors among 304. They are the vectors that do not turn when multiplied by the covariance matrix. The biggest eigenvalues correspond to those dimensions where the variance is the greatest; thus, they are the principal components (Abdi and Williams, 2010).

Table 1 Age categories

Year of birth interval	Age category	Age interval
1938–1958	1	80–60
1958–1968	2	60–50
1968–1978	3	50–40
1978–1988	4	40–30
1988–2000	5	30–18

Once the data processing phase is concluded, the cluster analysis is conducted. In this research, the *k*-means method, which is an algorithm that groups the individuals of equal variance while aiming for the minimization of inertia, is used (Sun et al., 2023). The method breaks a population down into a *k* number of clusters described by the centroids also known as means μ_j. The inertia can be interpreted as a measure of internal coherence of clusters where the less the inertia, the greater the coherence (Pedregosa et al., 2011). Given that, the algorithm chooses the initial centroids and loops between following two steps:

1. locate the nearest centroid for each individual,
2. assign new centroids by the mean of previous values.

This process is repeated until the centroids do not change significantly. Therefore, Eq. (1) is minimized.

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left(\left\| x_i - \mu_j \right\|^2 \right) \tag{1}$$

Furthermore, an analysis is conducted to investigate what would be the best value for *k*. The test consists of calculating the value of inertia in the dimensionally reduced dataset by using 148 principal components for *k* varying from 2 to 19 and verifying where the reduction curve, as shown in Fig. 1, starts to stabilize. The *k* range is limited to this interval as a primary investigation due to computational efficiency, but it showed to be enough for the analysis, since the number of clusters to be chosen were less than 10.

Additionally, it is important to highlight that the higher the number of clusters, the better the inertia. Still, if the inertia is taken to the extreme minimum, the clustering ends up in individuals and does not give any useful insight. Therefore, there must be a balance in choosing the value of *k*. Fig. 1 shows the results of the 18 runs and the variation in the inertia between the number of the clusters.

A first look at the inertia curve reveals that it has a smooth shape which makes it difficult to assess a point of stabilization in the reduction step. Thus, the curve of the variation between the clusters, shown in Fig. 2, is explored. Additionally, it is demonstrated that at the number of 7 clusters, the variation starts to stabilize suggesting a possible value for *k*.

The total number of participants is assigned to various groups, and distributions are generated and plotted for each variable thus revealing their characteristics.

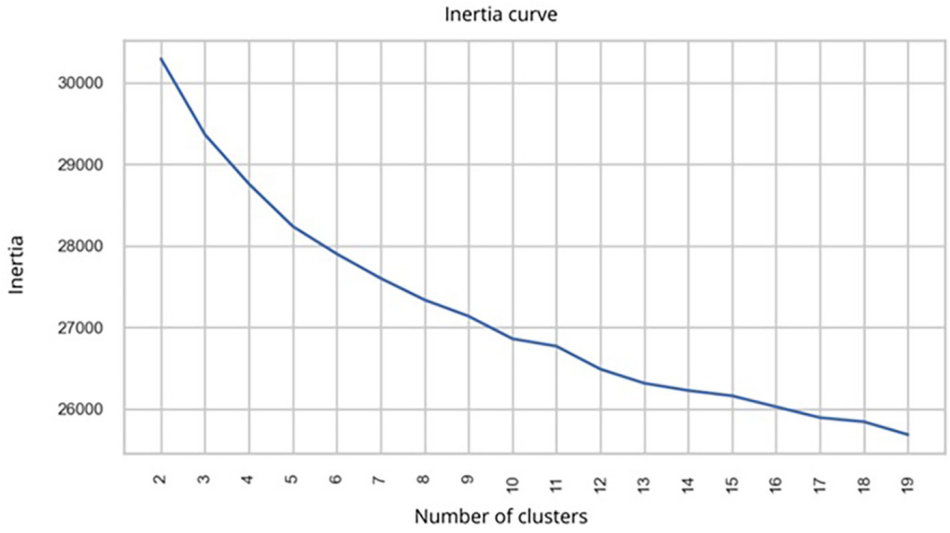

Fig. 1 Inertia curve

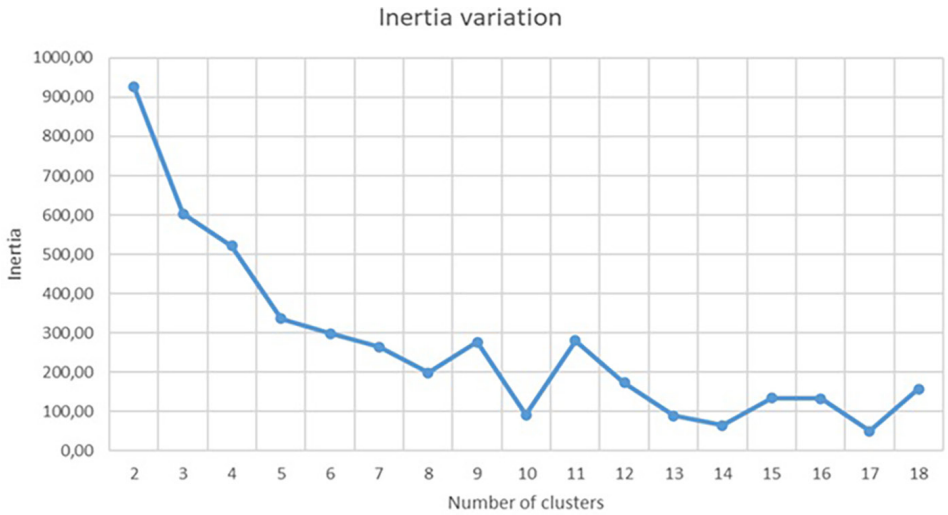

Fig. 2 Inertia variation

These plots, as demonstrated in Fig. 3, are to extract specific profiles that could explain most respondents' behavior. Although the assessment can be subjective, logical thinking is needed to establish a standard interpretation. Therefore, a major part of each cluster is analyzed with its neighbors, and one trait is extracted for the variable in specific clusters. Moreover, these traits constitute the characteristics of the clusters, and based on these, a profile, which specifies and describes the majority of the group, is formed for each group.

4 Results

In this section, the results of the *k*-means cluster analysis applied on the dataset derived from the survey are presented. The number of clusters is defined as seven, given the stabilization of the inertia variation curve, where a population of 781 is represented. This division shows a similar distribution between the clusters except for Cluster 1, as demonstrated in Table 2.

As the main objective of this research, the clusters can extract the participants' characteristics to build the group members' general preference and travel behavior profiles. As mentioned before, the characteristics are investigated, and the most common traits are collected in the descriptions of each cluster. A representative name is assigned to each group due to identification reasons.

4.1 Cluster 1 – Stability lovers

The members of this cluster have stable daily routines and use their free time to relax and meet friends or family. They prefer spending their leisure time in open spaces and avoid crowded places. While they are not flexible

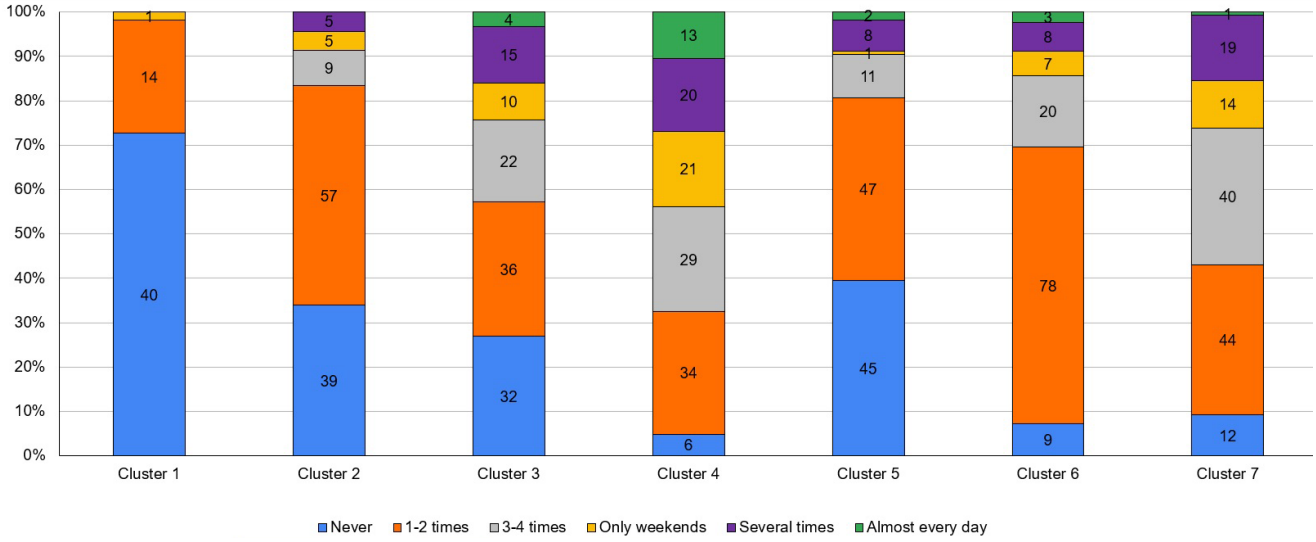

Fig. 3 Frequency clustering for bar, pub, and café

Table 2 Cluster distribution and participant profiles

Cluster number	Percent. of total	Flexibility	Age group	Leisure activities	Transp. modes
1	7.04%	fixed routines	elderly	open spaces	walking, public tr
2	14.72%	moderate	adults, elderly	shopping, family	car
3	15.24%	high	adults	sport	car
4	15.74%	very high	adults	restaurant, bar, park	walking, public tr
5	14.60%	extremely	elderly	not really	car
6	16.01%	not flexible	younger	sport, shopping, family	car
7	16.65%	partially	younger	restaurant, bar, sport, shopping, family	public tr

with their routines, they can adjust the time or location for activities outside their usual schedule. Moreover, when they are outside of their homes, they do not spend much time on activities and prefer nearby places to conduct them (Fig. 3). Despite owning cars, they often choose to walk or use public transportation for short distances. The members of this group live in small towns close to places they frequently visit, and belong to older generations with stable incomes.

4.2 Cluster 2 – Happiness seekers

The members of this cluster engage in a variety of activities during their leisure time, including shopping and visiting family and friends (Fig. 3). While they generally avoid crowded places, they occasionally make exceptions. In general, "happiness seekers" are flexible, except when it comes to activities related to attractions, which they prioritize and spend more time. They are also willing to travel longer distances to reach their destinations, typically driving their cars for these trips. This group consists of adults and elderly individuals with college degrees, living outside of big cities and having average incomes. Additionally, they have more cars, which are their favorite transport mode to use almost always (Table 2).

4.3 Cluster 3 – Sports practitioners

In this cluster (Fig. 3), members tend to realize leisure activities primarily related to sports, while they usually do not visit cinemas or concerts. The respondents are keen on shopping and spending time on visiting family and friends. They have a high degree of spatial and temporal flexibility, and they easily change their plans. Sports practitioners spend a reasonable time outdoors mainly conducting sports activities, but do not spend much time in bars and pubs. However, these participants like stay for longer times in restaurants and shopping facilities. They seem to move within a specific range, except for reaching their families, which can be far from them and take more time. Members in this group have college degrees and do not live in big cities. They consist of adults with average income, having more cars, and preferring driving these cars.

4.4 Cluster 4 – Parkgoers

In this cluster, members are enthusiastic about their leisure time (Fig. 3). They enjoy being outdoors, and meeting family and friends in parks, restaurants, and pubs. Sport activities and shopping are also important for them

due to their active lifestyle. Parkgoers are highly flexible with both the location and timing of their leisure activities, often based on the occasion. They favor outdoor spaces and parks, primarily seeking green areas. These respondents allocate a reasonable amount of time to leisure activities and handle routine tasks like visiting the ATM or post office efficiently. Parkgoers are willing to travel longer distances for obligatory tasks or special places but prefer leisure activities close to home. For nearby activities they usually walk or use public transport, reserving their cars for longer trips. The group is formed by younger people and adults with average incomes and college degrees, owning a car and residing in large towns (Table 2).

4.5 Cluster 5 – Weekend shoppers

Members of this cluster generally do not favor leisure activities (Fig. 3), but when they conduct such activities, they prefer visiting restaurants or parks. At the same time, they are very enthusiastic about shopping and spending time with family and friends. In case of any activity, they are extremely flexible regarding time and location. Weekend shoppers spend reasonable time outside of their homes but not very frequently. They enjoy dining out and often participate in shopping activities, which can be time-consuming. While they work at various distances from their homes, they prefer to spend their leisure time nearby, except when visiting museums or concerts, which may involve longer travel. This group consists of elderly individuals with average incomes. Some members live outside of big cities and typically have two cars per household.

4.6 Cluster 6 – Inflexible travelers

These travelers enjoy practicing sports, shopping, and visiting family and friends, but they rarely go to pubs, restaurants, or parks (Fig. 3). They also tend to avoid cinemas, theaters, and concerts, although they do spend time outside their homes when they choose to. These participants have limited spatial and temporal flexibility and are not fond of changes. The time they spend in pubs and parks varies by occasion, while shopping and eating out can take longer. Additionally, the time spent with family and friends is very important. They prefer nearby activities, making their travel time short, except for special occasions, like visiting museums and concerts, which may require longer trips. Moreover, their families usually live close by. This group consists of young individuals with average incomes and high school or college degrees, living in both small and big cities. They own more cars and often use them as their preferred transport mode (Table 2).

4.7 Cluster 7 - Routine enthusiasts

In this cluster, members enjoy practicing sports, shopping, and visiting family and friends. They often visit pubs, bars, and parks, but they do not prefer eating outside (Fig. 3). Having a calm social time seems important for these participants, so they typically avoid cinemas, theaters, and concerts. In case of social activities, they are flexible regarding time and space, but when it comes to ordinary tasks, their flexibility reduces especially in terms of time. Routine enthusiasts spend a reasonable time outside of their homes, and social events can be lengthy and distant. Going to the gym and doing sports are very important for this group. They prefer conducting ordinary activities quickly and close to their homes. Spending time with family and friends is crucial for them. They typically use public transport for all activities, but sometimes they also walk. This group includes young members living in metropolitan areas, where households generally own one car (Table 2).

5 Discussion

The results demonstrate some expected and some unexpected features within the clusters, which are worth highlighting. Stability lovers are represented by elder people who have car but prefer using public transport or walking instead of driving their cars. Therefore, this segment relies on the public transport network throughout the day to do their routine-related activities. However, happiness seekers love attractions in their leisure time, and the distance to travel does not matter for them. They travel by driving their cars; thus, weekends, holidays, and special dates mean that these people use the road network making the quality of infrastructure important for them.

Groups that prefer using cars are weekend shoppers, park goers, and sports practitioners, which is a surprising result. Weekend shoppers move from the suburbs and outskirts to the city to do shopping activities on weekends. Park goers use the public transport for their daily trips, but they prefer driving for long distances. Finally, sports practitioners prefer open-air activities and sports near their homes, but they are keen on using cars and no other transport modes.

A group that is likely to use public transport is inflexible travelers. As punctuality is of high importance for them, they may demand well-planned transportation services and reliable schedules. Moreover, these travelers prefer doing activities nearby. Similar to them are routine enthusiasts, who focus on their routines and ordinary tasks but are flexible when spending their leisure time. This segment may use public transport throughout the weekdays for going to

work and during the weekend for leisure activities. This pattern is also identified by Strömblad et al. (2022), who highlight a higher public transport usage in case of outdoor activities, but lower utilization in case of cultural events.

The survey demonstrates that stability lovers and routine enthusiasts are more likely to use bicycles due to their focus on their daily standard activities, but none of the segments show interest in shared transport services.

The environmental impacts due to the usage of private cars by weekend shoppers, park goers, sports practitioners increase traffic congestion, air pollution, and emissions. To mitigate these effects, promoting sustainable alternatives like public transport, cycling, and walking is crucial. For clusters that already prefer public transport, enhancing service quality and reliability, and realizing pedestrian-friendly measures can further reduce private car usage.

These results may provide adequate guide for planning transportation policies and investments in the fields of infrastructure and public services, considering the needs of all socioeconomic groups, particularly those who rely heavily on public transport or have limited access to private vehicles. The study can show demand for improvements in such areas as communication and support for elderly public transport users, service quality, road quality, the connection of the transportation network to public places, leisure areas, and dense commercial areas, as well as the deployment of sports facilities in the neighborhoods, and understanding the transportation needs and behavior changes. As an example, support for elderly people might be improved by offering dedicated assistance with boarding and alighting. The communication and support may be optimized by knowing when this group uses public transport.

Clustering algorithms are commonly used in mobility pattern investigation studies since they can group individuals with related characteristics. This research work pretreats the data to make them suitable to the algorithm. One of the limitations is solved by aggregating the data and eliminating the outliers, while the PCA use the categorical variables in the data (Koushik et al., 2020).

During the pre-tuning phase, choosing the k value is a challenging task because in this case, it is not clear how to identify the exact point where the decay starts to stabilize in the inertia curve. Since in cluster analysis, the objective is to look for common characteristics that form a group of individuals (Yadav and Sharma, 2013), the less the number of the groups that can describe the majority of the data is the better. Therefore, the bigger the number of the clusters the lower the inertia but at the same time, the higher the difficulty of identifying general patterns. Thus, there is a need to use a complementary approach to choose a limit for the inertia and consequently, to choose a k value.

Some limitations of the method are overcome by applying dimensional reduction techniques. However, it is challenging to extract the patterns from each cluster. In some cases, there are skewed distributions inside the clusters that reveal the main trait, but sometimes there are semi-uniform distributions, which make the process difficult. Moreover, the data presented in the analysis are from various countries, which makes it impossible to draw conclusions about a specific place thus helping the formation of plans and actions in a region.

To overcome limitations, future research should add a complementary qualitative or quantitative approach to the inertia method. The results demonstrate that some clusters have more latent differences than others. The use of data from specific countries or cities may give new possibilities in identifying the characteristics of local groups and in investigating reasons behind their behavior. Interviews, focus groups, and workshops, can provide deeper insights into the motivations and preferences of each cluster. Furthermore, the elaborated results can be used as a guideline to develop daily or weekly activity chains for the clusters. These qualitative insights can complement quantitative findings, offering a richer understanding of transportation behavior and supporting targeted, user-centered interventions.

6 Conclusion

The topic of current study is profile identification related to travel behavior where an unsupervised machine learning cluster analysis method is applied. A dataset built from an activity-based survey is used where the participants' travel habits and socio-demographic characteristics are gathered.

The analysis requires a pre-tuning of parameter k, which defines the number of clusters by describing how coherent the data are when examining their inertia. Therefore, the possibilities for the k value are assessed by merging the related variables and using PCA. As a result, seven clusters are defined. Afterward, the individuals are distributed to the clusters, and a description, which demonstrates the people's primary features and profiles, is elaborated for each group. The results demonstrate that the applied method is useful in distinguishing expected behavior in each group according to the members' traits, making it possible to generate profiles as mentioned in the results section.

In conclusion, using *k*-means cluster analysis to investigate travel behavior and user profiles not merely reveals the demand for improvements in areas ranging from support for elderly public transport users to infrastructure and service quality but empowers decision-makers to develop informed public policies. By utilizing a specific dataset and Python framework, the findings underscore the potential for tailored, locally applicable methods to enhance transportation systems thus bringing benefit to the society. with the improved public transport, infrastructure, and policies that address travelers' needs.

Acknowledgement
This research was supported by the János Bolyai Research Fellowship of the Hungarian Academy of Sciences (BO/00090/21/6). The linguistic revision of current paper is prepared by Eszter Tóth.

References

Abdi, H., Williams, L. J. (2010) "Principal Component Analysis", WIREs Computational Statistics, 2(4), pp. 433–459. https://doi.org/10.1002/wics.101

Cheng, Y. "Dimensionality reduction applied to logical judgments", [preprint] OSF Preprints, 03 August 2021. https://doi.org/10.31219/osf.io/k93fs

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., Blondel, V. D. (2013) "Exploring the mobility of mobile phone users", Physica A: Statistical Mechanics and its Applications, 392(6), pp. 1459–1473. https://doi.org/10.1016/j.physa.2012.11.040

Egan, R., Mark Dowling, C., Caulfield, B. (2022) "Planning for diverse cycling practices: A cycle-parking type preference typology", Case Studies on Transport Policy, 10(3), pp. 1930–1944. https://doi.org/10.1016/j.cstp.2022.08.007

Gascon, M., Marquet, O., Gràcia-Lavedan, E., Ambròs, A., Götschi, T., de Nazelle, A., …, Nieuwenhuisjen, M. J. (2020) "What explains public transport use? Evidence from seven European cities", Transport Policy, 99, pp. 362–374. https://doi.org/10.1016/j.tranpol.2020.08.009

Hibino, N., Yamashita, Y., Okunobo, N. (2020) "Fundamental analysis of trip patterns in urban area considering household composition in addition to gender and age", Transportation Research Procedia, 48, pp. 1583–1591. https://doi.org/10.1016/j.trpro.2020.08.200

Jain, D., Tiwari, G. (2019) "Explaining travel behaviour with limited socio-economic data: Case study of Vishakhapatnam, India", Travel Behaviour and Society, 15, pp. 44–53. https://doi.org/10.1016/j.tbs.2018.12.001

Koushik, A. N. P., Manoj, M., Nezamuddin, N. (2020) "Machine learning applications in activity-travel behaviour research: a review", Transport Reviews, 40(3), pp. 288–311. https://doi.org/10.1080/01441647.2019.1704307

Li, Z., Wang, W., Yang, C., Ragland, D. R. (2013) "Bicycle commuting market analysis using attitudinal market segmentation approach", Transportation Research Part A: Policy and Practice, 47, pp. 56–68. https://doi.org/10.1016/j.tra.2012.10.017

Manoj, M., Verma, A. (2015) "Activity-travel behaviour of non-workers belonging to different income group households in Bangalore, India", Journal of Transport Geography, 49, pp. 99–109. https://doi.org/10.1016/j.jtrangeo.2015.10.017

Matowicki, M., Pribyl, O., Pecherkova, P. (2021) "Carsharing in the Czech Republic: Understanding why users chose this mode of travel for different purposes", Case Studies on Transport Policy, 9(2), pp. 842–850. https://doi.org/10.1016/j.cstp.2021.04.003

Papadimitriou, E., Lassarre, S., Yannis, G. (2017) "Human factors of pedestrian walking and crossing behaviour", Transportation Research Procedia, 25, pp. 2002–2015. https://doi.org/10.1016/j.trpro.2017.05.396

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., …, Duchesnay, E. (2011) "Scikit-learn: Machine learning in Python", Journal of Machine Learning Research, 12, pp. 2825–2830.

Pronello, C., Camusso, C. (2011) "Travellers' profiles definition using statistical multivariate analysis of attitudinal variables", Journal of Transport Geography, 19(6), pp. 1294–1308. https://doi.org/10.1016/j.jtrangeo.2011.06.009

Rakha, T., Rose, C. M., Reinhart, C. F. (2014) "A framework for modeling occupancy schedules and local trips based on activity based surveys", In: Proceedings of ASHRAE/IBPSA-USA Building Simulation Conference, Atlanta, GA, USA, 2014, pp. 433–440. [online] Available at: https://publications.ibpsa.org/conference/paper/?id=simbuild2014_55 [Accessed: 04 January 2024]

Rashidi, T. H., Kanaroglou, P., Toop, E., Maoh, H., Liu, X. (2015) "Emissions and built form - An analysis of six Canadian cities", Transportation Letters, 7(2), pp. 80–91. https://doi.org/10.1179/1942787514Y.0000000036

Saha, R., Tariq, M. T., Hadi, M., Xiao, Y. (2019) "Pattern recognition using clustering analysis to support transportation system management, operations, and modeling", Journal of Advanced Transportation, 2019(1), 1628417. https://doi.org/10.1155/2019/1628417

Strömblad, E., Winslott Hiselius, L., Smidfelt Rosqvist, L., Svensson, H. (2022) "Characteristics of Everyday Leisure Trips by Car in Sweden – Implications for Sustainability Measures", Promet – Traffic&Transportation, 34(4), pp. 657–672. https://doi.org/10.7307/ptt.v34i4.4039

Subbarao, S. S. V., Swaroop, S. N. V., Shekhar, R. S. (2020) "Interrelationships between mode choice and trip-chain choice decisions: A case of Mumbai Metropolitan Region", Transportation Research Procedia, 48, pp. 3049–3061. https://doi.org/10.1016/j.trpro.2020.08.182

Sun, S., Bi, R., Wang, Z., Ji, Y. (2023) "Loading and Unloading Points Identification Based on Freight Trajectory Big Data and Clustering Method", Promet – Traffic&Transportation, 35(2), pp. 148–160. https://doi.org/10.7307/ptt.v35i2.106

Tokyo Metropolitan Area Transportation Planning Council (2018) "The 6th Tokyo Person-Trip Survey", [online] Available at: https://www.tokyo-pt.jp/person/01 [Accessed: 04 January 2024] (in Japanese)

Victoriano, R., Paez, A., Carrasco, J.-A. (2020) "Time, space, money, and social interaction: Using machine learning to classify people's mobility strategies through four key dimensions", Travel Behaviour and Society, 20, pp. 1–11. https://doi.org/10.1016/j.tbs.2020.02.004

Xu, W., Lin, W. (2016) "Selecting the public transit projects with PCA-DP technique: The example of Xiamen City", Transport Policy, 46, pp. 56–71. https://doi.org/10.1016/j.tranpol.2015.11.002

Yadav, J., Sharma, M. (2013) "A review of k-mean algorithm", International Journal of Engineering Trends and Technology, 4(7), pp. 2972–2975.

Zhao, C., Nielsen, T. A. S., Olafsson, A. S., Carstensen, T. A., Fertner, C. (2018) "Cycling environmental perception in Beijing – A study of residents' attitudes towards future cycling and car purchasing", Transport Policy, 66, pp. 96–106. https://doi.org/10.1016/j.tranpol.2018.02.004