Original Article

A Fuzzy Clustering Approach to Identify Pedestrians' Traffic Behavior Patterns

Parisa Saeipour (MSc)[1] , **Parvin Sarbakhsh (PhD)[1]*** , **Saman Salemi (MD)[2], Fatemeh Bakhtari Aghdam (PhD)[3]**

[1]Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Iran
[2]Department of Medicine, Islamic Azad University Tehran Medical Sciences, Tehran, Iran
[3]Road Traffic Injury Research Center, Tabriz University of Medical Sciences, Tabriz, Iran

Abstract

Background: Pattern recognition of pedestrians' traffic behavior can enhance the management efficiency of interested groups by targeting access to them and facilitating planning via more specific surveys. This study aimed to evaluate the pedestrians' traffic behavior pattern by fuzzy clustering algorithm and assess the factors related to higher-risk traffic behavior of pedestrians.

Study Design: This study is a secondary methodological study based on the data from a cross-sectional study.

Methods: The fuzzy c-means (FCM), as a machine learning clustering method, was conducted to identify the pattern of traffic behaviors by collecting data from 600 pedestrians in Urmia, Iran via "the Pedestrian Behavior Questionnaire" (PBQ) and using 5 domains of PBQ. Multiple logistic regression was fitted to identify risk factors of traffic behaviors.

Results: Results revealed two clusters consisting of lower-risk and higher-risk behaviors. The majority of pedestrians (64.33%) were in the lower-risk cluster. Subjects ≤ 33 years old (Odds ratio [OR] = 1.92, $P < 0.001$), subjects with ≤ 6 years of education (OR = 1.74, $P = 0.010$), males (OR = 1.90, $P = 0.001$), unmarried pedestrians (OR = 3.61, $P = 0.007$), and users of public transportation (OR = 2.01, $P = 0.002$) were more likely to have higher-risk traffic behavior.

Conclusion: We identified traffic behavior patterns of Urmia pedestrians with lower-risk and higher-risk behaviors via FCM. The findings from this study would be helpful for policymakers to promote safety measures and train pedestrians.

Please cite this article as follows: Saeipour P, Sarbakhsh P, Salemi S, Bakhtari Aghdam F. A fuzzy clustering approach to identify pedestrians' traffic behavior patterns. J Res Health Sci. 2023; 23(3):e00592. doi:10.34172/jrhs.2023.127

Background

According to the reports of the World Health Organization (WHO), traffic accidents are one of the most common causes of injury and death.[1-3] Existing research has found that a wide range of factors may contribute to the occurrence and severity of pedestrian crashes, including behavioral, roadway, and environmental factors.[4,5]

Traffic behavior of pedestrians has always been one of the most important issues affecting traffic accident mortality. The risky behavior of a pedestrian can directly increase the chances of a traffic crash. Pattern recognition of traffic behaviors of pedestrians and identifying homogeneous groups have been the research subjects in several prior studies, and the traffic behavior analysis has always attracted considerable interest from transport authorities since understanding pedestrians in terms of their traffic behavior patterns can help the planning and implementation of better services. In addition, identifying such patterns of traffic behavior can enhance management efficiency via more targeted access to groups of interest and facilitate planning through more specific surveys.

Furthermore, identifying clusters of pedestrians where each cluster presents a similar traffic behavior pattern can help to perform a more detailed evaluation through complementary analysis such as logistic regression to identify factors affecting their behavior and predict the probability of a pedestrian exhibiting more risky traffic behavior.[6]

On the other hand, pedestrians' traffic behavior data is too heterogeneous, and this heterogeneity makes it difficult to identify certain patterns and evaluate factors that significantly impact these behavioral patterns. Such heterogeneity may lead to biased estimation of parameters and potentially incorrect conclusions.[7,8] Therefore, by building and applying more efficient models for pattern recognition in traffic data, researchers try to overcome the heterogeneity in these data to some extent.[9-11] Furthermore, data segmentation techniques such as cluster analysis can help to explore hidden patterns in complex data sets[12] such as pedestrians' behavior data by reducing the heterogeneity of data.[13,14]

There are two kinds of clustering methods: hard and

soft clustering. Hard clustering methods similar to the k-means algorithm are suitable for limited clustering tasks in which each data point belongs to only one group. On the other hand, soft clustering method such as fuzzy clustering is appropriate for overlapping clustering task, so data can belong to all clusters with a certain value of membership and it can give descriptions of objects in clusters in more detail. Fuzzy clustering is appropriate for data with complex structures or when there are vague or overlapping class boundaries. Moreover, fuzzy clustering can be more robust to outliers and noise in data. Choosing an appropriate clustering algorithm solely depends on the data type to be clustered and the purpose of the clustering applications.[15]

Some data sets cannot be adequately split into some non-overlapping clusters, while partitions may overlap with each other to some degree, and some data points contribute to more than one cluster. Fuzzy clustering algorithms are helpful for datasets with subgroupings of points having indistinct boundaries and overlap between the clusters such as human behavior data. Human behavior (e.g., traffic behavior) is an inherently complex subject; so, it cannot be clearly clustered into completely separate clusters, and there will be overlap between the clusters so that a person's behavior can potentially belong to multiple clusters of behavior. For these cases, we can use fuzzy clustering as a membership value-based clustering method which allows an object to be a member of more than one cluster but with different membership degrees.[15]

As we stated, it is crucial to use machine learning methods such as clustering method to identify traffic behavior patterns and investigate their effective factors. Hence, this study aimed to identify the hidden pattern of pedestrians' behaviors in Urmia, Iran, by using fuzzy cluster analysis and assess the factors affecting this pattern.

Methods
Participants
This study is a secondary study based on the data collected by Bakhtari et al (preprint).[16] This descriptive cross-sectional study was carried out among participants aged 18 years and above (N = 600) living in Urmia, Iran, from May to October 2018. In this study, the cluster sampling method was applied for sampling, so the health centers were considered clusters, and some of them were randomly selected. Then, from each selected center, depending on the population covered, the participants were randomly selected in terms of the inclusion and exclusion criteria.

The inclusion criteria were being 18 years old or above, willing to participate in the study, and being capable of standing and walking. The exclusion criteria were having a history of severe mental illness, depression, Alzheimer's, dementia, restrictive musculoskeletal disorders, neurological deficits (stroke), Parkinson's and paralytic disease, acute myocardial infarction, uncontrolled hypertension, and severe hearing and visual impairment. These exclusion criteria were checked by their self-reported medical records.

Selected subjects that were willing to participate in the study, were invited to visit the health centers. Based on a previous study[16] and considering the standard deviation of pedestrian behaviors, cluster sampling design effect, as well as the rate of incomplete questionnaires, they estimated the sample size at 600.

Data collection and questionnaire
Some information including demographic characteristics (age, gender, marital status, and education), pedestrian traffic behavior, walking minutes/day, and transportation mode were gathered by relevant questioners.

Pedestrian traffic behavior data were collected using the "Pedestrians Behavior Questionnaire" (PBQ) which is a valid and reliable instrument.[17] PBQ includes 29 items measuring traffic behavior with a five-point Likert scale from 1 to 5 (1 = never, 2 = rarely, 3 = sometimes, 4 = often, and 5 = always). It represents 5 domains of traffic behavior: (1) adherence to traffic rules (7 items) (e.g., I cross the street after the vehicles are fully stopped), (2) traffic violations (10 items, reverse scored) (e.g., I don't use the pedestrian bridge because most people don't), (3) positive behaviors (2 items) (e.g., I let the vehicle pass even if I have the priority right), (4) traffic distraction (4 items, reverse scored) (e.g., I use my mobile phone while crossing the street), and (5) aggressive behaviors (2 items, reverse scored) (e.g., If I get angry at the behavior of a driver, I would kick or punch the car). The score of each domain was calculated as the mean score of its items, and the mean scores of all 29 items showed the total score of PBQ. Hence, the score of each domain of PBQ and the total score of PBQ ranged from 1 to 5. A higher score in all domains and total score indicated better traffic behavior and vice versa.

Statistical analyses
Statistical analyses were performed using R software version 4.2.2 (packages: fclust[18]). Quantitative and qualitative variables were presented as mean and standard deviation (SD) as well as number and percentage, respectively. This study used fuzzy c-means (FCM) clustering for clustering pedestrians' behavior. The silhouette index, partition entropy, partition coefficient (PC), and modified partition coefficient (MPC) were used to select the optimal number of clusters.[19-21] Furthermore, the chi-square test and multiple logistic regression were used to examine the factors related to the obtained clusters.

Clustering task
Cluster analysis tries to separate data into groups or clusters such that both the homogeneity within the clusters and the heterogeneity between clusters are maximized.[22] This technique is an unsupervised machine learning algorithm because it uses machine learning algorithms to cluster unlabeled data and discover unknown patterns.[23,24] Clustering methods are distinguished regarding how they allocate data to clusters and are divided into two categories:

soft clustering (overlapping clustering) and hard clustering (exclusive clustering). In classical or hard cluster analysis, each data must be allocated to exactly one cluster while soft cluster analysis techniques allow overlapping clusters.[25]

K-means clustering belongs to the hard clustering, and FCM clustering belongs to the soft clustering category. In k-mean cluster analysis, each data point belongs to only one cluster with membership values of zero or one, while in fuzzy cluster analysis, the membership value of assigning a data point to clusters is between [0, 1], so data can belong to more than one cluster simultaneously with certain membership values, and it can give descriptions of objects in clusters in more detail.[15,26-29]

Hard clustering approaches such as k-means are simple, easy to modify, and less complex to interpret, but they are sensitive to the centroid initialization and outlier,[30,31] and FCM is more flexible than conventional k-means. Although soft clustering such as FCM is supposedly slower and computation time increases more rapidly for FCM than for the k-means algorithm with the growing number of clusters and sample size, this should not be of concern with the power of today's computers.[32]

K-means clustering algorithm

The k-means algorithm is an iterative algorithm that tries to partition the dataset into the fixed number (k) of distinct non-overlapping clusters in a dataset where each data point belongs to only one cluster. This algorithm attempts to make the intra-cluster data points as similar as possible while also maintaining the clusters as different as possible. It allocates data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is the minimum. The less variation within clusters, the more homogeneous the data points are within the same cluster.[23,24]

Fuzzy C-means clustering algorithm

The FCM is the weighted sum of squared errors within clusters, which is defined as follows:

$$J_m(U,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c} u_{ik}^m \, ||x_k - v_i||_A^2, 1 < m < \infty, \quad (1)$$

where $V = (v_1, v_2, \ldots, v_c)$ is a vector of the unknown cluster (centers) $v_i \in \mathbb{R}^p$. The value of u_{ik} represents the degree of membership of the data point x_k of set $X = \{x_1, x_2, \ldots, x_n\}$ to the ith cluster.

A measure of similarity between a data point and the cluster prototypes as the inner product was defined by a norm matrix A. A fuzzy c-partition of X [33] is suitably represented by a matrix $U = [u_{ik}]$. That if $||x_k - v_i||_A \rangle 0$ for all i and k, then (U,V) may minimize J_m, when $m > 1$ and

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \, for \, 1 \le i \le c , \quad (2)$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{||x_k - v_i||_A^2}{||x_k - v_j||_A^2}\right)^{\frac{1}{m-1}}} \, for \, 1 \le i \le c, \le k \le n. \quad (3)$$

The Picard iteration approach minimizes J_m by initializing the matrix U randomly and computing the (Eq2) and (Eq3) after each iteration. The Picard iteration method is an iterative algorithm used in FCM clustering to update the membership values and cluster centers. It is an extension of the classic c-means clustering algorithm that allows for fuzzy membership values, indicating the degree of belongingness of each data point to each cluster. When iteration reaches a stable condition, it is terminated; that is, when the changes in the cluster centers or the membership values at two successive iterations are smaller than a predefined threshold value.

The FCM algorithm always converges to a minimum point. A different initial guess of u_{ij} may lead to a different minimum. Finally, to allocate each data point to a particular cluster, defuzzification is necessary, and this can be done by assigning a data point to a cluster for whh the value of the membership is maximal. Defuzzification is the process of converting fuzzy membership values into crisp values in FCM.[15,26-29]

Clustering validity indices

Clustering validity indices are employed for the quality evaluation of partitions produced by clustering algorithms as well as for determining the number of clusters in the dataset.[20,34-37] They are calculated for various numbers of clusters, and the optimal number of clusters is determined by comparing the values of an index for all possible numbers of clusters. The following section outlines a brief explanation of the more commonly used fuzzy clustering validity indices that are used to measure the clustering performance and determine the optimal number of clusters.

Partition coefficient

In fuzzy clustering, a PC that was initially designed by Bezdek[20] is defined as:

$$PC = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N} u_{ij}^2$$

The PC values range in [1/c, 1]. The closer the index value is to 1, the clearer the clustering, and 1/c indicates that there is no clustering tendency in the considered dataset.

Modified partition coefficient

A PC correction (MPC)[38] is defined using a linear transformation to remove the dependence of PC on c. The modified PC is expressed as:

$$MPC = \frac{c}{c-1}PC - \frac{1}{c-1}$$

The range of MPC is [0, 1] where MPC = 0 corresponds to maximum fuzziness and MPC = 1 to a hard partition.

Partition entropy coefficient

The partition entropy (PE) coefficient[19] is defined as:

$$PE = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}\,logu_{ij}$$

The PE index values range in [0, log c]. The closer the value of PE to 0, the clearer the clustering is. The index value close to the upper bound (i.e., log c) indicates the lack of any clustering structure in the data sets.

Silhouette index
The Silhouette coefficient combines the factors of intra-cluster polymerization and inter-cluster resolution to measure the clustering effect.[21] The Silhouette index gets the optimal clustering number by computing the difference between the average distance within the cluster and the minimum distance between the clusters; that is, the optimal clustering effect, which is defined as:

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}\right)$$

where a (i) represents the average distance of sample i to other samples in the cluster, and b (i) represents the minimum distance of the sample from the sample i to the other clusters.

Results
Table 1 depicts a summary of the data, including detailed information and descriptions of the traffic behavior of the participants, as well as the frequency and percentage of their demographic characteristics. Demographic characteristics are age, gender, marital status, education, transportation characteristics, and pedestrian traffic behavior variables are adherence to traffic rules, traffic violation, positive behavior, traffic distraction, aggressive behavior, and total score of PBQ.

The optimal number of clusters
Pedestrian behavior was clustered based on variables such as adherence to traffic rules, traffic violation, positive behavior, traffic distraction, aggressive behavior, and total PBQ score. To select the optimal number of clusters (c), different numbers of clusters were tested from two to six. The lower and upper limit of the number of clusters was determined according to the previous studies and validation indexes. The fuzzy silhouette index, PE, PC, and modified PC were used to select the optimal number of clusters, which is presented in Table 2. For c = 2, the fuzzy silhouette index, PC, and modified PC were larger than other cluster numbers (a higher score is better), and for c = 2, PE was lower (a lower score is better). As a result, the optimal number of clusters was determined "2".

Clustering description
Table 3 illustrates the traffic behavior of the participants according to revealed clusters and statistical differences between them. As shown, the first cluster (C1) had lower scores in adherence to traffic rules, violation, positive behavior, distraction, aggressive behavior, and total PBQ

compared to the second cluster; therefore, we can name it the higher-risk traffic behavior cluster. Of the investigated participants, 214 (35.66%) of them belonged to this cluster. The second cluster (C2) had higher traffic behavior scores for total PBQ and its domains compared to the first cluster, so we can regard this cluster as the lower-risk traffic behavior cluster. Most of the participants belonged to this cluster (64.33%).

The mean ± SD of the total PBQ score of the pedestrian in the lower-risk cluster was 4.10 ± 0.29, while it was

Table 1. Descriptive information of pedestrians' behavior and their demographic characteristics

Quantitative variables	Mean	SD
Adherence to traffic rules	3.48	0.73
Traffic violation	3.93	0.68
Positive behavior	3.61	0.74
Traffic distraction	4.04	0.79
Aggressive behavior	4.32	0.95
Total Score PBQ	3.80	0.51
Qualitative variables	**Number**	**Percent**
Gender		
Male	239	39.8
Female	361	60.2
Marital status		
Single	155	25.8
Married	423	70.5
Other	22	3.7
Education status		
Illiterate	11	1.8
1-6 grades	91	15.2
7-12 grade diploma	214	35.7
Associate	66	11.0
Bachelor	150	25.0
Master	49	8.2
Doctoral and higher	19	3.2
Walking (min)		
<30	263	43.8
30-60	199	33.2
60-120	92	15.3
120	46	7.7
Transportation		
Personal car	392	65.3
Taxi	84	14.0
Large vehicle	51	8.5
Motorcycles and bicycles	12	2.0
Walking	61	10.2
Age (y)		
18-25	123	20.5
26-33	185	30.8
34-41	153	25.5
≥42	139	23.2

Table 2. The value of cluster validity indexes for choosing the optimal number of clusters

Cluster Number (c)[a]	FSI	PE	PC	MPC
2	0.639772	0.477040	0.691147	0.382293
3	0.485307	0.850437	0.493321	0.239982
4	0.408518	1.096336	0.400192	0.200256
5	0.436820	1.295305	0.340591	0.175739
6	0.423527	1.465443	0.294850	0.153819

Note. FSI: Fuzzy silhouette index; PE: Partition entropy; PC: Modified partition; MPC: Modified partition coefficient.
[a] for c = 2, the FSI, PC, and MPC are larger than other cluster numbers (a higher score is better), and PE is lower (a lower score is better).

Table 3. Mean and SD of traffic behavior in each cluster

Fuzzy c-means	Cluster 1 = 214 (Higher risk)		Cluster 2 = 386 (Lower risk)		P value
	Mean	SD	Mean	SD	
Adherence to traffic rules	3.12	0.70	3.68	0.67	0.001
Traffic violation	3.34	0.61	4.26	0.46	0.001
Positive behavior	3.10	0.68	3.89	0.60	0.001
Traffic distraction	3.40	0.76	4.40	0.53	0.001
Aggressive behavior	3.36	0.98	4.86	0.31	0.001
Total score PBQ	3.25	0.30	4.10	0.29	0.001

Note. SD: standard deviation; PBQ: Pedestrian behavior questionnaire.

3.25 ± 0.30 in the higher-risk cluster ($P < 0.001$). Moreover, all domains of PBQ significantly differed in the two clusters ($P < 0.001$), as depicted in Table 3.

According to the results of the chi-square test, there were significant differences between lower-risk and higher-risk clusters according to the pedestrians' age, education, gender, marriage status, and kind of vehicle used for transportation which indicates a relationship between these factors and pedestrians' traffic behavior pattern; therefore, the proportion of female pedestrians in lower-risk cluster was higher than that in higher-risk cluster (65.3% vs. 50.9%). Married pedestrians were more prevalent in the lower-risk cluster (76.4% vs. 59.8%), and more than 71.2% of the pedestrians in the lower-risk cluster and 54.7% in the higher-risk cluster used personal cars. Moreover, more than 54.6% of the participants in the lower-risk cluster were over 34 years old, while this rate was 37.9% in the higher-risk cluster. About 51.1% of lower-risk participants vs. 38.8% of higher-risk pedestrians had academic education which indicates their safer traffic behavior (Table 4).

Results of multiple logistic regression to assess independent predictors of having higher-risk traffic behavior demonstrated the significant effects of age, gender, marital status, type of transportation in the city, and education on being in the higher-risk cluster of traffic behavior (compared to the lower-risk cluster).Hence, subjects ≤ 33 years old compared to > 33 years old were more likely to have higher-risk traffic behavior (Odds ratio [OR] = 1.92, 95% confidence interval [CI]: 1.33-2.75, $P < 0.001$). Subjects with primary education or less compared to secondary or higher-educated pedestrians

were more likely to be in the higher-risk cluster (OR = 1.74, 95% CI: 1.10- 2.74, $P = 0.010$). Furthermore, male pedestrians had higher odds of more risky traffic behavior compared to females (OR = 1.90, 95% CI: 1.31-2.75, $P = 0.001$). In addition, unmarried pedestrians compared to married people (OR = 3.61, 95% CI: 1.40- 9.23, $P = 0.007$) and users of public transportation compared to users of personal cars (OR = 2.01, 95% CI: 1.30-3.08, $P = 0.002$) were more likely to be in higher-risk traffic behavior cluster (Table 5).

Discussion

Pedestrian behavior plays an important role in pedestrian safety. In this study, we first clustered the behaviors of pedestrians based on PBQ domains using the fuzzy

Table 4. Distribution of underling variables according to the clusters

Variables	Cluster 1 = Higher-risk		Cluster 2 = Lower-risk		P value
	Number	Percent	Number	Percent	
Gender					0.001
Male	105	49.1	134	34.7	
Female	109	50.9	252	65.3	
Age (y)					0.001
18-25	55	25.7	68	17.6	
26-33	78	36.4	107	27.7	
34-41	44	20.6	109	28.2	
>42	37	17.3	102	26.4	
Marital status					0.001
Single	72	33.6	83	21.5	
Married	128	59.8	295	76.4	
Other	14	6.5	8	2.1	
Educational status					0.044
Illiterate	5	2.3	6	1.6	
1-6 grades	45	21.0	46	11.9	
7-12 grade diploma	77	36.0	137	35.5	
Associate	24	11.2	42	10.9	
Bachelor	44	20.6	106	27.5	
Master	15	7.0	34	8.8	
Doctoral and higher	4	1.9	15	3.9	
Transportation					0.001
Personal car	117	54.7	275	71.2	
Taxi	43	20.1	41	10.6	
Large vehicle	22	10.3	29	7.5	
Motorcycles and bicycles	9	4.2	3	0.8	
Walking	23	10.7	38	9.8	
Walking (min)					0.080
<30	86	40.2	177	45.9	
30-60	66	30.8	133	34.5	
60-120	42	19.6	50	13.0	
>120	20	9.3	26	6.7	

Note. SD: Standard deviation.

Table 5. Multiple logistic regression to assess independent predictors of having higher-risk traffic behavior

Variables	B	SE	*P* value	OR (95% CI)
Gender				
Male	0.64	0.18	0.001	1.90 (1.31, 2.75)
Female				1.00
Age (y)				
≤33	0.65	0.18	0.001	1.91 (1.33, 2.75)
>33				1.00
Marital status				
Unmarried	1.28	0.48	0.007	3.60 (1.40, 9.23)
Married				1.00
Educational status				
Illiterate or 1-6 grades	0.55	0.23	0.017	1.74 (1.10, 2.74)
>7 grades				1.00
Transportation				
Public transportation	0.69	0.21	0.002	2.01 (1.30, 3.07)
Walking or biking	0.42	0.28	0.142	1.52 (0.86, 2.68)
Personal car				1.00
Walking (min)				
<30	0.11	0.36	0.754	1.12 (0.54, 2.28)
30-60	-0.04	0.36	0.904	0.95 (0.47, 1.94)
60-120	0.27	0.39	0.485	1.31 (0.60, 2.85)
>120				1.00

Note. SE: Standard error, OR: Odds ratio; CI: Confidence interval.

clustering method. According to the validation indices, the optimum number of clusters was 2. Cluster analysis with two clusters revealed two behavioral patterns; that is, pedestrians in the first cluster had a lower score of PBQ, and their traffic behavior was riskier, while pedestrians' behaviors in the second cluster were safer, and they obtained a higher score of PBQ and its domains. Afterward, we assessed the association between underlying factors (e.g., demographic characteristics, type of transportation, and the like) and unsafe behavior using multiple logistic regression. The results demonstrated the significant effect of age, gender, marital status, type of transportation in the city, and education on being in the higher-risk cluster of traffic behavior. Clustering our pedestrians' behavior dataset into two homogeneous subsets helped to identify associated factors that are not easily detectable when using the dataset as a whole.

As demonstrated in this study, clustering techniques can be used not only for descriptive analysis but also as a prepossessing segmentation tool for a more detailed standard statistical analysis.[6] Clustered data rather than the raw dataset can provide clear and more meaningful information.[8,39-42]

The application of FCM clusters is not limited to the traffic behavior of pedestrians. We can apply it to all types of traffic data to find better solutions for improving traffic safety. Due to the complex nature of traffic data such as high dimensionality, spatial-temporal structure, and overlapping clusters, cluster fuzzy clustering has been applied frequently.[43-46] Furthermore, researchers have used k-means clustering algorithms to identify homogeneous coincidence clusters.[42,47,48] In a study,[41] Latent class cluster and multinomial logit models were used to investigate the statistical relationship between pedestrian injury severity outcome and contributing factors (e.g., pedestrian behavior, demographics, accident characteristics, and the built environment). According to the obtained results, there is a relationship between severe accidents and variables such as using alcohol or drugs, age over 65, and adverse weather conditions.

Depaire et al[14] succeeded in investigating the performance of latent class clustering for traffic accident segmentation. The clusters obtained from the types of traffic accidents were sensible and could examine the effect of variables such as the age and type of road on traffic accidents.

However, regarding the pattern of pedestrian behavior, to the best of our knowledge, this study is the first one that used cluster analysis to identify patterns of pedestrians in terms of their traffic behavior. For analyzing pedestrian behavior, other statistical methods such as binary logit, ordered logit or probit, mixed logit, and multinomial logit models have been used.[9,49-53] Hence, we did not find any similar study to compare our results with them due to the different nature of the results.

Regarding revealed clusters, according to the scores of PBQ and its domains in two clusters, the clusters were named lower-risk and higher-risk pedestrians, and about 35% of pedestrians belonged to the higher-risk cluster. According to the range of scores of the questionnaire and its dimensions, which are between 1 to 5, the pedestrians in both clusters had scored higher than the middle score (score 3) and had acceptable traffic behavior in comparison with each other. Therefore, we can conclude that there are two clusters of pedestrians in Urmia: one with safe and cautious traffic behavior (total PBQ score of 4.1) and the other with moderate traffic behavior (total PBQ score of 3.25).

This finding can help develop educational and intervention programs for pedestrians as we encounter two groups of the population with moderate and good traffic behavior, so planning and policymaking should be performed considering these two groups.

Regarding factors affecting these traffic behavior patterns, the current results indicated that age, gender, marital status, type of transportation in the city, and education are related to the pattern of traffic behavior. In this study, female pedestrians had safer traffic behavior than males, and this finding is consistent with previous studies, which show males have more risky traffic behavior.[54-57]

Furthermore, according to our results, age was a significant variable related to pedestrian traffic behavior. Hence, younger pedestrians were more likely to be in the higher-risk cluster of traffic behavior. Consistent with our findings, some studies[58-60] have shown that young pedestrians are more distracted than older people and

show aggressive behavior. This behavior may be due to the risk-taker nature of this age group or the use of cell phones.

Moreover, based on the findings of our study, education was associated with the traffic behavior of pedestrians; that is, higher-educated pedestrians had safer traffic behavior. In this regard, we can declare that the increase in the level of education makes their behaviors and decisions more reasonable, especially in adherence to traffic rules and aggressive behaviors, and this result is consistent with other studies.[61]

In our data, married pedestrians had safer behavior. It can be influenced by age-related changes, or having a family may make people more responsible and cautious. This result is qualitatively consistent with similar studies with increased risks of driver injury among never-married people.[62]

Results of previous investigations regarding the role of marriage in pedestrians' traffic behavior are consistent with the current study. In line with our findings, Ghahramani et al revealed that married people are better than single ones in terms of traffic behavior.[63]

Regarding the kind of transportation, our results indicated that transportation with personal car decreases the odds of being in the higher-risk traffic behavior cluster. Although we did not find any relevant studies, due to the experience of driving and having a better perception of crashes, the people with individual cars avoid risky behavior. The use of the self-reporting method for data collection was the major limitation of this study because it may lead to bias in reporting traffic behavior.

Conclusion

We identified traffic behavior patterns of Urmia pedestrians consisting of lower-risk and higher-risk behaviors with FCM. Understanding which group of pedestrians have more unsafe behaviors and what causes them may help planners and policymakers think of better training solutions for them. The current study showed that using statistical methods, including clustering, can provide us with more details in addition to statistical descriptions. The findings from this study would help promote safety measures and training pedestrians.

Highlights

- It is crucial to identify the pattern of traffic behaviors of pedestrians to enhance management efficiency.
- We evaluated the pedestrians' traffic behavior pattern with the fuzzy c-means algorithm.
- Two clusters, consisting of lower-risk and higher-risk behaviors, were revealed. The majority of pedestrians (64.33%) were in the lower-risk cluster.
- Age, gender, marriage, education, and kind of transportation were associated with traffic behaviors of pedestrians.
- These findings can help promote safety measures and train pedestrians.

References

1. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380(9859):2095-128. doi: 10.1016/s0140-6736(12)61728-0.
2. World Health Organization (WHO). Injuries and Violence: The Facts. Geneva: WHO; 2010.
3. Pfortmueller CA, Marti M, Kunz M, Lindner G, Exadaktylos AK. Injury severity and mortality of adult zebra crosswalk and non-zebra crosswalk road crossing accidents: a cross-sectional analysis. PLoS One. 2014;9(3):e90835. doi: 10.1371/journal.pone.0090835.
4. Hatamabadi H, Vafaee R, Hadadi M, Abdalvand A, Esnaashari H, Soori H. Epidemiologic study of road traffic injuries by road user type characteristics and road environment in Iran: a community-based approach. Traffic Inj Prev. 2012;13(1):61-4. doi: 10.1080/15389588.2011.623201.
5. Karbakhsh M, Zandi NS, Rouzrokh M, Zarei MR. Injury epidemiology in Kermanshah: the national trauma project in Islamic Republic of Iran. East Mediterr Health J. 2009;15(1):57-64.
6. Fan C, Chen M, Wang X, Wang J, Huang B. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Front Energy Res. 2021;9:652801. doi: 10.3389/fenrg.2021.652801.
7. Mannering FL, Bhat CR. Analytic methods in accident research: methodological frontier and future directions. Anal Methods Accid Res. 2014;1:1-22. doi: 10.1016/j.amar.2013.09.001.
8. Shaheed MS, Gkritza K. A latent class analysis of single-vehicle

motorcycle crash severity outcomes. Anal Methods Accid Res. 2014;2:30-8. doi: 10.1016/j.amar.2014.03.002.

9. Haleem K, Alluri P, Gan A. Analyzing pedestrian crash injury severity at signalized and non-signalized locations. Accid Anal Prev. 2015;81:14-23. doi: 10.1016/j.aap.2015.04.025.

10. Valent F, Schiava F, Savonitto C, Gallo T, Brusaferro S, Barbone F. Risk factors for fatal road traffic accidents in Udine, Italy. Accid Anal Prev. 2002;34(1):71-84. doi: 10.1016/s0001-4575(00)00104-4.

11. Haleem K, Gan A. Contributing factors of crash injury severity at public highway-railroad grade crossings in the U.S. J Safety Res. 2015;53:23-9. doi: 10.1016/j.jsr.2015.03.005.

12. Lee RCT. Clustering analysis and its applications. In: Tou JT, ed. Advances in Information Systems Science. Vol 8. Boston, MA: Springer; 1981. p. 169-292. doi: 10.1007/978-1-4613-9883-7_4.

13. Friedman JH. Data mining and statistics: what's the connection? Comput Sci Stat. 1998;29(1):3-9.

14. Depaire B, Wets G, Vanhoof K. Traffic accident segmentation by means of latent class clustering. Accid Anal Prev. 2008;40(4):1257-66. doi: 10.1016/j.aap.2008.01.007.

15. Heil J, Häring V, Marschner B, Stumpe B. Advantages of fuzzy k-means over k-means clustering in the classification of diffuse reflectance soil spectra: a case study with West African soils. Geoderma. 2019;337:11-21. doi: 10.1016/j.geoderma.2018.09.004.

16. Bakhtari Aghdam F, Sadeghi-Bazargan H, Sarbakhsh P, Pashaie T, Ponnet K, Nicknejad M. Pedestrians in Iran: determinants of unsafe traffic behaviors of pedestrians. Version 2. Res Sq [Preprint]. doi: https://doi.org/10.21203/rs.3.rs-66747/v2.

17. Sadeghi-Bazargan H, Haghighi M, Heydari ST, Soori H, Rezapur Shahkolai F, Motevalian SA, et al. Developing and validating a measurement tool to self-report pedestrian safety-related behavior: the Pedestrian Behavior Questionnaire (PBQ). Bull Emerg Trauma. 2020;8(4):229-35. doi: 10.30476/beat.2020.86488.

18. Ferraro MB, Giordani P, Serafini A. fclust: an R package for fuzzy clustering. R J. 2019:1-18.

19. Pakhira MK, Bandyopadhyay S, Maulik U. Validity index for crisp and fuzzy clusters. Pattern Recognit. 2004;37(3):487-501. doi: 10.1016/j.patcog.2003.06.005.

20. Bezdek JC. Cluster validity with fuzzy sets. Journal of Cybernetics. 1973;3(3):58-73. doi: 10.1080/01969727308546047.

21. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53-65. doi: 10.1016/0377-0427(87)90125-7.

22. Hair JF Jr, Anderson RE, Tatham RL, Black WC. Multivariate Data Analysis. New Jersey: Prentice Hall; 1998.

23. Vermunt JK, Magidson J. Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL, eds. Applied Latent Class Analysis. Cambridge University Press; 2002.

24. Xu R, Wunsch D, 2nd. Survey of clustering algorithms. IEEE Trans Neural Netw. 2005;16(3):645-78. doi: 10.1109/tnn.2005.845141.

25. Bock HH. Automatische klassifikation. In: Walter E, ed. Statistische Methoden II: Mehrvariable Methoden und Datenverarbeitung. Berlin: Springer; 1970. p. 36-80. doi: 10.1007/978-3-642-88253-1_10.

26. Xu D, Tian Y. A comprehensive survey of clustering algorithms. Ann Data Sci. 2015;2(2):165-93. doi: 10.1007/s40745-015-0040-1.

27. Rasyid LA, Andayani S. Review on clustering algorithms based on data type: towards the method for data combined of numeric-fuzzy linguistics. J Phys Conf Ser. 2018;1097(1):012082. doi: 10.1088/1742-6596/1097/1/012082.

28. Nerurkar P, Shirke A, Chandane M, Bhirud S. Empirical analysis of data clustering algorithms. Procedia Comput Sci. 2018;125:770-9. doi: 10.1016/j.procs.2017.12.099.

29. Döring C, Lesot M-J, Kruse R. Data analysis with fuzzy clustering methods. Comput Stat Data Anal. 2006;51(1):192-214. doi: 10.1016/j.csda.2006.04.030.

30. Sajana T, Sheela Rani CM, Narayana KV. A survey on clustering techniques for big data mining. Indian J Sci Technol. 2016;9(3):1-12. doi: 10.17485/ijst/2016/v9i3/75971.

31. Jiawei H, Micheline K, Jian P. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann Publishers; 2012.

32. Bora DJ, Gupta AK. A comparative study between fuzzy clustering algorithm and hard clustering algorithm. Int J Comput Trends Technol. 2014;10(2):108-13. doi: 10.14445/22312803/ijctt-v10p119.

33. Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. Springer Science & Business Media; 2013.

34. Gunderson RW. Application of fuzzy ISODATA algorithms to star tracker pointing systems. IFAC Proceedings Volumes. 1978;11(1):1319-23. doi: 10.1016/s1474-6670(17)66090-7.

35. Windham MP. Cluster validity for fuzzy clustering algorithms. Fuzzy Sets and Systems. 1981;5(2):177-85. doi: 10.1016/0165-0114(81)90015-4.

36. Windham MP. Cluster validity for the fuzzy c-means clustering algorithrm. IEEE Trans Pattern Anal Mach Intell. 1982;4(4):357-63. doi: 10.1109/tpami.1982.4767266.

37. Libert G, Roubens M. New experimental results in cluster validity of fuzzy clustering algorithms. In: New Trends in Data Analysis and Applications. Amsterdam: North-Holland Publishing; 1983. p. 205-18.

38. Dave RN. Validating fuzzy partitions obtained through c-shells clustering. Pattern Recognit Lett. 1996;17(6):613-23. doi: 10.1016/0167-8655(96)00026-8.

39. Tavakoli Kashani A, Shariat Mohaymany A. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Saf Sci. 2011;49(10):1314-20. doi: 10.1016/j.ssci.2011.04.019.

40. Yasmin S, Eluru N, Bhat CR, Tay R. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. Anal Methods Accid Res. 2014;1:23-38. doi: 10.1016/j.amar.2013.10.002.

41. Sun M, Sun X, Shan D. Pedestrian crash analysis with latent class clustering method. Accid Anal Prev. 2019;124:50-7. doi: 10.1016/j.aap.2018.12.016.

42. Mohamed MG, Saunier N, Miranda-Moreno LF, Ukkusuri SV. A clustering regression approach: a comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Saf Sci. 2013;54:27-37. doi: 10.1016/j.ssci.2012.11.001.

43. Liu Y, Zhi W, Wang S, Wen X, Li H, Xu W. Application of ACA-based fuzzy c-means clustering to division of traffic zones. In: 18th COTA International Conference of Transportation Professionals. Beijing, China: American Society of Civil Engineers; 2018. doi: 10.1061/9780784481523.247.

44. Zhu G, Chen J, Zhang P. Fuzzy c-means clustering identification method of urban road traffic state. In: 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). Zhangjiajie: IEEE; 2015. p. 302-7. doi: 10.1109/fskd.2015.7381958.

45. Javadi S, Rameez M, Dahl M, Pettersson MI. Vehicle classification based on multiple fuzzy c-means clustering using dimensions and speed features. Procedia Comput Sci. 2018;126:1344-50. doi: 10.1016/j.procs.2018.08.085.

46. Liu D, Lung CH. P2P traffic identification and optimization using fuzzy c-means clustering. In: 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011). Taipei, Taiwan: IEEE; 2011. doi: 10.1109/fuzzy.2011.6007613.

47. Hamzehei A, Chung E, Miska M. Traffic safety risks trends and patterns analysis on motorways. In: Kononov J, ed. Transportation Research Board (TRB) 93rd Annual Meeting Compendium of Papers. Washington, DC: Transportation

Research Board (TRB); 2014.

48. Kim K, Yamashita EY. Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in Honolulu, Hawaii. J Adv Transp. 2007;41(1):69-89. doi: 10.1002/atr.5670410106.

49. Lee C, Abdel-Aty M. Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida. Accid Anal Prev. 2005;37(4):775-86. doi: 10.1016/j.aap.2005.03.019.

50. Sze NN, Wong SC. Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accid Anal Prev. 2007;39(6):1267-78. doi: 10.1016/j.aap.2007.03.017.

51. Ivan JN, Garder PE, Zajac SS. Finding Strategies to Improve Pedestrian Safety in Rural Areas. 2001. Available from: https://rosap.ntl.bts.gov/view/dot/14677.

52. Islam S, Jones SL. Pedestrian at-fault crashes on rural and urban roadways in Alabama. Accid Anal Prev. 2014;72:267-76. doi: 10.1016/j.aap.2014.07.003.

53. Rothman L, Howard AW, Camden A, Macarthur C. Pedestrian crossing location influences injury severity in urban areas. Inj Prev. 2012;18(6):365-70. doi: 10.1136/injuryprev-2011-040246.

54. Peden MM, Knottenbelt JD, van der Spuy J, Oodit R, Scholtz HJ, Stokol JM. Injured pedestrians in Cape Town--the role of alcohol. S Afr Med J. 1996;86(9):1103-5.

55. Kumar A, Paul M, Ghosh I. Analysis of pedestrian conflict with right-turning vehicles at signalized intersections in India. J Transp Eng A Syst. 2019;145(6):04019018. doi: 10.1061/jtepbs.0000239.

56. Clifton KJ, Livi AD. Gender differences in walking behavior, attitudes about walking, and perceptions of the environment in three Maryland communities. In: Proceedings from the Conference on Research on Women's Issues in Transportation; 2005 Mar 31-Apr 2; Chicago, IL. p. 79-88.

57. Jamali-Dolatabad M, Sarbakhsh P, Sadeghi-Bazargani H. Hidden patterns among the fatally injured pedestrians in an Iranian population: application of categorical principal component analysis (CATPCA). BMC Public Health. 2021;21(1):1149. doi: 10.1186/s12889-021-11212-x.

58. Wells HL, McClure LA, Porter BE, Schwebel DC. Distracted pedestrian behavior on two urban college campuses. J Community Health. 2018;43(1):96-102. doi: 10.1007/s10900-017-0392-x.

59. Granié M-A, Pannetier M, Guého L. Developing a self-reporting method to measure pedestrian behaviors at all ages. Accid Anal Prev. 2013;50:830-9. doi: 10.1016/j.aap.2012.07.009.

60. Deb S, Strawderman L, DuBien J, Smith B, Carruth DW, Garrison TM. Evaluating pedestrian behavior at crosswalks: validation of a pedestrian behavior questionnaire for the U.S. population. Accid Anal Prev. 2017;106:191-201. doi: 10.1016/j.aap.2017.05.020.

61. Zheng T, Qu W, Ge Y, Sun X, Zhang K. The joint effect of personality traits and perceived stress on pedestrian behavior in a Chinese sample. PLoS One. 2017;12(11):e0188153. doi: 10.1371/journal.pone.0188153.

62. Whitlock G, Norton R, Clark T, Jackson R, MacMahon S. Motor vehicle driver injury and marital status: a cohort study with prospective and retrospective driver injuries. Inj Prev. 2004;10(1):33-6. doi: 10.1136/ip.2003.003020.

63. Ghahramani M, Fathi Shayan M, Nazemi S, Farah Bakhsh M, Dahim M, Sadeghi- Bazargani H, et al. Study of pedestrian behavior and its effective factors in Sahand new city. J Inj Violence Res. 2019;11(4 Suppl 2). doi: 10.5249/jivr.v11i2.1297.