Fuzzy C-Means clustering, GPS, Level of Service,
urban streets, cluster validation measures

Prasanta Kumar BHUYAN*, **K.V. Krishna RAO**
Department of Civil Engineering, IIT Bombay, India 400076
**Corresponding author*. E-mail: pkbtrans@gmail.com

FCM CLUSTERING USING GPS DATA FOR DEFINING LEVEL OF SERVICE CRITERIA OF URBAN STREETS IN INDIAN CONTEXT

Summary. At present speed ranges for Levels of Service (LOS) categories are not well defined for highly heterogeneous traffic flow on urban streets in Indian context. In this regard a study was carried out in the city of Mumbai in India. The objective of this research work is to define free-flow speed ranges of urban street classes and speed ranges of LOS categories. In this regard speed data were collected using GPS palmtop and Fuzzy C-Means (FCM) clustering is used to define the speed ranges. It is found from this study that speed-ranges for LOS categories in Indian context are lower than that mentioned in Highway Capacity Manual.

КЛАСТЕРИЗАЦИЯ В СООТВЕТСТВИИ С АЛГОРИТМОМ НЕЧЕТКИХ С-СРЕДНИХ С ИСПОЛЬЗОВАНИЕМ ДАННЫХ GPS ДЛЯ ОПРЕДЕЛЕНИЯ УРОВНЯ СЕРВИСА ГОРОДСКИХ УЛИЦ В УСЛОВИЯХ ИНДИИ

Аннотация. В настоящее время диапазоны скорости для уровня сервиса различных категорий не определены в связи с разнородностью транспортных потоков на городских улицах в условиях Индии. Этим было обусловлено проведение исследований в городе Мумбаи в Индии. Целью данной исследовательской работы является определение уровней скорости свободного потока на городских улицах и диапазонов скоростей для различных категорий уровня сервиса. В этой связи для определения диапазона скоростей были собраны данные о скорости с использованием GPS карманного персонального компьютера (палмтопа) и алгоритма нечетких с-средних. Из этого исследования определено, что диапазоны скоростей для различных категорий уровня сервиса в условиях Индии ниже тех, которые упомянуты американском руководстве по пропускной способности дорог.

1. INTRODUCTION

Level of service in the HCM (2000) is defined as ‘‘a quality measure describing operational conditions within a traffic stream, generally in terms of such service measures as speed and travel time, freedom to maneuver, traffic interruptions, comfort and convenience’’ [1]. The HCM designates six levels of service for each type of facility, from “A” to “F”, with LOS “A” representing the best operating conditions and LOS “F” the worst. Urban Street LOS is based on average through-vehicle travel speed for the segment or for the entire street under consideration. The use of moving observer method is the most commonly used technique for the collection of travel time data. However accuracy with this technique varies from technician to technician. Recent research has demonstrated the feasibility of using GPS receiver in recording location as longitude-latitude, travel time and travel speed. GPS receivers can record location and speed automatically at regular sampling periods. As a result, large amounts of reliable travel time and speed data can be collected and processed. Consequently, only one technician is required in the vehicle to operate the equipment. With GPS data the speed ranges of levels of service categories of urban street classes can be defined at higher accuracy.

Number of methodologies has been developed for level of service analysis. In its application of fuzzy set theory to analyze highway capacity and level of service it has been shown the limitations of the procedure traditionally followed to determine the level of service [2]. In this study a model to evaluate airport passenger services using fuzzy set theory techniques was developed. In their words the authors have stated, "The literature on transportation level of service evaluation indicates a strong impetus to move away from a strictly capacity/volume or time/space based measure to one that directly incorporates the perception of passengers". In a recent study, the researchers have examined the uncertainty associated with the measuring and mapping of existing six LOS categories. Hence six frameworks were developed to address the uncertainty lies within six levels of service categories [3]. In a similar kind of study, a methodology was developed using fuzzy set technique to define level of service of urban roads taking into account users' perceptions in Indian environment [4]. Further improvement to classification problems in its application three clustering techniques such as *K*-means, Fuzzy C-means and Self Organizing Map (SOM) were used for the development of a real-time inductive-signature-based level of service criteria for signalized intersection surveillance system [5]. Similar to the present study threshold speed was used to asses LOS for heavy traffic under platooing condition [6]. In this study the definition of threshold speed used "the minimum speed users consider acceptable in traveling on a uniform road section under heavy flows and platooning traffic" by the authors. The method used two measures of effectiveness, one reflecting percent time spent following (PTSF) and the second reflecting speed. However it has been suggested a threshold speed would be used to decide which MOE governs LOS in each period: if average travel speed (ATS) is higher than the threshold value, only PTSF would be examined, implying user consider the speed reasonable. If ATS is below the threshold speed, platooning would be behind speed in importance in the view of drivers.

After the collection of large set of travel speed data, the issue is how to determine the threshold values for partitioning different LOS categories. From literature it was found that Fuzzy C-means (FCM), a method derived from fuzzy logic is suitable for solving multi-class clustering problems. FCM clustering is used on speed data to define speed ranges of LOS categories of urban street classes. The method of FCM clusters analysis, various cluster validation parameters and their use in finding optimum number of clusters has been demonstrated in this study. The data set used in this study was obtained from 10 to12 travel runs taken on five major urban corridors in the city of Mumbai, India during March and April of the year 2005. The total length of these corridors is about 140 km. These corridors, on the whole, were divided into 100 street segments. Comprehensive data sets of free-flow speed, travel speeds during both peak and off- peak hours, inventory details and classified traffic volume data were used. Also, the same sets of data were collected from Kolkata, a city in the eastern part of India to check the validity of this method in application to Indian cities in general. This extensive data which covers the desired variation that is expected in the urban cities is an indication towards the data sufficiency for applying FCM clustering method in defining level of service criteria in Indian context.

2. FCM CLUSTERING

The connection between fuzzy logic and fuzzy cluster analysis is usually only through the application of membership coefficient, and not the more comprehensive theory. FCM clustering algorithm introduced by Bezdek is adopted in the present study [7], which is considered one of most popular and accurate algorithms in cluster analysis/pattern recognition. [8] Based on concepts, centers are as similar as possible to each other within a cluster and as different as possible from elements in other clusters.

A $N \times c$ matix U= [μ_{ik}] represents the fuzzy partitions, its conditions are given by:

$$\mu_{ik} \in [0, 1], 1 \le i \le N, \quad 1 \le k \le c, \tag{1a}$$

$$\sum_{k=1}^{c} \mu_{ik} = 1, \quad 1 \le i \le N, \tag{1b}$$

$$0 < \sum_{i=1}^{N} \mu_{ik} < N, \quad 1 \le k \le c. \tag{1c}$$

The FCM clustering algorithm is based on the minimization of an objective function called *C-means functional*. It is defined by Dunn as:

$$J(X;U,V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m \left\| X_k - V_i \right\|_A^2 \tag{2}$$

Where:

$$V = [V_1, V_2, V_3, \ldots\ldots\ldots\ldots, V_c], V_i \in R^n \tag{3}$$

is a vector of cluster centers, which have to be determined, and

$$D_{ikA}^2 = \left\| X_k - V_i \right\|_A^2 = (X_k - V_i)^T A (X_k - V_i) \tag{4}$$

is a squared inner-product distance norm.

Where: X is the data set, U is the partition matrix; V is the vector of cluster centers; V_i is the mean for those data points over cluster i; m is the weight exponent which determines the fuzziness of the clusters (default value is 2); n is the number of observations; $D^2{}_{ik}$ is the distance matrix between data points and the cluster centers; A_i is a set of data points in the i-th cluster; c is the number of clusters; N is the number of data points;

The stationary points of the objective function (2) can be found by adjoining the constraint (1b) to J by means of Lagrange multipliers:

$$\bar{J}(X;U,V,\lambda) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m D^2{}_{ikA} + \sum_{k=1}^{N} \lambda_k \left(\sum_{i=1}^{c} \mu_{ik} - 1 \right) \tag{5}$$

and by setting the gradients of ($\bar{J}$) with respect to U, V and λ to zero. If $D_{ikA}^2 > 0, \forall_{i,} k$ and m>1, then *(U, V)* may minimize (2) only if

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(D_{ikA} / D_{jkA} \right)^{2/m-1}}, \quad 1 \le i \le c, \quad 1 \le k \le N \tag{6}$$

and

$$V_i = \frac{\sum_{k=1}^{N} \mu_{ik}^m X_k}{\sum_{k=1}^{N} \mu^m{}_{i,k}}, \; 1 \le i \le c, \tag{7}$$

This solution also satisfies the remaining constraints (1a) and (1c). It is to be noted that equation (7) gives V_i as the weighted mean of the data items that belong to a cluster, where the weights are the membership degrees. That is why the algorithm is called C-means. It can be seen that the FCM algorithm is a simple iteration through (6) and (7).

2.1. Cluster Validation Measures

Cluster validity refers to the problem of whether a given partition fits to the data at all [9]. Different scalar validity measures have been proposed in the literature, none of them is perfect by oneself, and therefore several indices are used in this study, such as: Partition Coefficient (PC), Classification Entropy (CE), Partition Index (PI), Separation Index (SI), Xie and Beni's Index (XB), Dunn's Index (DI) and Silhouettes (S). The validation parameter Silhouette is discussed below.

Silhouettes

A silhouette value S is expressed for each object as follows.

$$S = \frac{b-a}{\max(a,b)} \tag{8}$$

Here, a particular object i is in cluster A and a is equal to the average dissimilarity of i to all other objects in A. For every other cluster not equal to A, cluster B has the smallest average dissimilarity between its objects and i which is equal to b. The cluster B is the nearest neighbor of object i.

3. STUDY CORRIDORS AND DATA COLLECTION

3.1. Study Corridors

Five important urban road corridors of the city of Mumbai of Maharastra State, India are selected for the present study. Grater Mumbai is an Island city with a linear pattern of transport network having predominant North-South commuter movements. Passengers move towards south for work trip in the morning hours and return towards north in the evening hours. Hence four north-south corridors and one east-west corridor were chosen. Major roads like Eastern express highway extending up to south (Corridor-1), LBS Road extending up to south via Ambedkar road (Corridor-2), Western express highway extending up to marine drive (Corridor-3), SV road extending up to south via Veer Savarkar road (Corridor-4) and Versova- Andheri- Ghatkopar- Vashi (VAGV) (Corridor-5) are included. These five corridors are overlapped on the GIS base map of Greater Mumbai are shown in fig. 1.

3.2. Data Collection

The probe vehicles used in this research work were mid-sized vehicles (car). These vehicles were fitted with Trimble Geo-XT GPS receiver. The GPS data provides both spatial and time based data from which various traffic parameters were derived, including travel time and travel speeds. In order to get unbiased data sets three mid-sized vehicles were used and help of different drivers on different days of the survey work was taken. Basically three types of data sets were collected. The first type is roadway inventory details. During the collection of inventory details proper segmentation technique was applied, which is the directional stretch of road section immediately after signalized intersections to the location point immediately after the next signal. The second type of survey conducted was to find the free flow speed. The third type of data collected was congested travel speed. Congested travel speed survey was conducted during peak and off-peak hours on both directions of travel on all these corridors. In order to show the applicability of this study in other cities of India a similar survey was carried out in Kolkatta City. Two corridors having varying geometric and surrounding environmental characteristics were taken into considerations i.e. one corridor was Airport to Joka and the other corridor was Airport to Ulberia. These two corridors are approximately 80 kilometer length; comprised of 50 street segments. The interesting fact on selecting these two cities for this study is that traffic composition and road geometric characteristics along with functionality brings the true variation that was required for this purpose.

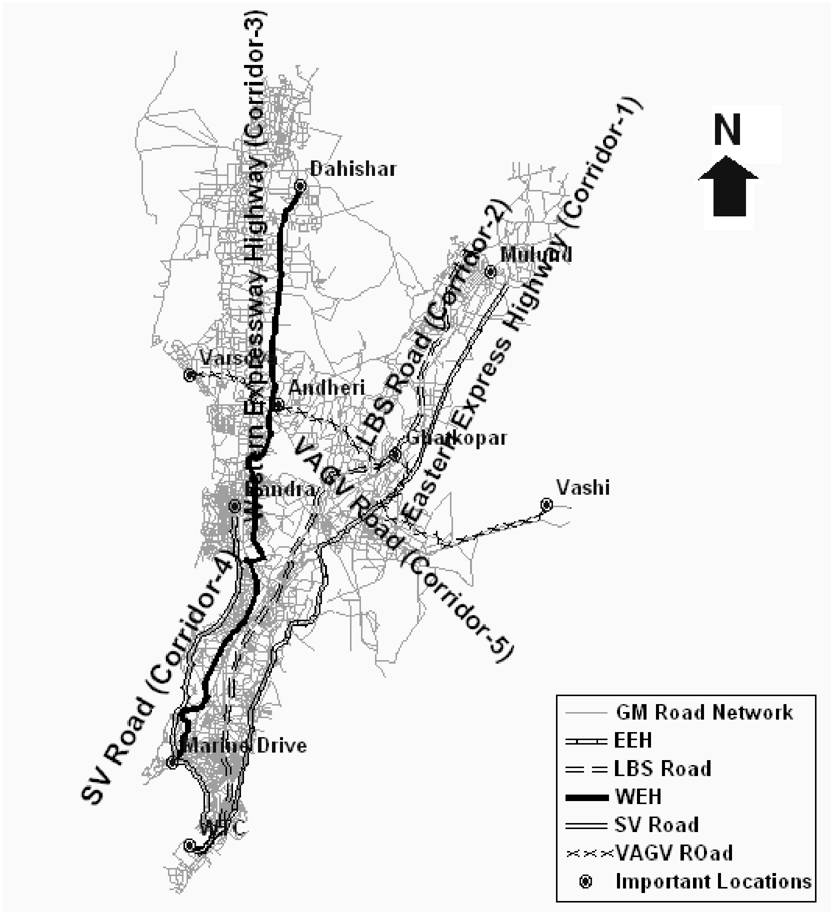
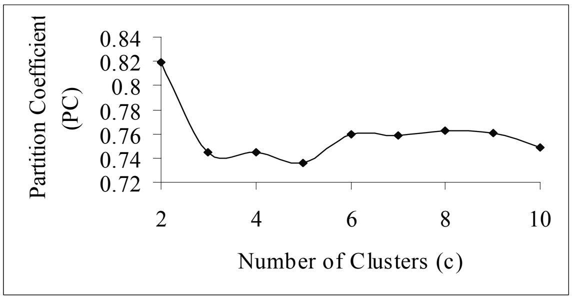


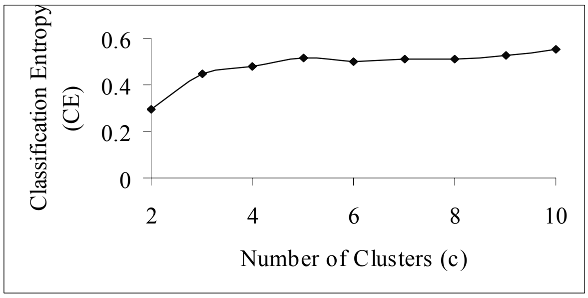
Fig. 1. Map showing selected corridors of greater Mumbai
Рис. 1. Карта, показывающая главные транспортные коридоры Мумбаи

4. RESULTS AND ANALYSIS

First, free flow speed data were used in FCM, and then input data (free flow speed) and output data (cluster centers) found from the analysis are used in computing validation parameters. The values of the validation measures obtained for 2 to 10 number of clusters are plotted in fig. 2. All the six validation parameters are interpreted to obtain the optimum number of clusters in deciding the classification of street segments into number of street classes.

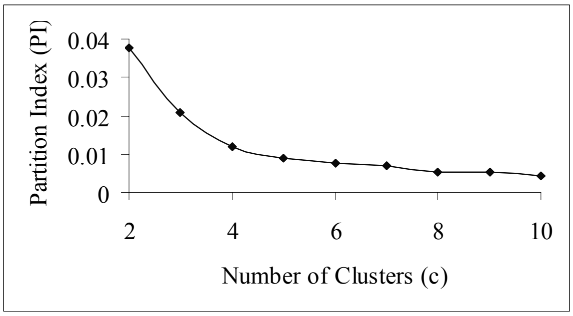
It is considered that partitions with lesser clusters are better, when the differences between the values of validation parameters are minor. From fig. 2 it is found that CE, PI and SI decrease beyond 4 clusters at a very slow rate. The rate of decrease in value of XB beyond 4 clusters is minimal and hence approaches its local minimum. Considering PC, CE, PI, SI, and XB and comparing these indices for different number of clusters, the optimal number of clusters was found to be 4; although PC and DI do not show a particular trend in their value. With the confirmations from all these validation measures, it has been decided to categorize the urban streets into four classes using free flow speeds on FCM clustering. Thus free-flow speed ranges of urban street classes are defined. It is observed from the collected data set that when a street segment falls under particular urban street class is agreed with the geometric and surrounding environmental condition of the road segments as well. It has been found that there is very good correlation between free flow speed and geometric and environmental characteristics of streets under considerations. Direction-wise average travel speeds calculated on street segments were used on FCM clustering to define speed ranges for level of service categories. Speed data falling under the six levels of service categories of urban street classes are shown in fig. 3.

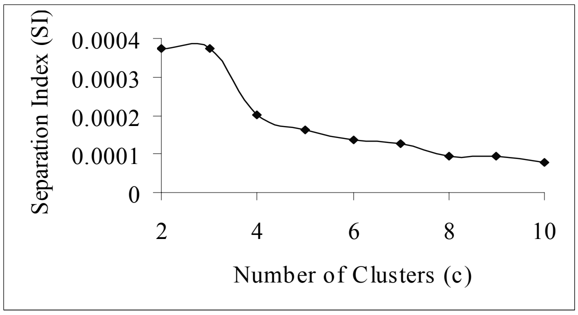




A: Partition coefficient vs. number of clusters

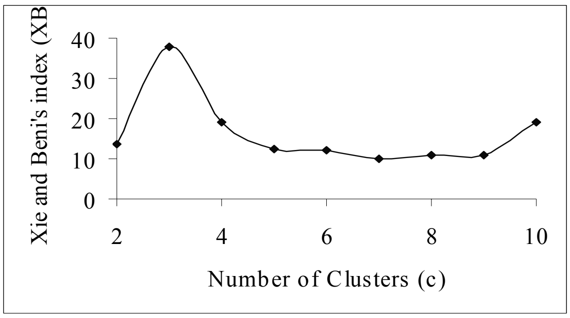
B: Classification entropy vs. number of clusters

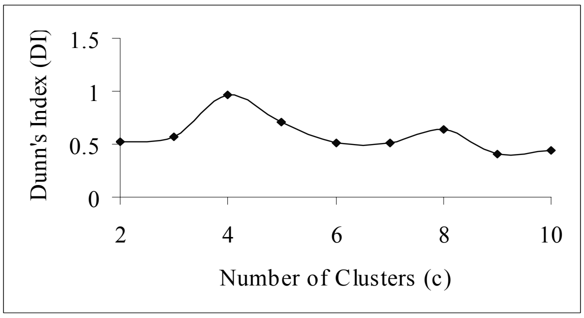




C: Partition index vs. number of clusters

D: Separation index vs. number of clusters





E: Xie and Beni's index vs. number of clusters

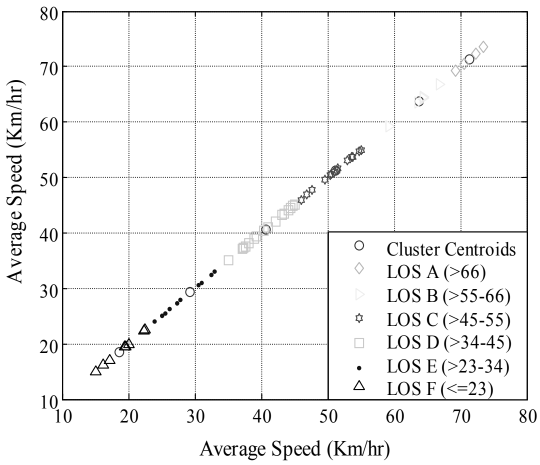
F: Dunn's index vs. number of clusters

Fig. 2. Validation measures for optimal number of clusters using FCM clustering
Рис. 2. Проверка измерения оптимального количества кластеров при использовании алгоритма нечетких с-средних

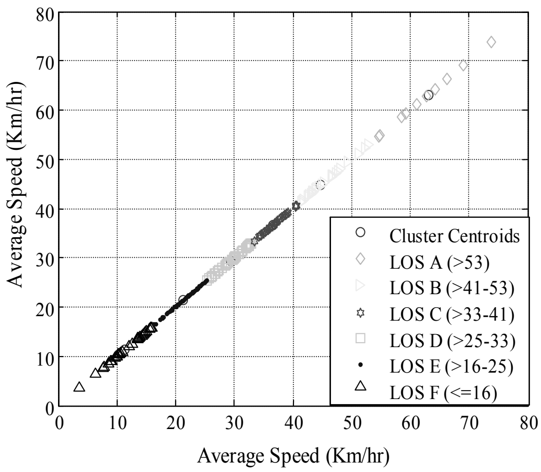
In fig. 3 the speed values are shown by different symbols depending on to which LOS category they belong. The legend in fig. 3 gives the speed ranges for the six LOS categories obtained by using FCM clustering. The speed ranges for LOS categories found using FCM clustering are also shown in table 1.

From table 1 it can be stated that free-flow speed ranges of urban street classes and speed ranges of LOS categories based on FCM clustering in Indian context are significantly different from the suggested values applicable to western countries with homogeneous traffic flow condition. Significant percentage of slow moving vehicles on Indian urban roads retards the average travel speed hence result in lower speed rages for urban streets and LOS categories. Fig. 4 shows Silhouettes plot for urban street classes using free flow speeds.

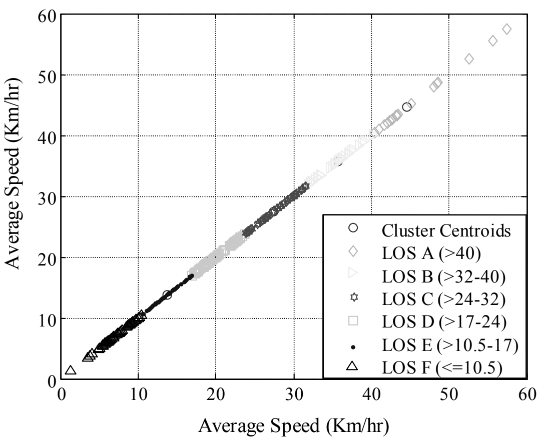
From this figure it is found that thickness of silhouettes of Urban Street class II and III are comparatively high which means large numbers of street segments are falling under urban street class II and III. Also it is found that Silhouette width of data points under urban street classes I to IV lies between 0.7 and 1.0. This indicates that free-flow speed data points are very strongly bonded within each urban street class. In other words urban streets are well classified into four numbers of street classes. Also it is found that speed data points are well bonded within each level of service categories using silhouette validation parameter. Speed ranges for level of service categories (A-F) expressed in percentage of free-flow speeds were found to be 85, 75, 60, 45, 33 and 20-30 respectively in the present study. Whereas, in HCM (2000) it has been mentioned these values are 90, 70, 50, 40, 33 and 25-33 percent respectively.



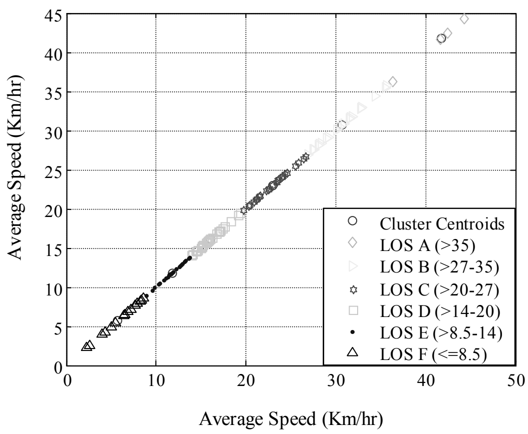
A: LOS of urban street class I



B: LOS of urban street class II



C: LOS of urban street class III



D: LOS of urban street class IV

Fig. 3. Level of service of urban street classes (I-IV) using FCM clustering on average travel speeds
Рис. 3. Уровень сервиса городских улиц классов (I-IV) для кластеризации с использованием алгоритма нечетких c-средних для средней скорости поездки

In order to check the application of this level of service criteria; data collected from Kolkata city were tested. Free flow speed and average travel speed during both peak and off-peak hours on each of segments on both corridors were calculated. The street segments were classified into four classes based on free-flow speed, geometric and surrounding environmental characteristics. Also, levels of service provided by the street segments during peak and off peak hours were estimated using Table 1 shown above. The percentage of travel runs under different levels of service categories found for urban street classes in Kolkata city during the survey period are shown in Table 4.

Table 1

Urban street speed ranges of LOS categories using FCM clustering

Urban street class	I	II	III	IV
Range of free-flow speed (FFS)	90 to 70 km/h	70 to 55 km/h	55 to 45 km/h	45 to 25 km/h
Typical FFS	75km/h	60km/h	50km/h	40 km/h
LOS	Average travel speed (km/h)			
A	>66	>53	>40	>35
B	>55-66	>41-53	>32-40	>27-35
C	>45-55	>33-41	>24-32	>20-27
D	>34-45	>25-33	>17-24	>14-20
E	>23-34	>16-25	>10.5-17	>8.5-14
F	≤ 23	≤ 16	≤ 10.5	≤ 8.5

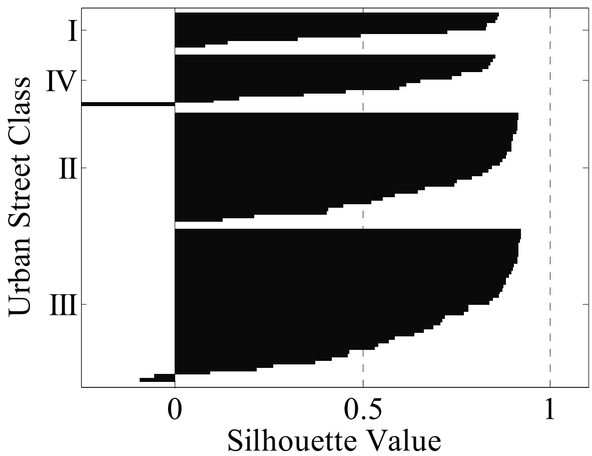


Fig. 4. Silhouette plot for urban street classes using FCM clustering

Рис. 4. График силуэта кластеров для категорий городских улиц при кластеризации с использованием алгоритма нечетких c-средних

Table 2

Percentage of travel runs under levels of service categories of urban street classes in Kolkatta city

Level of Service	Urban Street Class			
	I	II	III	IV
A	38.10	21.43	6.25	11.29
B	19.05	21.43	25.00	22.58
C	19.05	7.14	15.63	25.81
D	14.29	21.43	25.00	17.74
E	9.52	25.00	15.63	11.29
F	9.52	3.57	12.50	11.29

From this table it has been observed that the probe vehicle traveled at better quality of service under urban street class I, whereas under other urban street classes the observed vehicle traveled at average quality of service during the observed period.

5. CONCLUSION

After thorough interpretation of the several plots of cluster validation parameters, it was concluded to classify urban streets into four classes (I-IV) in Indian context. Free-flow speed range for urban street class IV obtained in Indian context is significantly lower than that mentioned in HCM (2000). The presence of high percent of slow moving vehicles on Indian urban roads can be one of the major causes for this lower free-flow speed range. The speed ranges for LOS categories are proportionately lower than those suggested in HCM (2000). These lower values are due to highly heterogeneous traffic flow on urban road corridors with varying geometry in India. Also, speed expressed in percentage of free flow speed is higher, especially, at lower LOS categories is that the absolute value of FFS themselves were found to be at the lower side. Using the FCM clustering methods on "Silhouette" validation parameter it is inferred that, in Greater Mumbai region less number of road segments are of high speed design (street class-I) or highly congested (street class-IV). More number of road segments is of suburban (street class II) or intermediate (street class III) type. It can be suggested that Greater Mumbai region needs substantial geometric improvements to provide better quality of service to the road users.

References

1. *Highway Capacity Manual.* Transportation Research Board (TRB), Washington. DC, USA, 2000.
2. Ndoh N.N. Ashford N.J.: *Evaluation of transportation level of service using fuzzy sets.* 73rd Annual Meeting of TRB, Transportation Research Board, Washington, DC, 1994.
3. Kikuchi S., Chakroborty P.: *Framework to represent uncertainty when level of service is determined.* Transportation Research Record 1968, TRB, National Research Council, Washington, D.C., 2006, p. 53–62.
4. Pattnaik S.B., Kumar K.R.: *Level of service of urban roads based on users perception.* Civil Engineering Systems, Vol. 14, 1996, p. 87-110.
5. Cheol O., Stephen G.R.: *Real-time inductive-signature-based level of service for signalized intersections.* Transportation Research Record 1802, TRB, National Research Council, Washington, DC, 2002, p. 97–104.
6. Romana M.G., Perez I.: *Measures of effectiveness for level-of-service assessment of two-lane roads: An alternative proposal using a threshold speed.* Transportation Research Record 1988, TRB, National Research Council, Washington, DC, 2006, p. 56–62.
7. Bezdek J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithm.* Plenum, New York, 1981.
8. Jain A.K., Dubes R.C.: *Algorithms for Clustering Data.* Prentice Hall, Englewood Cliffs, 1988.
9. Bezdek J.C., Pal N.R.: *Some new indexes of cluster validity.* IEEE Transactions on System, Man, and Cybernetics-Part B: Cybernetics, Vol. 28, No.3, 1998, p. 301-315.

Received 28.10.2009; accepted in revised form 9.12.2010