European Transport \ Trasporti Europei (2017) Issue 66, Paper n° 1, ISSN 1825-3997

Clustering Technique: An Analytical Tool in Traffic Engineering to Evaluate the Performance of Two-Lane Highways

Amardeep Boora[1], Indrajit Ghosh[2], Satish Chandra[3]

[1]Ph.D. Candidate, Department of Civil Engineering, Indian Institute of Technology (IIT) Roorkee, Roorkee 247667, Uttarakhand, India
[2] Assistant Professor, Department of Civil Engineering, Indian Institute of Technology (IIT) Roorkee, Roorkee 247667, Uttarakhand, India. E-mail: indrafce@iitr.ac.in
[3] Director, Central Research Road Institute (CRRI) New Delhi, New Delhi 110025, Delhi, India. Email: satisfce@gmail.com

Abstract

The performance of a highway can be evaluatedin terms oflevel of service (LOS). Latest U.S Highway Capacity Manual (HCM 2010) proposed six LOS thresholds i.e. A to F for two-lane as well as for multilane highways. Besides this, several studies have been conducted across the world to evaluate the performance of different road type facilities. However, calibration of different LOS thresholds is a classification related problem i.e. it is required to classify the data sample into different LOS ranges. Consequently, thek-median algorithm is used forclustering analysis to calibrate different LOS ranges with the help of three distance metrics, namely, Euclidean, Square Euclidean and Cityblock distance metrics. In the present study, results obtained from the k-median algorithm using three different metrics are discussed. Later, two statistics namely, Cohen kappa and Silhouette plot are also used to identify the best distance metrics among these three metrics. Additionally, different LOS thresholds were calibrated using the relationship of two performance measures named as follower density and the number of followers as the proportion of capacity. In the present study, clustering analysis technique is identified as an effective analytical tool for the analysis ofperformance measures for differenttraffic facilities.

Keywords: Centroid; Clustering; Metrics; Level of Service

1. Introduction

Clustering is a technique used to prepare different groups of similar types of data sets which incorporates different classification algorithms. There are a number of algorithms to

(corresponding author) Email: amardeepboora9@gmail.com

do the clustering i.e. k-means, fuzzy clustering, hierarchy clustering and manyothers. A general problem faced by variousresearchers in different research fields across the world is to organize the observed data sets into a meaningful structure for developing the taxonomies.Earlier, Jain et al. (1999) reviewed different clustering techniques and identified the clustering technique as a useful tool for various scientific studies. According to this study, Cluster analysis can be defined as "a process of grouping the different data sets based on similarity."Clustering is very useful in defining pattern analysis, grouping of the similar data sets, pattern classification and many more. Generally, clustering technique groups the data sets such that the similar instances are consolidated in the same cluster, while dissimilar data samples will belong to different cluster groups. The main objective of the study is to identify the best distance metrics to calibrate different LOS thresholds for two-lane highways. For the same purpose two different techniques, namely, Cohen Kappa and Silhouette plotwere used in the present study. While Cohen Kappa measures inter-rater agreement for qualitative items, Silhouette is used for interpretation and validation of the internal consistency of data within a cluster.

2. Literature Review

Past studies related to traffic engineering as well as other disciplinesused clustering as an analytical tool. Peng and Li (2006) proposed a two-stage approach for clustering through the evaluationof therelationship between traditional distances and modified distances. From the results, extended data clustering was found more suitable than the traditional one. Saha et al. (2015) conducted a study to develop different LOS ranges based on data obtained from several two-lane highwaystudy sites. For doing clustering analysis, three distance metrics namely, Euclidean, Square Euclidean and Chebyshev distance metrics were used. Findings of the study revealed K-median clustering as a suitable algorithm with the Chebyshev distance and K-mean clustering with the Euclidean distance.Bhuyan and Mohapatra (2014) used an affinity propagation (AP) clustering technique to define different LOS ranges for urban streets under heterogeneous traffic condition. Different LOS ranges proposed in the study were found lower than U.S. HCM (2000) proposed LOS ranges. Ben-Hur and Guyon (2003) used principal component analysis (PCA) to identify the best cluster data by using a hierarchical clustering algorithm. On the basis of results, PCA technique was found appropriate to identify best clusters with respect to stability, modification, and coincidence. Singh et al. (2013) conducted a study to identify the best distance metrics using k-means algorithm. Euclidean distance metrics was identified as the best one for clustering analysis as the distortion in k-means using Euclidean distance metric was found less than that of k-means using Manhattan distance metric.Bhuyan and Rao (2010) used Fuzzy C-Mean (FCM) clustering technique to define LOS criteria using free-flow speed (FFS) for urban streets in the Indian context. FFS ranges for all classes of urban streets were found lower than LOS ranges proposed by U.S. HCM (2000) and the main reason could be due to the highly heterogeneous traffic condition and the presence of slow moving vehicles. Bora and Gupta (2014) examined the performance of k-means algorithm using different distance metrics. From the study, it was observed that City block distance performed in a better way in terms of less computation time while the Silhouette plot revealed that the correlation distance metrics interpret the cluster data very well in

comraison to other metrics. Jiang et al. (2003) conducted a study to examine the suitability of fuzzy clustering analysis technique for identification of road traffic state. Murugesan and Moorthy (1997) used fuzzy clustering to examine the level of public transport service (LOPTS). It was concluded that the fuzzy approach could be used to assess the LOS of public transport. Prassas et al. (1996) also found that clustering technique could be used as an effective tool in transportation engineering. Qiu et al. (2012) used clustering analysis method (CAM) to make a layout for urban transport hub (UTH). It was observed that CAM provided a theoretical foundation for UTH layout planning. In another study, Bradley et al. (1997) identified k-median as a superior algorithm over k-means algorithm.

3. Methodology of the Study

Recently, Boora and Ghosh (2016) conducted a study to identify the performance measures for two-lane intercity highways. In that study five study sites, namely, site 1 (NH-47), site 2 (NH-58), site 3 (NH-4) site 4 (SH-31) and site 5 (SH-59) were selected which are located on two-lane intercity highways.Follower density (FD) and the number of followers as a proportion of capacity (NFPC) were identified as the most suitable performance measures. The same data was used in the present study to do the cluster analysis. Cluster analysis technique is discussed in details in next section of the research paper.

3.1 Cluster analysis technique

It is a data mining function which can be used to examine the characteristics of each group. Clustering helps in theexamination of thestructure of a particular data set. Clustering analysis is used to categorize the data samples or object in such a manner that the objects belong to a particular group (i.e. due to thesimilarity in data sample) rather than another data samples of different group. In order to identify the similarities among different data samples of the various groups, a specific clustering algorithm is needed. Consequently, k-median clustering algorithm was adopted in the present study. Different distance metrics used in the clustering analysis are as follows:

Euclidean Distance

It is the distance which helps to simplify the distance between two points in the plane or space. Let us assume $X = (X_1, X_2, X_3, \ldots X_N)$ and $Y = (Y_1, Y_2, Y_3, \ldots Y_N)$ are two points in Euclidean *n*-space, then it can be computed mathematically by using Equation 1.

$$\text{Distance (X, Y)} = \sqrt{\sum_i (X_i - Y_i)^2} \quad (1)$$

Squared Euclidean Distance

In order to place progressively greater weight to the points that are further apart, thestandardeuclidean distance can be squared and calculated as Square Euclidean distance. It can be represented in a mathematical form as shown in Equation 2.

$$\text{Distance (X, Y)} = \left(\sqrt{\sum_i (X_i - Y_i)^2}\right)^2 \quad (2)$$

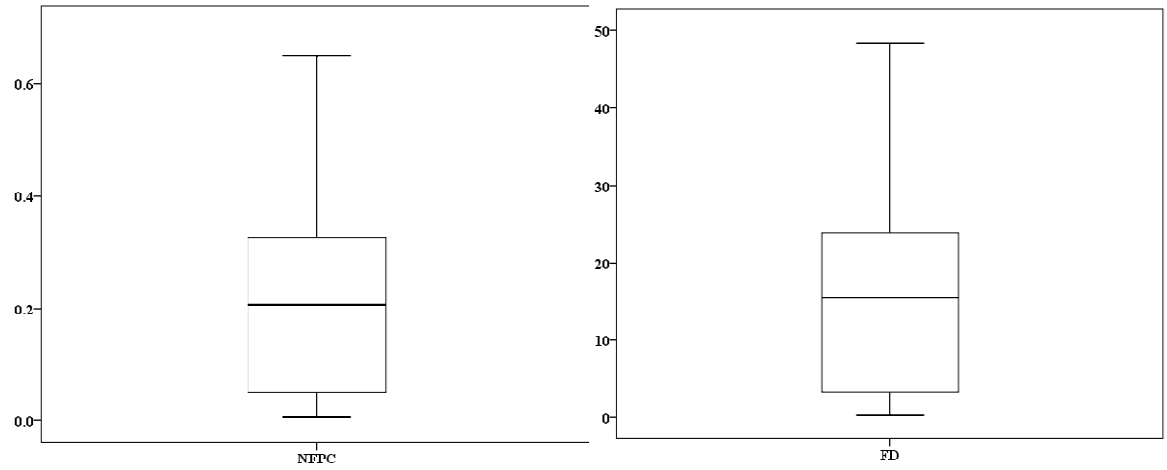
City-block Distance

It is the absolute average difference across dimensions, and mostly it exhibits the result similar to the standard Euclidean distance. It is to be noticed that the outliers are diminished (as they are not squared) and it is computed by Equation 3:

$$\text{Distance (X, Y)} = \sum_i abs\,(X_i - Y_i) \quad (3)$$

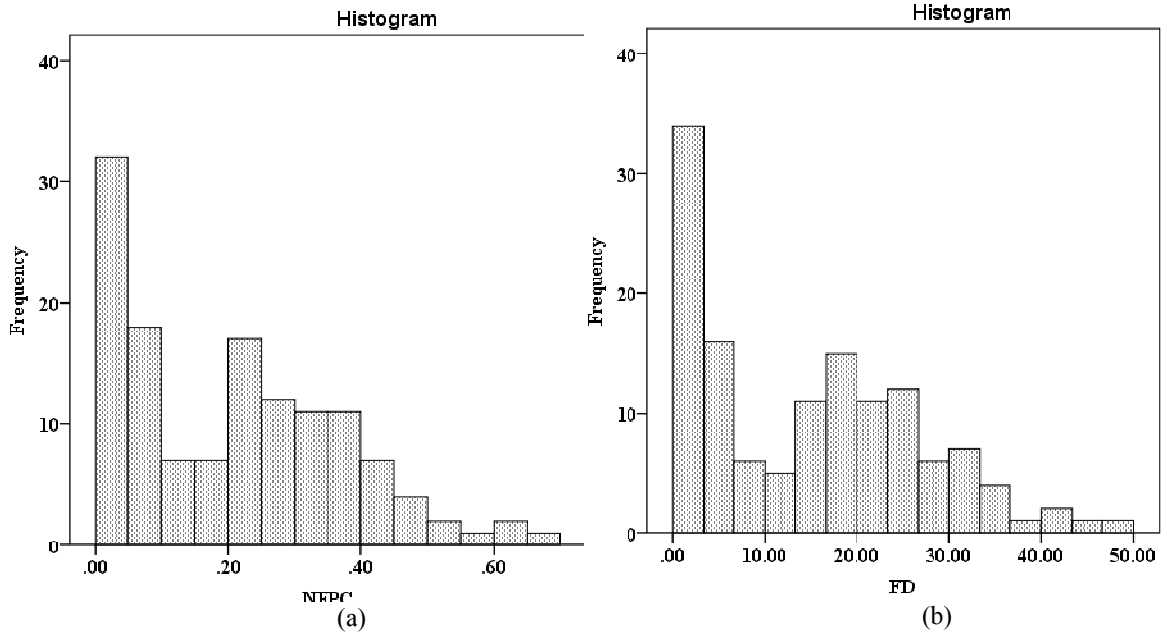
3.2 Identification of outliers

Due to the movement of different types of vehicles on two-lane intercity highways (having unique characteristic of heterogeneous traffic condition), results can differ from the remainders. These data sets are termed as outliers and ultimately affect the data clustering. Therefore it is necessary to remove outliers from the datasetsto develop good clusters. Earlier, Tukey (1977) proposed a rule of thumb for labeling the outliers. In the study,box plot was used to identify the outliers with the help of 1stquartile and 3rd quartile values. A factor "g" with avalue of 1.50 was also proposed to determine the lower limit and upper limit for labelling the outliers. Later, Hoaglin et al. (1986) used the proposed guidelines for labelling the outliers. By doing some simulation analysis, it was concluded that use of 1.50 as multiplier was not accurate as the 50 percent of the time it labeled the data sample as outliers which were not actually. In a subsequent research Hoaglin et al. (1987) proposed a value of 2.20 to use as a multiplier instead of 1.50 and the proposed value of multiplier was identified as the best one to identify the outliers. Beside this in a recent study, Saha et al. (2016) also used the PCA technique todetectthe outliers so that a good cluster can be produced. Ben-Hur and Guyon (2003) proposed a new methodology to examine the merit of clustering. Principal component analysis (PCA) technique was used to identify the best clusters. The samemethodology was adopted in the current research work as used by Hoaglin et al. (1987) to identify the outliers and produce a reliable cluster. For the present study, SPSS was used to identify the outliers. While Figure 1 exhibits the box plot of two performance measures (NFPC and FD) of two-lane intercity highways, Figure 2 shows the distribution of data for both the performance measures.From Figure 1 and 2, no outlier was observed visually. However, onlyon the basis of visual inspection conclusion could not be made. Consequently, lower and upper limits for labelling the outliers were calibrated by using the guidelines proposed by Hoaglin et al. (1987).



(a) (b)

Figure 1: Box plot of (a) NFPC and (b) FD



(a) (b)

Figure 2: Histogram of (a) NFPC and (b) FD

Table 1 exhibits the percentiles (i.e. quartiles) for both measures (i.e. NFPC and FD). Table 2 shows the lower and upper limit for identifying the outliers for NFPC and FD. Thereafter, lower and upper extreme values were identified for NFPC and FD as shown in Table 3. For labelling the outliers, lower and upper limit obtained from the analysis were compared with the extreme values identified for NFPC and FD.From Table 2 and 3, it can be observed that no data sample was identified beyond or below the upper and lower limit which implies that NFPC and FD data sample are free from outliers and can be used for clustering. The similarity and dissimilarity between different data samples are identified

with the help of distance measure. Both (k-means and k-median) clustering techniques are quite similar except the selection procedure for initial cluster center as mentioned earlier.

Table 1**:** Percentiles for NFPC and FD

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average(Definition 1)	NFPC	0.0206	0.0275	0.05	0.207	0.3267	0.4322	0.501
	FD	1.275	1.609	3.211	15.476	23.97	32.092	35.688
Tukey's Hinges	NFPC			0.05	0.207	0.326		
	FD			3.2433	15.4763	23.8923		

Table 2: Lower and Upper Limit Calibration for NFPC and FD

	NFPC	*FD*
Q1	0.05	3.211
Q3	0.3267	23.97
g	2.2	2.2
Q3-Q1	0.2767	20.759
g'	0.60874	45.6698
Lower limit	-0.55874	-42.4588
Upper Limit	0.93544	69.6398

Table 3: Lower and Upper Extreme values for NFPC and FD

			Case Number	*Value*
NFPC	Highest	1	1	.65
		2	2	.61
		3	3	.60
		4	4	.56
		5	5	.54
	Lowest	1	132	.01
		2	131	.01
		3	130	.01
		4	129	.01
		5	128	.01
FD	Highest	1	1	48.43
		2	2	43.60
		3	3	42.41
		4	4	41.58
		5	5	39.03
	Lowest	1	132	.25
		2	131	.50
		3	130	.56
		4	129	.57
		5	128	.71

Different steps are followed during the clustering analysis by considering the both algorithms. Clustering analysis methodology is explained in the form of a flow chart as shown in Figure 3.

Step 1: Select the appropriate algorithm based on field data (i.e. if data are normally distributed then use k-means algorithm otherwise k-median).
Step 2: Identify the number of clusters (K) as per the requirement of the study.
Step 3: Measure the initialcentres for K number of clusters.
Step 4: Calculate the distance between each object or sample w.r.t their cluster centroid by using the different distance metrics (i.e. Equation 1, 2 and 3).
Step 5: Assign different clusters to all the samples or objects based on minimum distance measurement in the previous step.

In the past several cluster validity indexes were used to identify the number of clusters named as Dunn's index, Hartigan index, Calinski-Harabasz index, and Silhouette index and manyothers. Rousseeuw (1987) concluded that silhouette could be used to define the number of clusters (k) for clustering analysis. In an earlier study,Bezdek and Pal (1998) indicated that there is no sole criterion or index (i.e. having advantages over other indices) to validate the number of clusters. A few studies (Arbelaitz et al. 2012, Pollard and Lann 2002) identified that silhouette width index performed well in many comparativeexperiments. Therefore, in the present study, the number of clusters (k) was identified using silhouette plot in MATLAB.

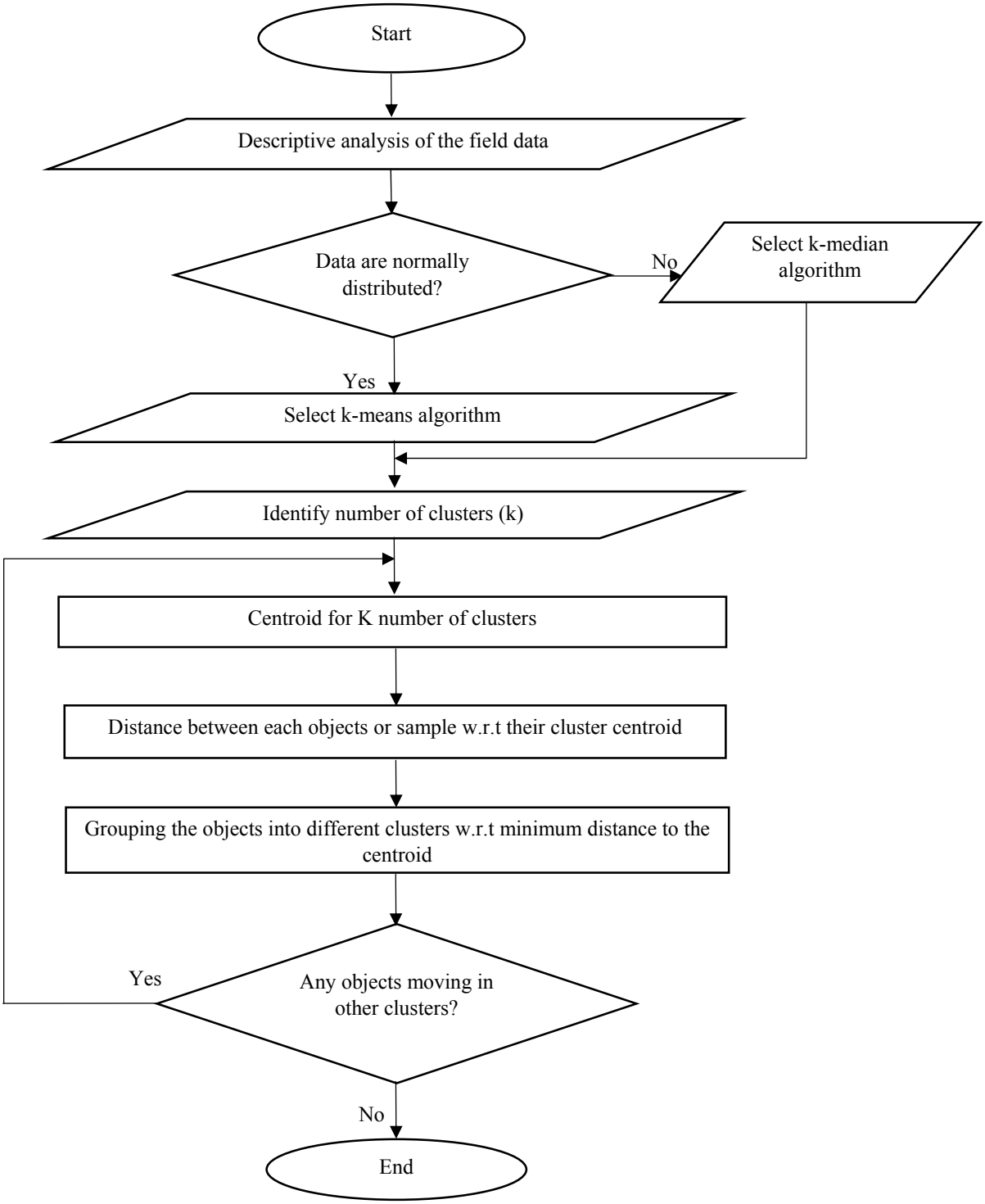


Fig 3.Clustering Analysis Methodology

In this technique number of the clusters are finalized based on the highest average silhouette value. Silhouette can be represented in mathematical form as exhibited in Equation 4 and 5. In the present study, a different number of clusters ranging from 2 to 6 were analysed with the help of silhouette plot as shown in Figure 4 (a) to (e).

Table 4:Validation Criteria of Cluster

Average Silhouette Value	*Interpretation*
0.71-1.0	A strong structure has been found
0.51-0.70	A reasonable structure has been found
0.26-0.50	The structure is weak
<0.25	No substantial structure has been found

$$s\,(i) = \begin{cases} 1 - a(i)/b(i) & if\ a(i) < b(i), \\ 0 & if\ a(i) = b(i), \\ b(i)/a(i)\ - 1 & if\ a(i) > b\ (i). \end{cases}$$

(4)

Silhouette value can also be written as,

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

(5)

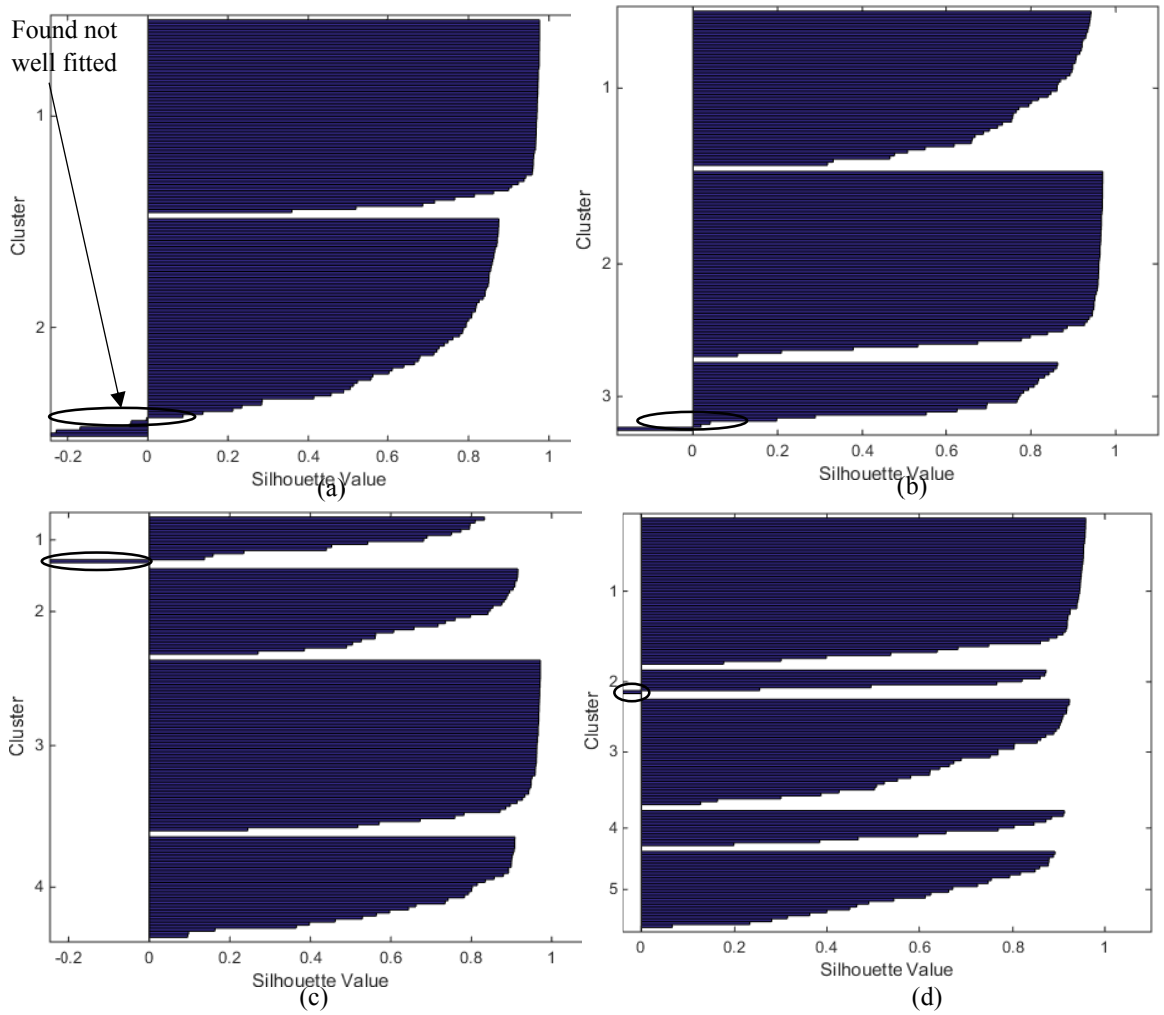
From the Equation 4 and 5 it is clear that

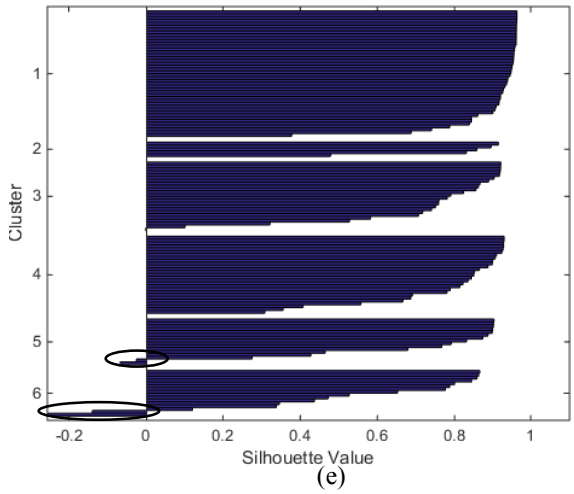
S(i) should be in range of $-1 \leq s(i) \leq 1$

Silhouette plot can be interpreted by using the thumb rule as shown in Table 4. Different silhouette plots exhibited different average silhouette value ranging from 0.77 to 0.81 as shown in Table 5. Silhouette plot having 3 number of cluster is showing highest average silhouette value.

Table 5: Average Silhouette value observed with Different Number of Clusters

Number of Clusters	*Average Silhouette Value*
2	0.78
3	0.81
4	0.78
5	0.75
6	0.77





(e)

Figure 4: Silhouette plot with varying number of cluster

From the silhouette plot, it was observed that three number of clusters could be the best option for clustering analysis. However, a very low variation in silhouette values wereobserved among all the plots. It is to be noted that in all the plots few samples were found not well fitted (i.e. they do not belong to that cluster). Among all the plots, silhouette plot having five number of cluster shows more consistency regarding grouping all the objects in the same cluster where they belong (i.e. very few samples were found to be grouped in another cluster) as the width of wrongly grouped samples is minimum. However,the latest edition of HCM 2010 and many other International standards defined six ranges of LOS for two-lane highways as well as other traffic facilities. Therefore, in this study too5 number ofclusters were adopted to define LOS for two-lane intercity highways under heterogeneous traffic condition. Amorim and Henning (2015) identifyed number of thecluster using adifferent index. It was concluded that silhouette index gives the bestresult. A descriptive analysis was also done to know about the distribution of data samples for both measures as shown in Table 6. From Figure 2 and Table 6 it is clear that data are positively skewed. Therefore, k-median clustering analysis technique was used by considering both measures along with different distance measures.

Table 6: Descriptive Analysis of the NFPC and FD

Summary	*NFPC*	*FD*
Mean	0.208	15.01
Median	0.206	15.47
Skewness	0.56	0.53
Kurtosis	-0.54	-0.61
St. Deviation	0.160	11.88

Figure 5 is exhibits the number of sample size which was grouped in thedifferent cluster during the clustering with the help of Euclidean, Squared-Euclidean and City-block distance metrics. It was observed that among all the metrics Euclidean and Squared-

Euclidean distance metrics grouped similar samples in adifferent cluster. While in thecase of City-block metric data sample grouped in cluster 3 and 4 were found different from other data which were grouped by other distance metrics as shown in Table 7. The main reason was the different algorithm of City-block distance while Square Euclidean distance is the simply square of the Euclidean distance.

Table 7: Output of Clustering Analysis

Cluster Number	*Sample Size*		
	Euclidean	*Sqeuclidean*	*Cityblock*
Cluster 1	611	611	611
Cluster 2	1686	1686	1686
Cluster 3	1960	1960	2035
Cluster 4	1713	1713	1638
Cluster 5	751	751	751

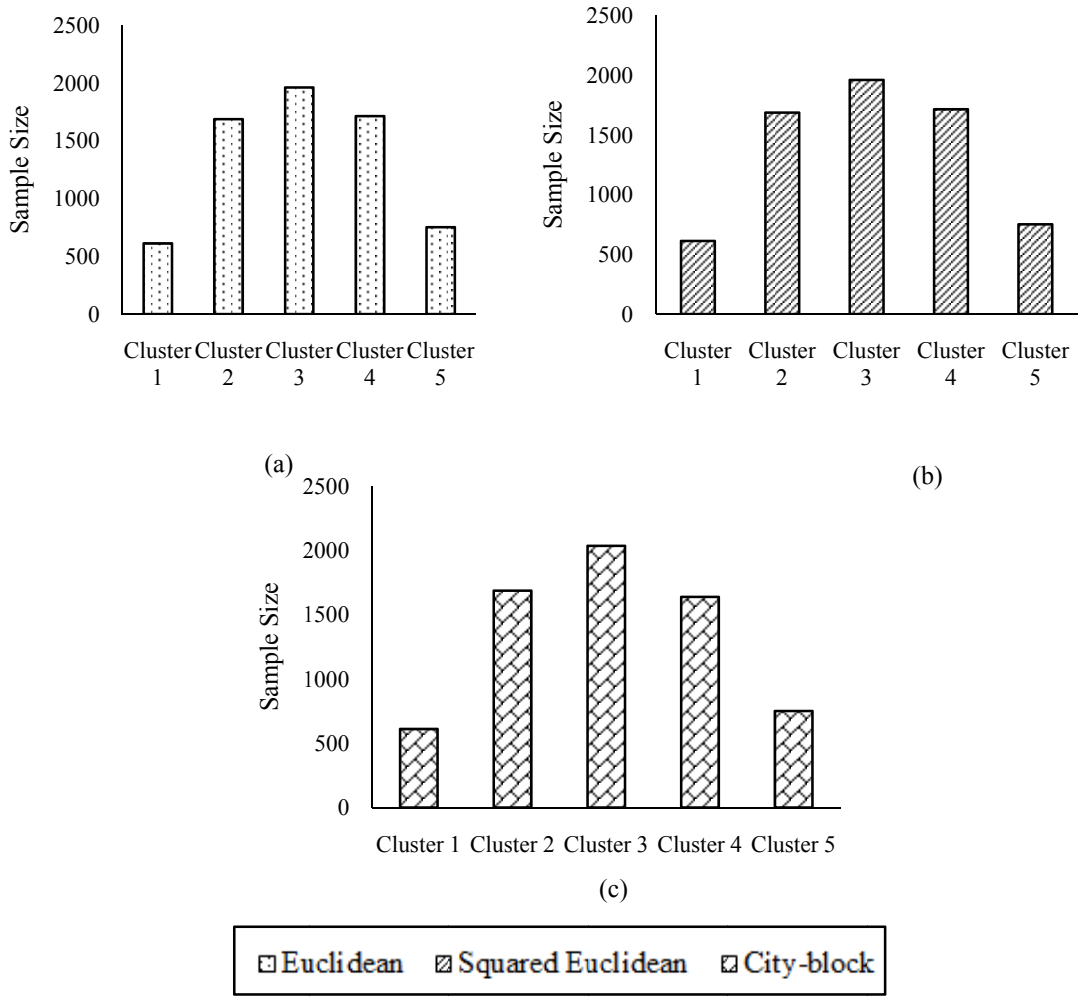


Figure 5: Output from Clustering analysis using different distance metrics

3.3 Identification of Best Distance Metric

The main aim of the current research paper is to identify the best distance metric for clustering analysis which will further help to calibrate different LOS thresholds for two-lane intercity highways under heterogeneous traffic condition. For the same purpose two statistics namely, Cohen Kappa statistics and Silhouette plot were used. Initially, Kappa statistics, is a statistic which measures inter-rater agreement for qualitative objects used to identify the best distance metric for clustering analysis. Kappa coefficient can be represented in mathematical form as shown in Equation 6.

$$K = \frac{P_o - P_e}{1 - P_e} \quad (6)$$

where,
P_o = observed agreement among raters
P_e = observed agreement by chance

Earlier,Brenner and Kliebsch (1996) and Saha (2013)used Kappa statistics to identify the best distance metric. In these studies,linear weighted kappa was found less sensitive to the number of categoriesin comparison to weighted Kappa statistics.Therefore,in the present study quadratic weighted Kappa statistics was used to identify the best distance for clustering analysis. Squared-Euclidean and Cityblock distance were found suitable in the current study because of the highest agreement score as shown in Table 8. The Kappa coefficient can be interpreted by using universal thumb rule as shown in Table 9.

Table 8: Quadratic Weighted Cohen Kappa Expected Agreement Obtained by Partitional Clustering Methods and Sum of Agreement Scores for Each Method

	K-Median[1]	*K-Median[2]*	*K-Median[3]*	*∑Agmt. scr*
K-Median[1]	-	$-0.1176^{(0)}$	$-0.1176^{(0)}$	0
K-Median[2]	$-0.1176^{(0)}$	-	$1^{(4)}$	4
K-Median[3]	$-0.1176^{(0)}$	$1^{(4)}$	-	4
∑Agmt. scr	0	4	4	

- Superscripts of different clustering methods indicate the following distance measures: [1]Euclidean distance; [2]Squared Euclidean distance and [3]Cityblock distance.
- Subscripts to the Kappa Coefficient indicates the agreement score

Table 9: Interpretation of Kappa and Proposed Agreement Score

Kappa Coefficient	*Strength of Agreement*	*Agreement Score*
<0.20	Poor	0
0.21-0.40	Fair	1
0.41-0.60	Moderate	2
0.61-0.80	Good	3
0.81-1.0	Very Good	4

Consequently, the best distance metric selection criteria for clustering analysis was validated with the help of the silhouette plot (Rousseeuw 1987; Saha 2013)to strengthen the results of the study as shown in Figure 6. The significance of silhouette plot can be examined with the help of thickness and width of the silhouette plot curve for each level of service. A large width of the silhouette for each cluster shows the strength of a particular cluster while thickness indicatess the number of data or objects in that particular group. Interpretation of silhouette value can be made with the help of universal thumb rule shown in Table 4. After examination of the silhouette plots, an average value of silhouette 0.80 obtained in thecase of square-Euclidean distance while cityblock distance showed 0.65. Therefore, on the basis ofresults,square-Euclidean distance was identified as the best metric to do the clustering. A relationship was established between NFPC and FD as shown in Figure 7. Later, different LOS thresholds were calibrated using NFPC and FD relationship as shown in Table 10.

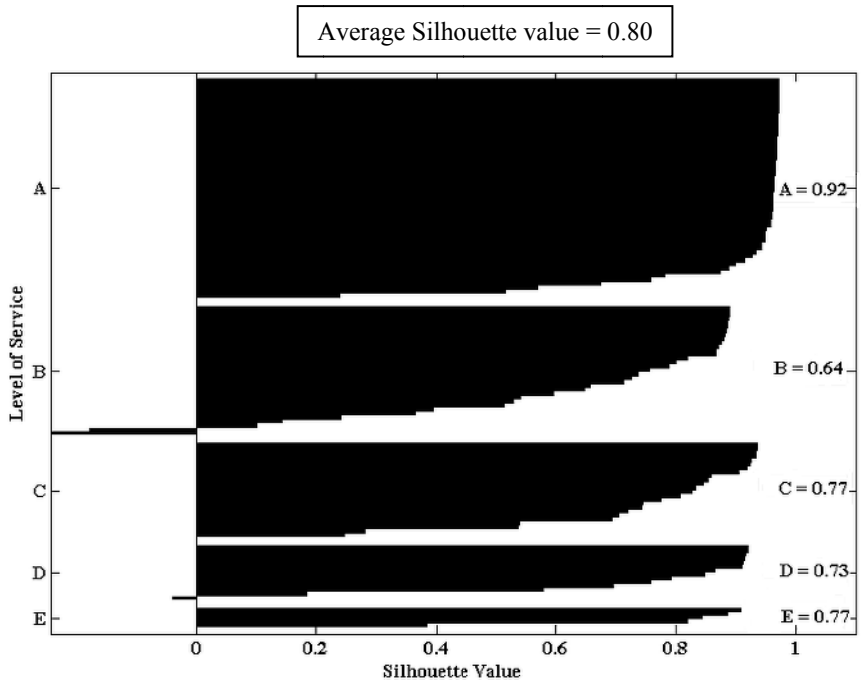


Fig 6. Silhouette Plot by using Square-euclidean Distance for NFPC and FD

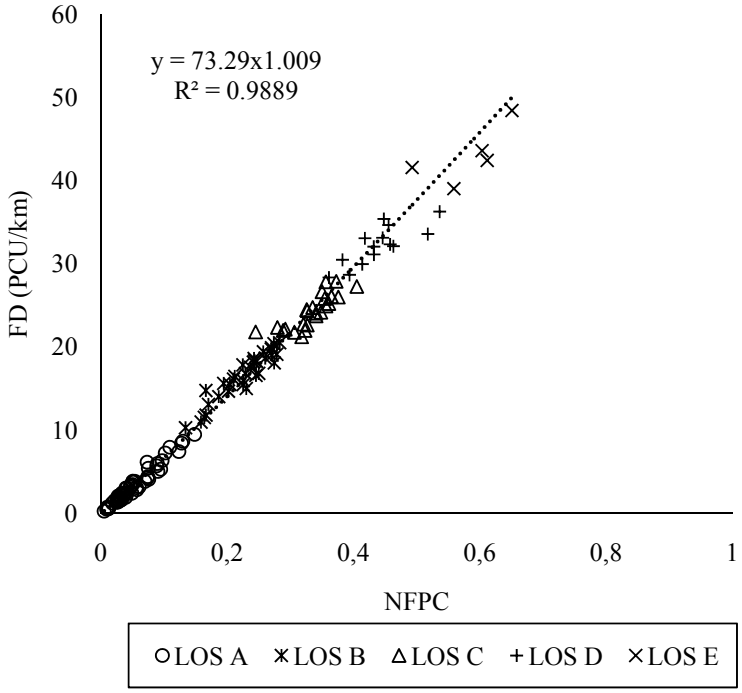


Fig 7.Relationship between NFPC and FD

Table 10: Proposed LOS Ranges Obtained from Partitional Data Clustering

LOS	***NFPC***
A	≤0.13
B	>0.13-0.28
C	>0.28-0.39
D	>0.39-0.50
E	>0.50-0.66

4. Discussion

In the present study, different distance metrics were examined to do the clustering analysis. For the same purpose, data was collected from a previous study conducted by Boora and Ghosh (2016). Initially, the number of clusterswere identified for clustering analysis. Silhouette width index was used to determine the number of cluster. However, ahigher average value of silhouette was obtained when the number of cluster were 3. Due to the identification of wrongly clustered samples (i.e. considering the width of wrongly clustered samples) and the general practice (i.e. for calibrating LOS) in traffic engineering, five number of clusterswere finalized for the current study. Later, outliers were also

detected using previously proposed methodology by Hoaglin et al. (1987) in statistical software SPSS. No outliers were identifiedby carrying out the analysis. On the basis ofdescriptive analysis data samples were found positively skewed.Consequently, k-median clustering analysis was done with the help of three distance metrics named Euclidean, Squared Euclidean and City-block distance. Two statistics named as Cohen Kappa and Silhouette plot were used to identify the best distance metric for calibrating different LOS thresholds for two-lane intercity highways. On the basis of statistical analysis, Squared Euclideanwas identified as the best metrics for clustering. After that different LOS thresholds were calibrated using NFPC and FD relationship and clustering analysis.

5. Conclusions

Based on the findings of the present paper several conclusion has been made which are as follows:

- Previously proposed methodology (Hoaglin et al. 1987) to detect the outliers was found suitable as the both measures showed good correlation with each other.
- Silhouette plot was found as a suitable method to identify the number of cluster.
- Use of Cohen Kappa and Silhouette plot results in the benefit for the study to identify the best distance metric.
- Clustering analysis is identified as an analytical tool for traffic engineering.

The current study suggestresearchers to extend the study by taking more methods for finalizing the number of cluster for clustering analysis and make a comparison by considering all the methods including silhouette index method. More clustering technique can be used to assess different LOS thresholds and comparison can be madeby explaining their merits and demerits. The present study encourages the researchers to use the clustering analysis technique in transportation engineering field for designing and planning of highways under heterogeneous traffic condition.

Acknowledgements

The work reported in this paper is the part of an on-going research project on "Development of Indian Highway Capacity Manual (INDO-HCM)," sponsored by CSIR-CRRI, New Delhi, India. The financial assistance provided by the sponsoring agency for traffic studies is gratefully acknowledged.

References

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J.M., Perona, I.(2013)"An extensive comparative study of cluster validity indices." *Pattern Recognition*, *46*(1), pp.243-256.

Ben-Hur, A.,Guyon, I.(2003)"Detecting stable clusters using principal component analysis." *Functional Genomics: Methods and Protocols*, pp.159-182.

Bezdek, J.C.,Pal, N.R.(1998)"Some new indexes of cluster validity." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *28*(3), pp.301-315.

Bhuyan, P.K.,Mohapatra, S.S.(2014)"Affinity propagation clustering in defining level of service criteria of urban streets." *Transport*, *29*(4), pp.401-411.

Bhuyan, P.K., Rao, K.K. (2010)"FCM clustering using GPS data for defining level of service criteria of urban streets in Indian context." *Transport problems*, *5*(4), pp.105-113.

Boora, A., Ghosh, I.(2016)"Performance indicator for two-lane intercity highways under heterogeneous traffic condition." Paper ID, 129, 19th EURO Working Group on Transportation Meeting (EWGT2016), Istanbul, Turkey.

Bora, M., Jyoti, D., Gupta, D., Kumar, A.(2014)"Effect of different distance measures on the performance of K-means algorithm: an experimental study in Matlab." *arXiv preprint arXiv: 1405.7471*.

Bradley, P.S., Mangasarian, O.L., Street, W.N.(1997)"Clustering via concave minimization." *Advances in neural information processing systems*, pp.368-374.

Brenner, H.,Kliebsch, U. (1996)"Dependence of weighted kappa coefficients on the number of categories." *Epidemiology*, pp.199-202.

deAmorim, R.C.,Hennig, C.(2015)"Recovering the number of clusters in data sets with noise features using feature rescaling factors." *Information Sciences*, *324*, pp.126-145.

Hoaglin, D.C.,Iglewicz, B., 1987. Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *82*(400), pp.1147-1149.

Hoaglin, D.C., Iglewicz, B.,Tukey, J.W.(1986)"Performance of some resistant rules for outlier labeling." *Journal of the American Statistical Association*, *81*(396), pp.991-999.

Jain, A.K., Murty, M.N. and Flynn, P.J.(1999)"Data clustering: a review." *ACM computing surveys (CSUR)*, *31*(3), pp.264-323.

Jiang, G.Y., Wang, J.F., Zhang, X.D., Gang, L.H. (2003)"The study on the application of fuzzy clustering analysis in the dynamic identification of road traffic state." In *Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE* (Vol. 2, pp. 1149-1152). IEEE.

Murugesan, R.,Moorthy, N.V.(1998)"Level of public transport service evaluation: a fuzzy set approach." *Journal of advanced transportation*, *32*(2), pp.216-240.

Peng, W., Li, T.(2006)"November. Interval data clustering with applications." In *Tools with Artificial Intelligence, 2006. ICTAI'06. 18th IEEE International Conference on* (pp. 355-362). IEEE.

Pollard, K.S., Van Der Laan, M.J.(2002)"A method to identify significant clusters in gene expression data."

Prassas, E., Roess, R.,McShane, W.(1996)"Cluster analysis as tool in traffic engineering." *Transportation Research Record: Journal of the Transportation Research Board*, (1566), pp.39-48.

Qiu, Y., Ma, S.,Xiong, D. (2012)"Layout planning towards urban transportation hub based on clustering analysis method." In *Service Systems and Service Management (ICSSSM), 2012 9th International Conference on* (pp. 706-709). IEEE.

Rousseeuw, P.J.(1987)"Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics*, *20*, pp.53-65.

Saha, P., Bhadra, A., Reddy, N.S., Sarkar, A.K.(2013)"Method of Identifying Low Performance Vehicles in Heterogeneous Traffic on Two-Lane Highways." *Procedia-Social and Behavioral Sciences*, *104*, pp.526-532.

Saha, P., Roy, N., Mukherjee, D., Sarkar, A.K.(2016)"Application of Principal Component Analysis for Outlier Detection in Heterogeneous Traffic Data." *Procedia Computer Science*, *83*, pp.107-114.

Saha, P., Sarkar, A.K., Pal, M. (2015)"Assessment of level-of-service of two-lane highways with heterogeneous traffic." In *Transportation Research Board 94th Annual Meeting* (No. 15-2723).

Singh, A., Yadav, A.,Rana, A.(2013)"K-means with Three Different Distance Metrics." *International Journal of Computer Applications*, *67*(10).

TRB. (2010). "Highway Capacity Manual (2010)" *Transportation Research Board, National Research Council, Washington, D.C*.

Tukey, J.W., 1977. Exploratory data analysis.