Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 104 (2013) 430 – 439

Procedia
Social and Behavioral Sciences

2nd Conference of Transportation Research Group of India (2nd CTRG)

A data mining approach to creating fundamental traffic flow diagram

Jalil Kianfar[a,b], Praveen Edara[a,1]

[a]University of Missouri, C2640 Lafferre Hall, Columbia, Missouri 65211, USA
[b]Crawford, Bunte, Brammeier (CBB), 1830 Craig Park Court, Suite 209, St. Louis, Missouri 63146, USA

Abstract

This paper investigates application of clustering techniques in partitioning traffic flow data to congested and free flow regimes. Clustering techniques identify the similarities and dissimilarities between data, and classify the data into groups with similar characteristics. Such techniques have been successfully used in market research, astronomy, psychiatry, and transportation. A framework is proposed for clustering traffic data based on fundamental traffic flow variables. Three types of clustering techniques are investigated: 1) connectivity-based clustering, 2) centroid-based clustering, and 3) distribution-based clustering. Specifically, hierarchical clustering, K-means clustering and general mixture model (GMM) were investigated.
Traffic sensor data from three freeway bottleneck locations in two major U.S. metropolitan areas, St. Louis, Missouri, and Twin Cities, Minnesota, were used in the study. Various combinations of traffic variables were investigated for all three clustering techniques. The results indicated that the clustering is an effective way to partition traffic data into the free flow and congested flow regimes. Partitioned traffic data can be used to create fundamental traffic flow diagrams and macroscopic traffic stream models. Using speeds, or both speeds and occupancies as input variables produced the best clustering results. The performance of K-means and hierarchical clustering techniques were comparable to each other and they outperformed GMM clustering.

Keywords: Clustering; Fundamental diagram; Traffic flow

1. Introduction

This paper developed a new framework based on clustering techniques to create flow-occupancy diagrams and to identify critical occupancy. Traffic flow theory literature commonly categorizes traffic flow conditions to two

[1] * Corresponding author. Tel.: +1-573-882-1900; fax: +1-573-882-4784.
E-mail address: edarap@missouri.edu

doi:10.1016/j.sbspro.2013.11.136

regimes: free flow and congested flow (Dervisoglu et al., 2009; Muralidharan et al., 2011; Thankappan & Vanajakshi, 2012). Traffic data is partitioned into different regimes by identifying breakpoints for traffic variable(s) in the data. In two-regime traffic models, critical occupancy is used to separate free flow and congested flow conditions. Observations with occupancy values smaller than critical occupancy are assumed to be in the free flow regime and observations with occupancy values greater than critical occupancy are assumed to be in the congested flow regime. Identifying the critical occupancy value from field observations is not trivial. There is no universally recommended procedure for determining critical occupancy from traffic data, and oftentimes it is up to researchers' judgment (Sun & Zhou, 2005).

A procedure that does not entirely rely on researchers' judgment would allow for consistency in identifying critical occupancy values, and therefore the flow regimes, across different study locations. With this motivation, this paper investigates the application of clustering techniques in partitioning traffic flow data to free flow and congested regimes. Clustering techniques identify the similarities and dissimilarities between data and classify the data into groups with similar characteristics. Such techniques have been successfully used in market research, astronomy, psychiatry, and transportation (Kianfar & Edara, 2010). A framework is suggested for clustering traffic data based on fundamental traffic flow variables. Three categories of clustering techniques are investigated: 1) connectivity-based clustering, 2) centroid-based clustering, and 3) distribution-based clustering. Hierarchical clustering, K-means clustering and general mixture model (GMM) are examples of each category of clustering techniques that were investigated.

Clustering techniques are used to partition traffic flow data into free flow and congested flow. The clustering algorithm result is used to create regression fits and to develop a flow-occupancy diagram. One limitation of this type of approach or any other automated flow-occupancy plot generation method, is that the outputs of the method are difficult to validate. There is no correct answer to serve as a benchmark for validation of automated method outputs. One crude way to validate is to use a dataset that is easy to visually determine the critical occupancy value, which can then be compared to the value obtained from clustering. In terms of clustering effectiveness, there are several measures for evaluating outputs of clustering techniques. These measures are used here to investigate whether data is well-partitioned. Performance measures such as the Davies–Bouldin index (Davies & Bouldin, 1979), the Dunn index (Bezdek & Pal, 1998), and the Silhouette coefficient (Tan et al., 2007) are used to evaluate clustering results.

In the next section, the state of the art in automated methods for creating flow-occupancy diagrams and applications of pattern recognition techniques in traffic engineering are reviewed. Three clustering algorithms used in this paper and three measures for evaluating clustering outputs are presented in section 3. A framework for selecting the clustering technique and selection of input variables is proposed in section 4. In section 5, the developed framework is applied to analyze traffic flow data at three freeway bottleneck locations. The results of clustering algorithm (i.e. free flow and congested flow partitions) are used to create flow-occupancy fits and to identify the critical occupancy. Performance of different techniques and input variable are compared in section 5. Section 6 presents the summary and conclusions.

2. Literature review

This section reviews the state of the art in automated methods for creating flow occupancy diagram and reviews the relevant applications of pattern recognition in traffic engineering. Pattern recognition and clustering techniques have been used in various areas of transportation engineering such as safety analysis (Mohamed et al., 2013), pavement maintenance (Sandra & Sarkar, 2013), transit (Ma et al, 2013), and identifying locations with similar traffic pattern on a road network (Chunchun et al., 2011).

2.1. Automated methods for creating flow-occupancy diagram

Dervisoglu et al. (2009) suggest a framework for calibrating flow-occupancy diagram for a freeway section. The objective of their study is to calibrate a cell transmission model for the freeway section. To do this, a regression line is fitted to flow-occupancy data with speeds greater than 88 km/h (55 mph) to create the free flow fit. Next, the Highway Capacity Manual (HCM, 2000) definition of capacity is used to determine the capacity of freeway segment. Following this, the determined capacity is horizontally projected to the free flow fit to determine the tip of flow-occupancy diagram and the critical occupancy is identified. Finally, an approximate quantile regression technique is used to create the overcritical flow occupancy fit. The suggested framework is applied to traffic sensor data from Interstate-880 in San Francisco Bay Area, California.

Li and Zhang (2011) suggest a method for creating flow-occupancy diagram. In the first step, traffic data is partitioned to free flow and congested flow based on characteristics of fluctuations in traffic data time series. Next, a minimum principle technique is used to identify equilibrium states followed by the application of a mixed integer optimization technique to create piecewise linear flow-occupancy fits. The objective of the optimization procedure is to obtain fits with minimum absolute deviation. The suggested methodology was applied to three freeway locations in California; the results of which complied with well-known traffic engineering principles.

In summary, the aforementioned studies suggested automated methods for creating flow-occupancy diagrams. In the methodology suggested by Dervisoglu et al. (2009), free flow speed should be determined by researchers to partition the traffic data to free flow and congested flow regimes. The methodology of Li and Zhang (2011) is based on an optimization model to create the best-fit lines; however, several parameters should be identified by the user during the optimization process. In the clustering-based methodology presented in this paper, there is no need to identify the free flow speed or any other parameter by the user and thus the methodology is less dependent on user judgment. As we later discuss, some user knowledge is still needed for verifying if the clustering results are in agreement with the basic constructs of traffic flow theory.

2.2. Application of pattern recognition in traffic engineering

Sun and Zhou (2005) apply clustering techniques in the modeling of multi-regime speed-density relationships. Clustering techniques are used to identify the breakdown points in a speed-density diagram, speed-density data is then partitioned based on the identified breakpoints, and linear regression techniques is used to create multi-regime speed-density relationships. K-means clustering algorithm is applied to three data sets from three highway sections in San Antonio, Texas. Speed-density data is then clustered to two and three groups. Linear, logarithmic and exponential regression techniques are used to create speed-density models. The authors conclude that clustering techniques provide modelers with a natural method for partitioning the data into different regimes and are helpful in identifying the breakpoints in traffic data sets.

Xia and Chen (2007) classify freeway operating conditions using an agglomerative clustering algorithm. Flow, speed, and occupancy data from a freeway detector in California are used as a case study. Bayesian Information Criterion (BIC) and dispersion measurement techniques are used to identify the number of clusters. The analysis suggest that each cluster could represent a freeway flow phase. They found out that occupancy and flow are the most significant variables affecting the clustering results. The study also suggests that occupancy is likely the most significant parameter in partitioning the data into congested and uncongested clusters. Occupancy is also the most important factor in further partitioning the congested cluster. Flow is the most important factor in further clustering the uncongested partition.

Azimi and Zhang (2010) use K-means, fuzzy C-means and CLARA algorithms to classify traffic conditions on a freeway segment in Austin, Texas. Traffic data is partitioned to six clusters, and the outputs are compared with Highway Capacity Manual thresholds for level of service. The clustering process is repeated with two different sets of input variables: 1) (flow, speed) and 2) occupancy. The outputs of K-means algorithm are found to be the most consistent with Highway Capacity Manual level-of-service thresholds. Since there is only one level of service for oversaturated flow conditions in the Highway Capacity Manual, the authors suggest using clustering techniques to further analyze oversaturated flow conditions.

In another study, Xia et al. (2012) suggest a clustering method for online identification of traffic states. Their approach is based on an agglomerative clustering algorithm modified in such way that instead of storing all the historical traffic data, it stored statistical features of data. Data from two traffic detectors at Interstate 80 in California are considered in the study. Flow, speed and occupancy are used as input variables of the clustering algorithm. Traffic data is partitioned into congested and uncongested clusters; the congested cluster then being further partitioned into three sub-clusters, and the uncongested cluster partitioned into two sub-clusters.

Previous research has utilized clustering techniques to model speed-density relationship, to cluster freeway operating conditions, and to cluster traffic conditions to six categories representing six freeway level of services (LOS). K-means clustering, hierarchical clustering (agglomerative), fuzzy C-means and CLARA techniques have been considered in these previous studies. Some studies have used Bayesian Information Criterion for evaluating the clustering outputs. This paper investigates application of K-means, hierarchical and general mixture model (GMM) clustering techniques for creating flow-occupancy diagram; the results of which are compared using several cluster validation indexes. The next section introduces the clustering techniques used to cluster traffic data into free flow and congested flow traffic regimes.

3. Pattern recognition techniques and validity indexes

This paper investigates the application of hierarchical clustering, K-means clustering, and general mixture model (GMM) clustering for partitioning traffic data to free flow and congested flow regimes. The results of clustering algorithms are then evaluated using Davies–Bouldin index, Dunn index, and Silhouette coefficient. For brevity, a brief description of clustering techniques and validity indexes is presented in this section. Additional details on the algorithms and validation indexes can be found in the cited literature.

3.1. Clustering techniques

The objective of clustering techniques is to classify a set of data into groups or clusters. Data elements are grouped such that similar elements are assigned to the same group (Cios et al, 1998). Data elements are arranged into clusters so that members of each cluster are homogeneous while clusters are heterogeneous (Hair et al., 2005). In hierarchical clustering, the initial number of clusters is assumed to be equal to the number of available data points, and each data point (same as a data element) is considered a pattern. Similar clusters are then merged together based on their similarity in a step-by-step procedure until the required number of clusters is achieved (Cios et al., 1998).

In K-means clustering, means stands for the average location of all the members of a particular cluster. K is the number of clusters. K-means technique is an iterative procedure involving the computation of cluster centroids. A centroid is an artificial point in space that represents the average location of the particular cluster. Each data point is assigned to the closest centroid. Location of each centroid is updated in each iteration until no significant change is observed in the location of centroids (Kianfar & Edara, 2010).

In General Mixture Model clustering (GMM), clusters are modeled with parametric distribution functions, and the entire dataset is represented by a mixture of those distributions. The most common model in GMMs is the mixture of Gaussians. In order to identify clusters, the parameters of Gaussian mixture are usually estimated using the Expectation Maximization (EM) algorithm (Law et al. 2004).

3.2. Clustering validity indexes

This section introduces three validity indexes used to evaluate the quality of clustering: Davies–Bouldin index, Dunn index and Silhouette coefficient. These three measures are among the most common internal cluster evaluation measures used for pattern recognition (Maulik & Bandyopadhyay, 2002; Pakhira et al., 2004; Žalik & Žalik, 2011). The results of hierarchical, K-means and GMM techniques in clustering traffic data are compared using these three measures.

The Davies–Bouldin (DB) index is defined based on the ratio of "sum of within-cluster scatter to between-cluster separation" (Pakhira et al, 2004). A smaller DB index shows a better clustering output (Davies & Bouldin, 1979).

The Dunn index (DI) is a cluster validity index that that represents "within cluster and between cluster separations" (Pakhira et al, 2004). Higher Dunn index values represent better clusters (Bezdek & Pal., 1998).

The Silhouette coefficient (SC) considers the impact of each cluster member on cohesion or separation of the cluster. The average silhouette value of all data points is called the Silhouette coefficient (Tan et al., 2007). The silhouette coefficient value varies between -1 and +1; a value of +1 indicates strong clustering structure whereas a value of -1 means that data points are misclassified.

4. Framework for clustering traffic data

Clustering techniques can be used to facilitate the process of partitioning traffic data into free flow and congested flow datasets. However, clustering techniques are not fully automated, and the user has to select the number of clusters and the input variables for each clustering technique. In the process of clustering traffic data, the following questions should be answered:

1. Which clustering technique is capable of distinguishing different features of free flow and congested flow regimes?
    - To answer this question, K-means, GMM and hierarchical clustering techniques are investigated using different case studies. Based on case studies' results, performance of clustering techniques for partitioning traffic data is evaluated.
2. What are the input variables for clustering techniques? Flow, occupancy (as a surrogate for density) and speed are three traffic variables commonly available from traffic sensors. To determine input variable(s) for clustering techniques, all possible combinations of traffic variable(s) were investigated; these input variable combinations are:
    - Three variables: (flow, occupancy, speed)
    - Two variables: (flow, occupancy), (flow, speed), or (speed, occupancy)
    - Single variable: (flow), (occupancy), or (speed)

All possible combinations of clustering techniques and input variables are used to partition traffic data into free flow and congested flow datasets. A combination of 1) clustering technique, and 2) input variable(s) forms a clustering scenario. For example, using K-means technique to cluster flow and speed data to two partitions is a clustering scenario. Another clustering scenario is to use K-means technique to cluster speed and occupancy data to two partitions. Each clustering scenario provides a free flow dataset and a congested flow dataset. Clustering scenarios results are evaluated using validity indexes described in Section 3.2 and the best clustering scenario is identified to partition traffic data to free flow regime and congested flow regime. The next step is to use the free flow and congested flow datasets to create the flow-occupancy diagram.

Past studies have demonstrated that there are two different types of maximum flow observed at freeway bottlenecks: queue discharge flow (QDF) and pre-queue flow (PQF) (Banks, 1999; 2002; 2009a; 2009b; Hall & Agyemang-Duah 1991; Persaud et al. 2001). The two types of maximum flow values are usually obtained from flow-occupancy plots (Banks, 2006) – 1) pre-queue flow (PQF) which is the maximum flow observed during the free flow conditions, and 2) queue discharge flow (QDF) which is the maximum flow observed during congested flow conditions.

Figure 1 shows the different stages of creating a flow-occupancy diagram from raw data. Free flow and congested flow datasets are classified using the clustering technique (Figure 1(a)). Linear regression is used to create the best-fit line for free flow cluster observations. Similar to the work of Dervisoglu et al. (2009), free flow fit is then drawn up to match the highest observed flow value (Figure 1(b)). This value is identified as the PQF. The occupancy corresponding to PQF (on the best-fit line) is selected as the critical occupancy. The next step is to create the congested flow plot. Linear regression is used to create a best fit-line from congested flow cluster observations. The congested flow plot is then drawn from critical occupancy to maximum observed occupancy (Figure 1(c)). The QDF value can then be inferred from the congested flow plot (flow at critical occupancy). In the next section, data from two bottleneck locations in St. Louis, Missouri and one bottleneck location in the Twin Cities, Minnesota are used to evaluate different clustering scenarios and to create flow-occupancy diagrams.

5. Case study

This section presents the results of clustering traffic data from study locations A and B in St. Louis, and study location C in the Twin Cities. Location A was between the second exit ramp and second entrance ramp of a full cloverleaf interchange. Location B was between the exit ramp and entrance ramp of a diamond interchange. Location C was downstream of a directional interchange and upstream of a partial cloverleaf interchange. Figure 2 shows the geometrics of study locations.

Traffic data is partitioned to two clusters (representing two-regime traffic models). Different sets of input variables are investigated for each clustering technique:

- flow, speed, occupancy
- flow, occupancy
- flow, speed
- speed, occupancy
- flow
- speed
- occupancy

There are 21 possible clustering scenarios for each study location. Table 1 reports validity indexes for clustering scenarios for study location A. Davies–Bouldin index, Dunn index, and Silhouette coefficient are reported for each clustering scenario. Smaller Davies–Bouldin index (DB), higher Dunn index (DI) and Silhouette coefficient (SC) close to 1.0 represent better clustering results. In terms of input variables, results are very close when single variable speed or occupancy, or two variables speed and occupancy are used as input

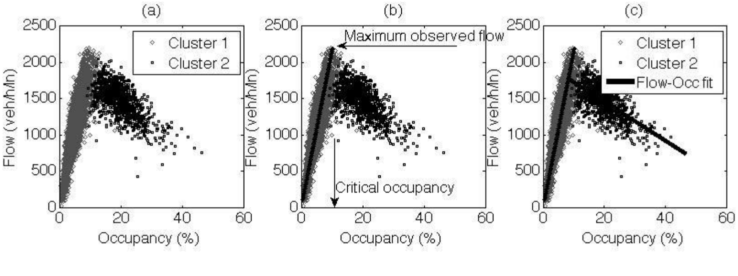

Fig. 1. Creating the flow-occupancy diagram

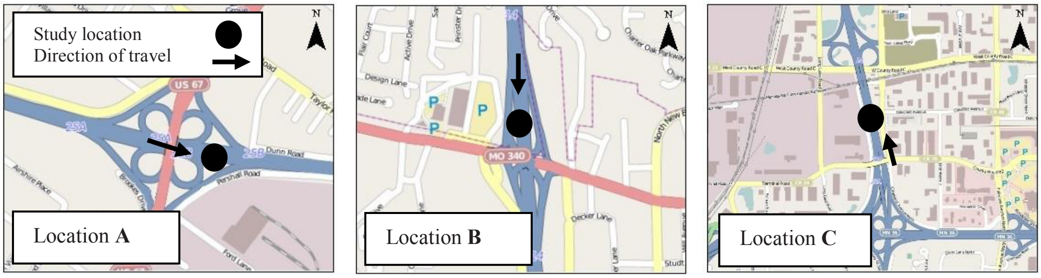

Fig. 2. Geometrics of study locations (Source: ArcMap OpenStreetMaps)

variable(s). For example when K-means technique is used to cluster location A data, the Silhouette coefficient is: 1) 0.97 when speed is used as the only input, 2) 0.94 when occupancy is used as the only input, and 3) 0.96 when both speed and occupancy are used as input. This trend is consistent for all three clustering techniques. In terms of clustering techniques, results of K-means and hierarchical techniques are comparable and outperform GMM technique. Based on the cluster validity indexes' for different scenarios, the best results for location A data set are obtained when single variable speed is clustered using the K-means technique. Speed is identified as the best variable for clustering traffic data at location A to two regimes. This finding is in agreement with other studies (Lorenz & Elefteriadou, 2001; Ozbay & Ozguven, 2007; Shawky & Nakamura, 2007) that used speeds to distinguish congested and free-flow conditions. However, one study conducted by Xia and Chen (2007) found occupancy to be the most significant variable in partitioning traffic data. In the current study, however, the clustering results when using 'speed' were slightly better than the results when 'occupancy' was used (see Table 1 rows 5 and 6 for K-means).

It is noted that the result of a clustering technique should not be thoughtlessly accepted, and some amount of engineering judgment is still needed to verify the compliance of results with traffic flow theory principles. Figure 3(a) to 3(c) show the result of clustering data at location A for K-means (flow) scenario. Figure 3(d) to 3(f) show the result of clustering the same data set for K-means (speed) scenario. It is observed that the result of K-means (flow) scenario do not follow accepted traffic flow theory principles. The clustering validity indexes presented in Table 1 for K-means (flow) scenario are notably worse than validity indexes for K-means (speed) scenario.

To further illustrate the clustering performance, time series of traffic variables at location A for one day are shown in Figure 4. Plots 4(a), 4(b), and 4(c) show time series of speed, flow, and occupancy from 6:30 a.m. to 9:30 p.m. on April 3, 2008. The clustering technique accurately identified congestion in both morning and afternoon peak periods.

Table 1. Clustering validity indexes for location A

Input Data	K-means			GMM			Hierarchical		
	DB	DI	SC	DB	DI	SC	DB	DI	SC
flow, speed, occupancy	0.62	0.00	0.70	1.06	0.00	0.24	0.46	0.00	0.62
flow, occupancy	0.62	0.00	0.70	1.05	0.00	0.25	0.44	0.00	0.63
flow, speed	0.62	0.00	0.70	1.05	0.00	0.25	0.53	0.00	0.60
speed, occupancy	0.29	0.01	0.96	0.43	0.00	0.92	0.25	0.01	0.95
occupancy	0.35	0.00	0.94	0.45	0.00	0.91	0.20	0.02	0.89
speed	0.26	0.00	0.97	0.40	0.00	0.93	0.31	0.00	0.95
flow	0.62	0.00	0.70	0.88	0.00	0.62	0.44	0.00	0.63

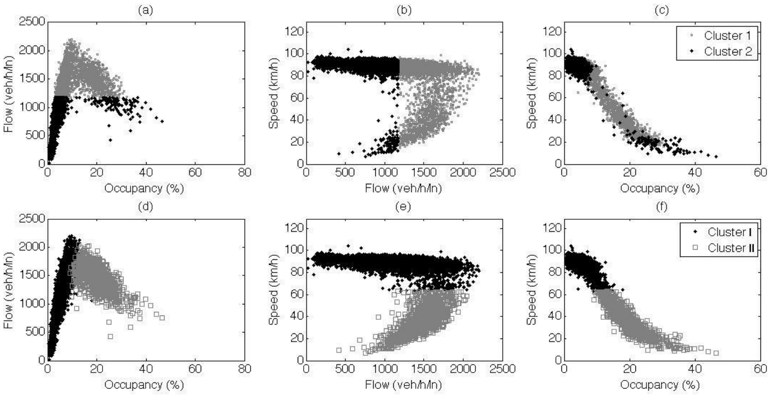

Fig. 3. Comparison of K-means (flow) and K-means (speed) scenarios for location A

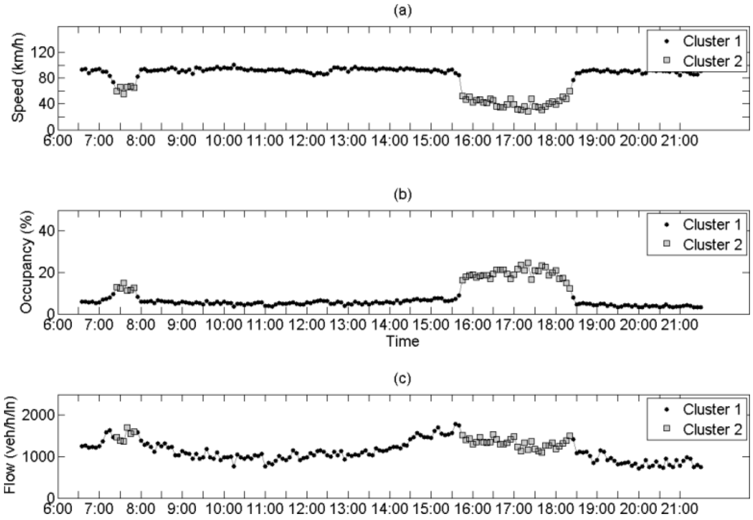

Fig. 4. Time series of traffic at location A on April 3, 2008

Table 2 summarizes best clustering scenarios for study locations A and B in St. Louis and C in the Twin Cities. In Table 2 each row is referred to as a case and represents the best scenario out of 21 evaluated scenarios for each location. In all three cases, K-means technique provided the best results. In terms of input variable(s), single variable speed produced the best results in two cases. Speed and occupancy when used together produced the best results in one case.

After identifying the two regimes, curve fitting was used to fit flow-occupancy plots for each regime as previously discussed. Linear regression is used to create best-fit lines for free flow and congested flow regimes. Figure 5 shows the best-fit lines for locations A, B in St. Louis and location C in the Twin Cities. Table 3 presents critical occupancy, pre-queue flow (PQF), and queue discharge flow (QDF) for locations A, B, and C. Critical occupancy, PQF, and QDF values for location C in the Twin Cities are higher than the respective values observed for locations A and B is St. Louis. As reported in previous research, the PQF value was greater than the QDF value at all sites.

Table 2. Summary of clustering before and after data sets at locations A, B and C

Study location	Clustering technique	Input variable(s)	DB	DI	SC
A	K-means	speed	0.26	0.00	0.97
B	K-means	speed, occupancy	0.29	0.01	0.95
C	K-means	speed	0.38	0.01	0.95

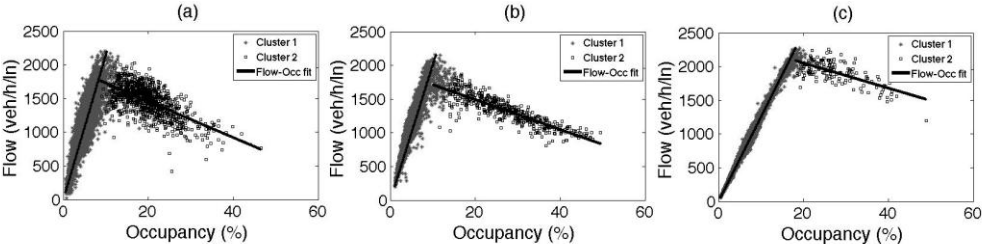

Fig. 5. Flow-occupancy fits for locations A and B in St. Louis and location C in the Twin Cities

Table 3. Critical occupancy, queue discharge flow and pre-queue flow

Study location	Critical occupancy (%)	Pre-queue flow (veh/h/ln)	Queue discharge flow (veh/h/ln)
A	10.25	2195	1734
B	11.00	2155	1699
C	18.00	2271	2089

6. Conclusion

This paper proposed a clustering-based automated method for creating flow-occupancy diagrams. The proposed method aims to alleviate some of the researcher's judgment involved in generating flow-occupancy diagrams by not requiring assumptions on any parameters, and thus allowing for consistency in identifying key traffic stream characteristics such as critical occupancy, pre-queue flow, queue discharge flow, across different study locations. A framework for clustering traffic data based on fundamental traffic flow variables was proposed. Hierarchical clustering, K-means clustering and the general mixture model (GMM) were evaluated. Performance measures Davies–Bouldin index, Dunn index, and Silhouette index were used to evaluate clustering results. Traffic sensor data from three freeway bottleneck locations in two major U.S. metropolitan areas, St. Louis, Missouri, and the Twin Cities, Minnesota, were used as case studies. Various combinations of traffic variables and clustering techniques were investigated. Partitioned traffic data was used to create flow-occupancy diagrams. Linear regression was used to create flow-occupancy fits. The results indicate that the clustering is an effective way to partition traffic data into free flow and congested regimes. Using speeds, or both speeds and occupancies, as input variables produced the best clustering results. The outputs of K-means and hierarchical clustering techniques were comparable to each other and outperformed GMM clustering.

One additional advantage of using clustering techniques for the intended application is that many commercial and open-source programming languages have in-built routines for different clustering techniques as well as the common validity indexes. In future research, data from additional bottleneck locations may be investigated to further validate the findings of current study.

References

Azimi, M., & Zhang, Y. (2010). Categorizing freeway flow conditions by using clustering methods. *Transportation Research Record, 2173*, 105–114.

Banks, J. H. (1999). Investigation of some characteristics of congested flow. *Transportation Research Record, 1678*, 128–134.

Banks, J. H. (2002). Review of empirical research on congested freeway flow. *Transportation Research Record,* 1802, 225–232.

Banks, J. H. (2006). New approach to bottleneck capacity analysis: Final report. California PATH Program, Institute of Transportation Studies, University of California at Berkeley.

Banks, J. H. (2009a). Automated analysis of cumulative flow and speed curves. *Transportation Research Record, 2124*, 28–35.

Banks, J. H. (2009b). Flow breakdown at freeway bottlenecks. *Transportation Research Record, 2099*, 14–21.

Bezdek, J. C., & Pal, N. R. (1998). Some new indexes of cluster validity. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 28(3)*, 301–15.

Chunchun, H., Nianxue, L., Xiaohong, Y., & Wenzhong, S. (2011). Traffic flow data mining and evaluation based on fuzzy clustering techniques. *International Journal of Fuzzy Systems, 13(4),* 344-349.

Cios, K. J., Pedrycz, W., & Swiniarski, R. W. (1998). *Data Mining Methods for Knowledge* (p. 520). Springer.

Davies, D. L. & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2)*, 224–227.

Dervisoglu, G., Gomes, G., Kwon, J., Horowitz, R., & Varaiya, P. (2009). Automatic calibration of the fundamental diagram and empirical observations on capacity. In: Proceedings of the Transportation Research Board 88th Annual Meeting (DVD), Washington, D. C.

Hair, J. F., Black, B., Babin, B., Anderson, R. E., & Tatham, R. L. (2005). *Multivariate Data Analysis (6th Edition)*. Prentice Hall.

Hall, F. L., & Agyemang-Duah, K. (1991). Freeway capacity drop and the definition of capacity. *Transportation Research Record, 1320*, 91-98.

Kianfar, J., & Edara, P. (2010). Optimizing freeway traffic sensor locations by clustering global-positioning-system-derived speed patterns. *IEEE Transactions on Intelligent Transportation Systems, 11(3),* 738–747.

Law, M. H. C., Figueiredo, M. A. T., & Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9)*, 1154–66.

Li, J., & Zhang, H. M. (2011). Fundamental diagram of traffic flow. *Transportation Research Record, 2260*, 50–59.

Lorenz, M. R., & Elefteriadou, L. (2001). Defining freeway capacity as function of breakdown probability. *Transportation Research Record, 1776*, 43–51.

Ma, X., Wu, Y. J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. In: Proceedings of the Transportation Research Board 92nd Annual Meeting (DVD), Washington, D. C.

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(12), 1650–1654.

Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F., & Ukkusuri, S. V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. *Safety Science, 54*, 27–37.

Muralidharan, A., Dervisoglu, G., & Horowitz, R. (2011). Probabilistic graphical models of fundamental diagram parameters for simulations of freeway traffic. *Transportation Research Record, 2249*, 78–85.

Ozbay, K., & Ozguven, E. E. (2007). A comparative methodology for estimating the capacity of a freeway section. *2007 IEEE Intelligent Transportation Systems Conference*, 1034–1039.

Pakhira, M. K., Bandyopadhyay, S., & Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition, 37(3)*, 487–501.

Persaud, B., Yagar, S., & Brownlee, R. (1998). Exploration of the breakdown phenomenon in freeway traffic. *Transportation Research Record, 1634,* 64–69.

Sandra A. K., & Sarkar, A. K. (2013). Clustering of pavement stretches and determining optimum number of clusters for pavement maintenance. In: Proceedings of the Transportation Research Board 92nd Annual Meeting (DVD), Washington, D. C.

Shawky, M., & Nakamura, H. (2007). Characteristics of breakdown phenomenon in merging sections of urban expressways in Japan. *Transportation Research Record, 2012*, 11–19.

Sun, L., & Zhou, J. (2005). Development of multiregime speed – density relationships by cluster analysis. *Transportation Research Record, 65*, 64–71.

Tan, P.-N., Steinbach, M. & Vipin Kumar, V. (2007). *Introduction to data mining*. Pearson Education, India.

Thankappan, A., & Vanajakshi, L. (2012). Development of optimized traffic stream models under heterogeneous traffic conditions. In: Proceedings of the Transportation Research Board 91st Annual Meeting (DVD), Washington D.C.

Transportation Research Board. (2000). *Highway Capacity Manual*. Washington D.C.

Xia, J., & Chen, M. (2007). A nested clustering technique for freeway operating condition classification. *Computer-Aided Civil and Infrastructure Engineering, 22(6),* 430–437.

Xia, J., Huang, W., & Guo, J. (2012). A clustering approach to online freeway traffic state identification using ITS data. *KSCE Journal of Civil Engineering, 16(3),* 426–432.

Žalik, K. R., & Žalik, B. (2011). Validity index for clusters of different sizes and densities. *Pattern Recognition Letters, 32(2)*, 221–234.