Hindawi
Journal of Advanced Transportation
Volume 2019, Article ID 1628417, 12 pages
https://doi.org/10.1155/2019/1628417

Review Article

Pattern Recognition Using Clustering Analysis to Support Transportation System Management, Operations, and Modeling

Rajib Saha, Mosammat Tahnin Tariq, Mohammed Hadi, and Yan Xiao

Department of Civil and Environment Engineering, Florida International University, 10555 West Flagler Street, EC 3678, Miami, FL 33174, USA

Correspondence should be addressed to Rajib C. Saha; rsaha005@fiu.edu

Received 9 January 2019; Accepted 29 October 2019; Published 30 December 2019

Academic Editor: Jose E. Naranjo

There has been an increasing interest in recent years in using clustering analysis for the identification of traffic patterns that are representative of traffic conditions in support of transportation system operations and management (TSMO); integrated corridor management; and analysis, modeling, and simulation (AMS). However, there has been limited information to support agencies in their selection of the most appropriate clustering technique(s), associated parameters, the optimal number of clusters, clustering result analysis, and selecting observations that are representative of each cluster. This paper investigates and compares the use of a number of existing clustering methods for traffic pattern identifications, considering the above. These methods include the K-means, K-prototypes, K-medoids, four variations of the Hierarchical method, and the combination of Principal Component Analysis for mixed data (PCAmix) with K-means. Among these methods, the K-prototypes and K-means with PCs produced the best results. The paper then provides recommendations regarding conducting and utilizing the results of clustering analysis.

1. Introduction

For a long time, decisions associated with transportation systems planning and planning for operations have been based on a limited amount of data collected for a few days. This data was used to obtain information that is supposed to represent the transportation system conditions for the whole year. Thus, the assessment of management and operation strategies; signal control optimization; and the use of analysis, modeling, and simulation (AMS) has been limited in most cases to one scenario of transportation system conditions.

With the advancements of transportation system management and operations programs and the associated intelligent transportation system technology deployments, quantitative and detailed traffic, and event data have become available from multiple sources that allow better identification of system performance and assessment of improvement alternatives under different congestion, incident, and weather scenarios utilizing data analytics and advanced AMS tools. This is particularly important in the assessments of alternative strategies since the main benefit of these strategies is their ability to adapt to different system conditions. The Federal Highway Administration (FHWA) realizing this need has updated their guidance for utilizing AMS to include clustering analysis to identify operation scenarios as an important component of AMS. Decision support systems developed to support transportation system operation decisions also require the identification of traffic patterns, for which response plans are developed for potential real-time activations.

There has been an increasing interest in using clustering analysis for the identification of traffic patterns that are representatives of traffic conditions in support of the above mentioned applications. However, there has been limited information to support agencies in their selection of the most appropriate clustering technique(s), clustering result analysis, and selecting observations that are representatives of each cluster. A large proportion of the conducting clustering analysis to support the applications mentioned above have utilized the K-means clustering method. This method is also widely used in other disciplines and is very efficient in analyzing large datasets. However, its application is limited to datasets with only quantitative variable as it utilizes the Euclidian distance as the dissimilarity matrix [1]. Some of the contributing factors to traffic patterns are categorical variables or are usually

converted to categorical variables before use. Thus, the use of the K-means method with these variables considered is not appropriate. There is a clear need to review existing clustering methods and provide recommendations regarding their applicability and performance as related to various applications.

The goal of the study is to support transportation agencies in their selection of a clustering technique and associated parameters for identifying operational scenarios. This paper investigates and demonstrates the use of a number of existing clustering methods for traffic pattern identifications. These methods include the K-means, K-prototypes, K-medoids, four variations of the Hierarchical method, and the combination of Principal Component Analysis (PCA) for mixed data (PCAmix) with K-means. In providing this investigation, this paper aimed to motivate agencies and researchers in transportation engineering to explore and understand various available clustering methods and apply the methods that are most suitable and perform best for their applications.

2. Review of Clustering Applications

Clustering analysis is an unsupervised learning technique and refers to a grouping or segmenting technique applied to a collection of objects to subgroup them in a way where the objects within a cluster are closely related compared to objects in different clusters [2]. Clustering methods usually utilize a dissimilarity measure to cluster the objects. Although clustering analyses have been used for a long time in other disciplines, the use of the approach in the transportation engineering field has been limited. However, there has been an increasing interest in this use in recent years due to the increasing availability of detailed data and the identified needs for scenario identification for AMS and decision support system applications, as mentioned earlier.

Xia and Chen [3] used K-means clustering to identify the traffic flow phases based on traffic density and speed data aggregated in 15 minutes. The authors also used a nested clustering technique, where each cluster at a level is further subclustered to classify the operating conditions of the freeway into several tiers [4]. Park [5] found that several clustering methods such as the K-means and Fuzzy clustering were effective in traffic volume forecasting. Chen et al. [6] used the K-means clustering along with Davies-Bouldin Index and Silhouette Coefficient to capture the distinct groups in the vehicle temporal and spatial travel behaviors using license plate recognition data. Fuzzy C-means clustering, a probability-based clustering was found successful in recognizing congestion patterns on urban roads based on GPS trajectory [7]. Spectral clustering, a method that allows clustering using fewer dimensions, was used to analyze the traffic state variation based on the quantitative speed data [8]. Other studies that used clustering include Oh, Tok, & Ritchie [9] and Alvarez & Hadi [10].

The most extensive example of the utilization of clustering analysis in transportation engineering is its use in the AMS testbed effort funded by the FHWA [11–13]. As this effort involved six testbeds to pilot test the use of AMS for the evaluation of advanced strategies. These are the San Mateo (US 101), Pasadena, Dallas, San Diego, Phoenix, and Chicago testbeds. The effort emphasized the importance of multiscenario assessments that involve modeling days with different traffic patterns rather than an average day. The testbeds used clustering analysis to identify the traffic patterns based on measurements such as volume, speed/travel time, and event data (e.g., incident and weather).

In the Dallas AMS Testbed [14]; four operational conditions were identified for each period using the K-means clustering based on vehicle miles traveled (VMT), travel time, incident severity, and precipitation. It is interesting to note that the study converted the quantitative variables to categorical variables before using them in clustering. For example the VMT was categorized into three levels and the precipitation was categorized into wet and dry.

The San Diego Testbed [15] used incident duration, demand, travel time, and incident impact on delay in the clustering. After applying the K-means clustering, four operation scenarios were selected for each peak period.

The Pasadena Testbed [16] also utilized the K-means clustering based on VMT, travel time, the total number of incidents, total duration of incidents, and precipitation to identify three operation scenarios for the AMS during the weekday peak periods.

The Phoenix Testbed [17] used the hourly traffic count, hourly travel speed, hourly precipitation, hourly incident frequency, hourly traffic counts, and travel speed in clustering. The Hierarchical clustering method was applied first to find the minimum number of clusters. The K-means clustering was then used for determining four traffic patterns in each period. Although Hierarchical clustering itself is applicable for finding the traffic patterns, the analysis team did not explain the rationale of using the K-means after utilizing the Hierarchical clustering. Since all four patterns selected using clustering represent dry days, a fifth pattern representing a rainy day was selected for the analysis using an additional K-means clustering.

The Chicago Testbed [18] utilized a two-step joint K-means clustering procedure. In the first step, weather patterns were identified based on the precipitation type (Rain, Snow, and Clear) and intensity using the K-means algorithm. In the second step, the K-means algorithm based on traffic data was used to identify subpatterns under each weather condition. This was done because weather impacts was a main focus of the testbed.

The San Mateo Testbed [19] used K-means clustering analysis based on travel time, VMT, weather, and incident frequency (categorized by three duration categories). Five clusters representing five operational conditions were recommended.

3. Review of Clustering Methods

As indicated above, the K-means clustering was used in most reviewed applications in transportation engineering. There are several other clustering methods available; the methods can be classified under four major approaches: the centroid based methods, hierarchical clustering, distribution-based clustering, and density-based clustering as shown in Figure 1 [20].

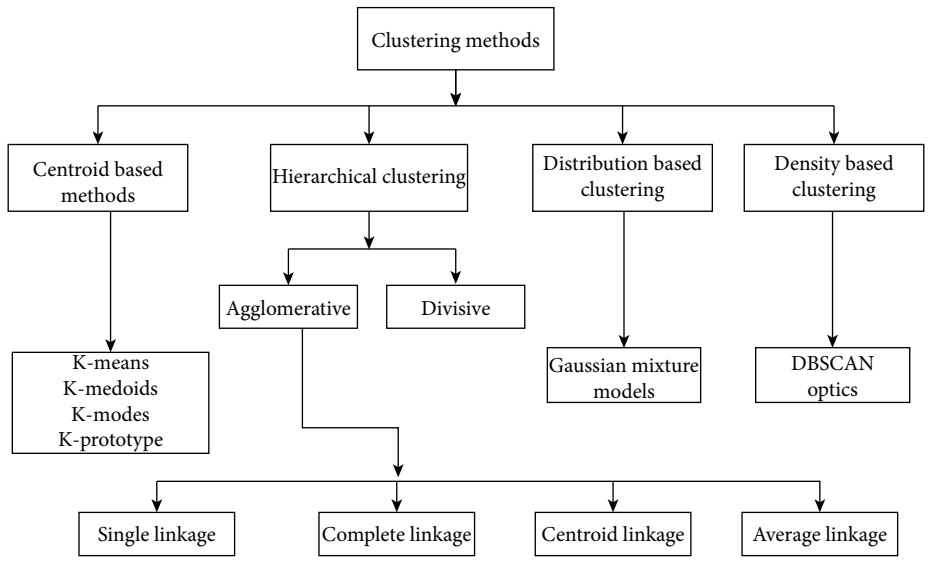

Figure 1: Different types of clustering methods.

Figure 1 also shows examples of clustering methods for the four major approaches. This section presents a brief review of methods that are relevant to the study since they are important to be reviewed to determine their performance when conducting clustering.

3.1. K-means. The K-means algorithm is a widely used method that is applicable for clustering data based on quantitative variables [21]. The method is based on an iterative algorithm in which the process is initiated by providing a fixed set of centroids [22]. Each data point to be clustered is then assigned to its closest centroid using a squared Euclidian distance measure [23]. To assign a point to a cluster, the goal is to minimize the sum of average pair-wise distance within-cluster dissimilarity. The centroids are then updated by computing the average of all the points assigned to each cluster. These steps are iterated until the assignment of the data points to each centroid does not change significantly. This method is efficient to analyze large datasets however, its application is limited to clustering based on the quantitative variable as it utilizes the Euclidian distance as the dissimilarity matrix [1].

3.2. K-prototypes. To perform clustering analysis on datasets that contain both categorical and quantitative data, a method referred to as the "K-prototypes" was proposed [1]. The K-prototypes algorithm works in a similar fashion to the K-means algorithm but applies a combined dissimilarity measure. For quantitative variables, it uses Euclidean distance while for categorical variables, it uses a simple matching dissimilarity [1].

3.3. K-medoids. The K-medoids clustering algorithm is very similar to the K-means algorithm, except that it uses dissimilarity measures to allow clustering based on both quantitative and categorical variables [2, 24–26]. In addition, it has been reported that the squared Euclidean distance measure used in the K-means lacks robustness against outliers that produce very large distances [2]. K-medoids overcomes this issue at the expense of computational efficiency, by placing the outlier into separate clusters.

3.4. Hierarchical Clustering. Hierarchical clustering does not require the specification of the number of clusters initially as required by the K-means and K-medoids. However, it requires the user to specify a dissimilarity measure between groups of observations known as "linkage" [2]. In this method, the clusters at each level of the hierarchy are created by merging clusters at the next lower level. At the lowest level, each cluster contains a single observation while at the highest level there is only one cluster containing all observations. Commonly used linkages are Single, Complete, Average, and Centroid. Respectively, these four linkages consider the dissimilarity between two groups as the smallest dissimilarity between two points in the groups, largest dissimilarity between two points in opposite groups, average dissimilarity over all pairs in opposite groups, and the dissimilarity between the centroids of the groups. The Euclidean distance is generally used as the dissimilarity measure for quantitative variables, while other dissimilarity measures such as Gower metric is used for other variables.

3.5. Principal Component Analysis (PCA) Combined with Clustering. PCA is a statistical approach for dimension reduction and compression while retaining most of the variation in the data set [27]. The purpose of PCA is to convert the observations to an orthogonal system of Euclidean space and thus reduce the dimensionality by retaining only those characteristics of the data set that contributes most of its variance. PCA is effective in reducing the noise in the data set, in addition to reducing the computational cost by reducing the dimensions. In particular, PCA was found effective in capturing the cluster structure in the data set when used along with clustering methods instead of clustering methods by themselves [28]. Ding and He [29] found that K-means clustering on high dimension data was affected by the noise

in the data set and applying K-means clustering in the PCA subspace improved the results significantly. However, the applicability of PCA is limited to quantitative variables. PCAmix is an extended PCA approach for mixed data set combining quantitative and categorical variables [30, 31] and will be investigated in this study for use in combination with clustering. Clustering with the reduced dimensions from PCA was found very effective in the recognition of the patterns in the data set and widely used approach in other fields [32–34]. Because of the mix of quantitative and categorical variables, the PCAmix approach instead of the basic PCA approach was used in the analysis.

4. Utilized Data

To investigate the performance of different clustering methods in identifying traffic patterns, data was retrieved for a corridor of about 16-miles of the I-95 freeway in Fort Lauderdale, Florida. This facility is one of the busiest and strategically important routes in the South [35]. The analysis horizon is the time period from January 1st, 2017 to December 31st, 2017 excluding holidays and weekends. Traffic data including volume, speed, and occupancy was collected from five microwave detectors placed on average at a half-mile interval along the corridor. The data was retrieved from the regional data warehouse, which is a part of the Regional Integrated Transportation Information System (RITIS) [36] in 15-minute intervals for the morning (AM) peak period (7:00 AM–9:30 AM) in the Southbound Direction of I-95.

Incident data for the analysis horizon was retrieved from the incident management database managed by the Florida Department of Transportation (FDOT) District 4. incident data from several sources such as data sharing with the police, automatic detection based on microwave sensors and verification based on closed-circuit television cameras, and service patrol reports. The collected incident data is very detailed and includes several useful attributes, including the start and end times, lane blockage duration, total incident clearance time, number of blocked lanes, severity, time stamps of emergency vehicle arrivals, number of vehicles involved in the incident, and so on. Weather data was collected from the National Centers for Environmental Information–National Oceanic and Atmospheric Administration website [37]. The data was collected from the Pompano Beach Airpark weather station, which is within a 10-mile radius of the study corridor. The weather station measures the precipitation using an 8-inch gauge that is of standardized design used throughout the world for official rainfall measurements. The data set includes the hourly precipitation (in inches) for each 15 min observations. All three types of data (traffic, incident, and weather) were converted to 15minutes resolution and assembled for clustering analysis.

5. Analysis Methodology

This section presents the methodology utilized in this study. The method consists of data retrieval, data preparation, application of the clustering algorithms, performance assessment of the clustering algorithms, and operational scenarios selection. Figure 2 shows the different steps of the methodology, which are explained in detail in the subsequent sections. The statistical package "R-Studio" was used for data assembly, as well as clustering and PCAmix analysis.

5.1. Utilized Variables. Six variables: volume, speed, occupancy, travel lane blockage due to incidents, incident severity, and precipitation estimated based on the retrieved data that were selected for potential utilization in the clustering since these variables are considered to cover various influencing factors on congestion. Similar variables were considered in the FHWA AMS testbeds discussed earlier. There was no construction work on the corridor in the study period. The clustering was initially done using measures calculated based on the average of detector data for all locations of the corridor. However, it was found that better results can be obtained if the detector measurements at each of the five detection locations are used individually in the clustering. Prior to analysis, the precipitation data was categorized in the same manner utilized in the Chicago AMS testbed into four groups as follows: 0 (0 inch/hr), 1 (0–0.1 inch/hr), 2 (0.1–0.3 inch/hr), and 3 (≥0.3 inch/hr). In the analysis, volume, speed, and occupancy were used as quantitative variables, while travel lane blockage, incident severity, and precipitation were used as categorical variables. Thus, the most widely used clustering technique (K-means) cannot be used directly due to the mix of these two types of variables [21]. Instead, this study examines and compares the results from a number of clustering approaches.

Initial clustering without normalization in this study indicated the need for normalization of the variables to a common scale since without it, the clustering was dominated by the variables that have higher magnitudes such as volume. Several methods of normalization have been proposed in the literature including the Min-Max, Z-score, Decimal Scaling, among others. In the study, the variables were normalized using Min-Max normalization shown in Equation (1) as the method performs well to bring all the data within a common scale [37].

$$X' = \frac{X - MinX}{MaxX - MinX}, \tag{1}$$

where, X' = normalized value,
X = attribute value,
$MinX$ = lowest value of the attribute,
$MaxX$ = highest value of the attribute.

5.2. Determining the Number of Clusters. One of the challenges with clustering is to determine an adequate number of clusters to be able to identify all frequent patterns. Widely used clustering methods such as the K-means and K-medoids require the specification of the number of clusters. Some of the clustering methods such as Hierarchal clustering can automatically recommend the number of clusters. There are several empirical methods available to identify the required number of clusters based on the results of clustering analysis, such as the Elbow method, average Silhouette method, and Gap Statistics method, among others. In this study, the process

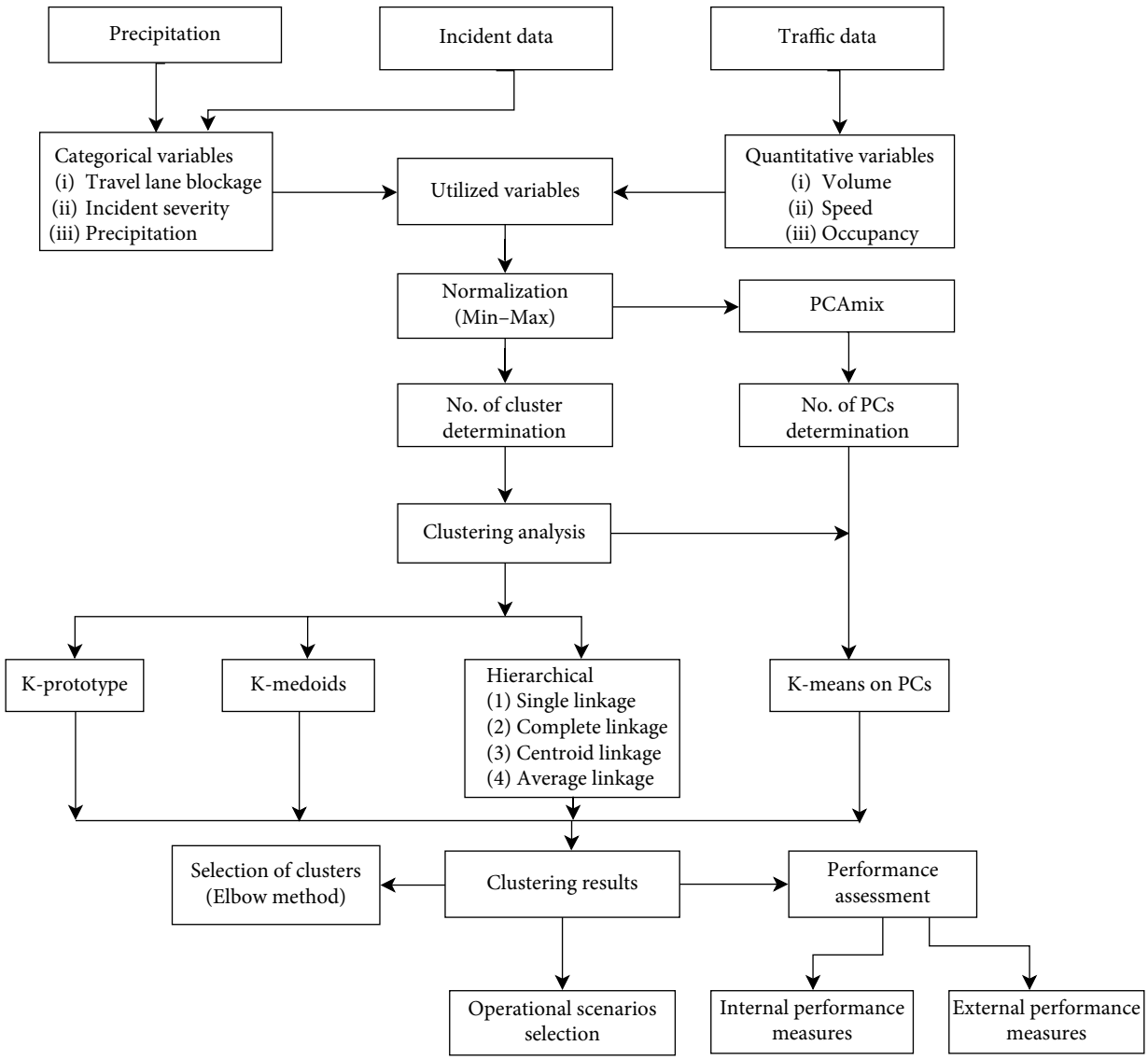

Figure 2: Flow chart of the analysis methodology.

of identifying the number of clusters starts with specifying a maximum of 20 clusters, and the optimal number of clusters were selected based on the Elbow method, which is explained briefly below.

The Elbow method is an empirical method that provides an objective approach to determine the optimal number of clusters. The method requires minimal prior knowledge about the dataset and the attributes of the dataset. The Elbow method determines the number of clusters based on the total within-cluster sum of square (WSS) for each investigated number of clusters [38]. A graph is drawn between the total WSS, and the number of clusters and the location of the bend in the plot is considered as an indicator of the appropriate number of cluster. In this study, the number of clusters using this method was determined, and the results were examined to determine if this method is adequate or additional clusters are needed to identify all patterns of interest, will be discussed when presenting the results of clustering in this paper.

5.3. Conducting Clustering Analysis. A number of methods for clustering were examined, and their results were compared to determine how well they can cluster the traffic conditions into different patterns. The examined methods include the K-prototypes, K-medoids, and Hierarchical clustering with different linkage types (single, complete, centroid, and average linkage) using the optimal number of clusters identified based on the Elbow method.

The PCAmix approach, combined with K-means clustering was also investigated. PCA is a technique to reduce the data dimensionality by geometrically projecting onto lower dimensions called the principal components (PCs), which are defined as a linear combination of the data's original variables [39]. The first PC is identified by minimizing the total distance between the data and their projection onto the PC while retaining the maximum variance of the projected points. Similarly, all other PCs are formed based on the same condition in addition to no correlation among different PCs. PCs try to retain

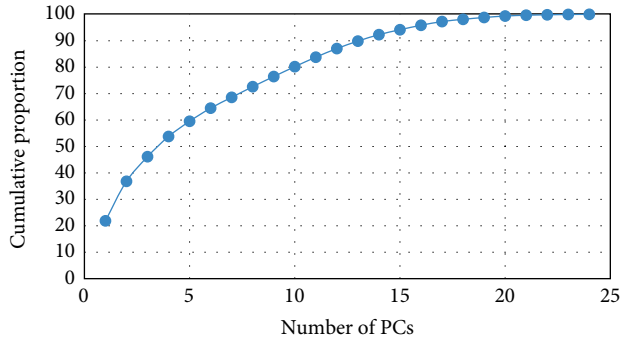

Figure 3: Proportion of variance of data explained by principal components (PCs).

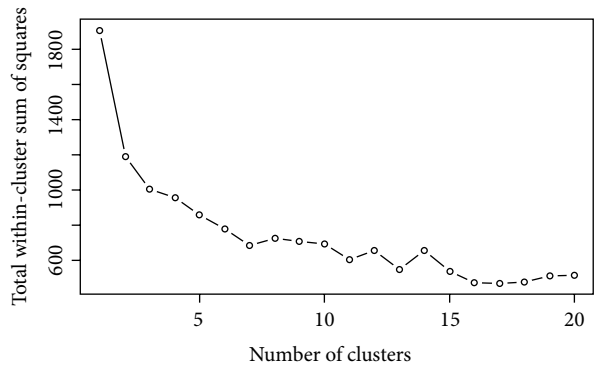

Figure 4: The relationship between the number of clusters and the total within-cluster sum of squares.

the maximum variation within the dataset with a small number of PCs capable of explaining the entire dataset. Six variables: volume, speed, occupancy, travel lane blockage due to incidents, incident severity, and precipitation from five detectors are used in the study. Thus, each PC is a linear combination of the six variables from all five detectors used in the analysis. The optimal number of PCs were determined based on the plot of the cumulative proportion of variance in the data explained by the utilized number of PCs. Finally, the K-means method is applied for clustering. This is possible since the resulting PCs from the PCAmix are quantitative variables, allowing the use of the Euclidean distance as a dissimilarity measure in the K-means clustering. The subset of PCs to be used in the clustering was determined by plotting the cumulative proportion of explained variance in the dataset against the number of PCs for the project case study, as shown in Figure 3. The data was projected into a total of twenty-five PCs through the application of the PCAmix algorithm while reporting the proportion of the explained variance by each PC. Based on this figure, a subset of ten PCs were selected for use in the analysis since it can explain a large proportion of the variation of the data and is able to retain the distinctive features of the data set as well. Although more PCs can be considered in the analysis, the increase in the number of PCs can eliminate the advantages of the use of PCA [29]. Besides, a smaller number of PCs in the analysis reduces the computational cost and removes the noise in the data.

5.4. Assessment of Clustering Method Performance. The performance of the investigated clustering methods were assessed utilizing the external and internal performance measures. External measures evaluate the purity of the clusters [40] while the internal measures evaluate the compactness of a clustering structure by determining how close the attributes of each data point are without considering additional information about the data [41]. As clustering is an unsupervised technique, there is no ground truth data associated with this technique to compare to. Thus, assessment based on quantitative external performance measures was not possible based on ground truth data. However, all data within clusters were visualized to evaluate the distribution of the data among all the clusters.

Unlike the external performance measure described above, Silhouette coefficient and connectivity were two internal performance measures chosen in the study for their capability to assess the performance of the clustering algorithms. These measures do not need ground truth data and allow easy interpretation of the results [42]. A higher Silhouette coefficient depicts a dense and well-separated cluster. A lower connectivity coefficient depicts a higher degree of connectedness of the clusters [43].

6. Clustering Results

6.1. Number of Cluster Selection. The number of clusters were selected by applying the Elbow method, as mentioned earlier. For this purpose, the total within-cluster sum of square (WSS) was plotted against the number of clusters after the initial runs of the K-prototypes clustering, as shown in Figure 4. The figure indicates that seven clusters are the optimal number of clusters based on the location of the kink in the elbow of the plots. the optimal number of clusters depends on the attributes of the dataset, and it can vary between locations. If sufficient data is used, then the number of clusters can be fixed for the analyzed location. This is the case for the case study since the data used in the study is for an entire year; therefore, it automatically considered the variation of traffic for season, weather, incident, and so on. It should be mentioned here that this study is concerned with the developing of the methodology rather than finding the optimal number of clusters. Seven clusters were found to produce good relative WSS for the other method. To allow fair comparison, the same number of cluster was used for all methods.

6.2. Evaluation of the Clustering Methods Based on Internal Measures. As stated earlier, the two utilized internal measures to assess the methods performance were the average Silhouette coefficient and connectivity. The comparison of the Silhouette coefficient of the investigated methods in Figure 5 indicates the superiority of the K-means with PCs over other clustering methods used in the study since this method produced the higher value of this coefficient. The K-medoids performance was the worst among the four tested algorithms. The assessment based on the connectivity measure, as depicted in Figure 6, shows a similar result with the best performance achieved with the use of K-means with PCs, as this produced the lowest value of the connectivity measure.

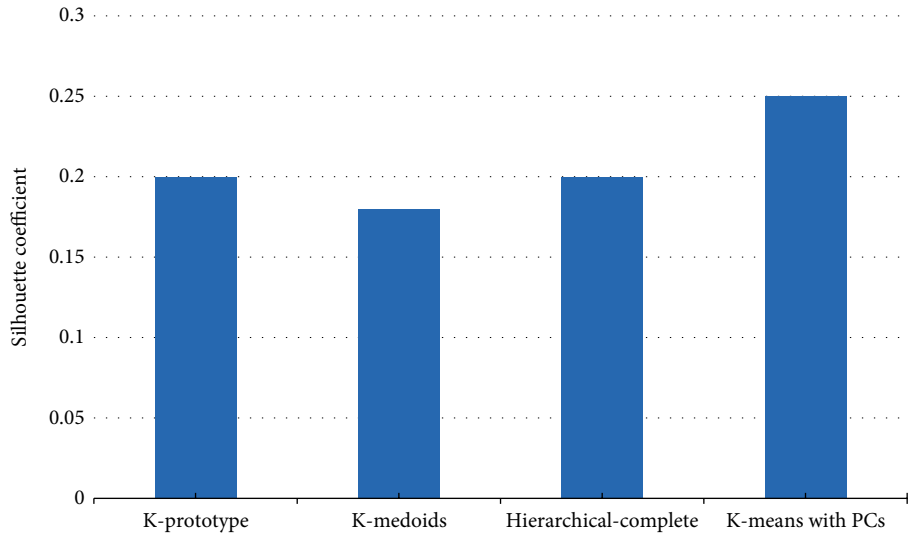

Figure 5: Silhouette coefficient for different clustering methods for the seven clusters.

6.3. Evaluation of Clustering Methods Based on External Measures. External measure assessment of the clustering methods evaluates the purity of each cluster. The distribution of the average volume, speed, and occupancy on all detectors (as indicated by the 10^{th}, 50^{th}, and 90^{th} percentiles of these variables); incident severity variation; and precipitation variation within each cluster were examined to be sure that the method is able to separate the traffic into distinctive patterns based on the clustering variables. Such separation will allow the analysis and modeling of these patterns.

Examining the results from the K-medoids indicated that the algorithm was not able to discern between the incident and non-incident observations as well as precipitation and non-precipitation observations. As an example, Figure 7 shows that the algorithm places incident and non-incident observations into the same clusters (Clusters 1, 3, 4, 5, 6, and 7; as shown in Figure 7). This is possible because the Gower dissimilarity metric used with the K-medoids was dominated by the quantitative variables (volume, speed, and occupancy) at the expense of the categorical variables (incident and precipitation attributes).

When using the Hierarchical clustering, it was found that the "Complete" linkage produced the best results among the four investigated linkages. The remaining three linkages grouped almost all of the observations into a single cluster irrespective of the differences in the attributes associated with different period patterns. Although the Complete linkage performed a little bit better than the other three linkage types, it still had the problem of assigning a big proportion of patterns to one cluster. The complete linkage grouped 81% of observation in one cluster while distributing the remaining 19% of the observations between the other six clusters, as shown in Figure 8. Based on the above, it can be concluded that the Hierarchical method was not able to produce good results.

Unlike the K-medoids and Hierarchical clustering, both the K-prototypes and K-means with PCs produced distinct clusters that separate different congestion levels and the causes of congestion (incidents, "recurrent conditions", and rainy conditions). The observations in each clusters produced by both methods are shown in Figures 9 and 10 respectively. The K-prototype produced three clusters with recurrent condition observations (Clusters 1, 5, and 6), three clusters (Clusters 2, 3, and 4) with incidents that have different levels of severity/lane blockages, and one cluster (Cluster 7) with combined incident and rain events. In contrast, the K-means with PCs produced two clusters (Clusters 3 and 5) with recurrent conditions, two clusters (Clusters 2 and 4) with incidents that have different levels of severity/lane blockages, and one cluster (Cluster 7) with incident and rain conditions. The K-means with PCs produced two more clusters which represent very unusual conditions. Cluster 6 contains only four Severity Level 3 incidents with all lanes blockage. The other cluster (Cluster 1) includes observations during special conditions such as days during hurricane preparation and Black Friday. Both the K-prototypes and K-means with PCs produced good clustering of the traffic patterns for the investigated case study. With both methods, the observations were clustered in three distinctive groups: normal (recurrent condition) clusters, incident clusters, and precipitation (rainy conditions) that may include incident clusters. It is important for the analyst to examine the clustering results in more detail to determine what each cluster actually represents. Further comparison of the clusters produced by the two methods is presented below.

6.4. Rainy Day Cluster. Rainy day cluster is one of the distinctive clusters produced by both the K-prototypes and K-means with PCs methods that include observations with increased congestion due to precipitation in some cases combined with incidents. Although not automatically separated into two clusters, the analyst may want to separate the precipitation observations into two groups for the analysis: precipitation observations with no incidents and precipitation observations with incidents; depending on the purpose of the study.

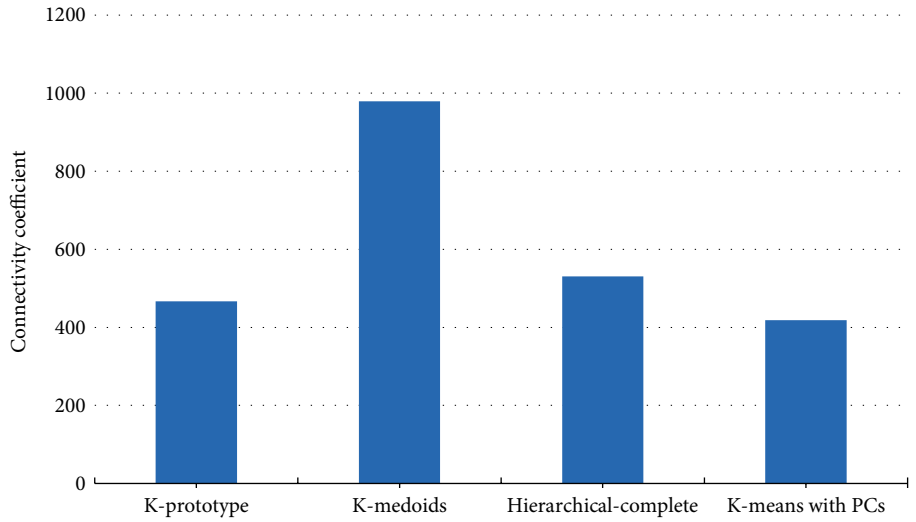

Figure 6: Connectivity coefficient for different clustering methods for the seven clusters.

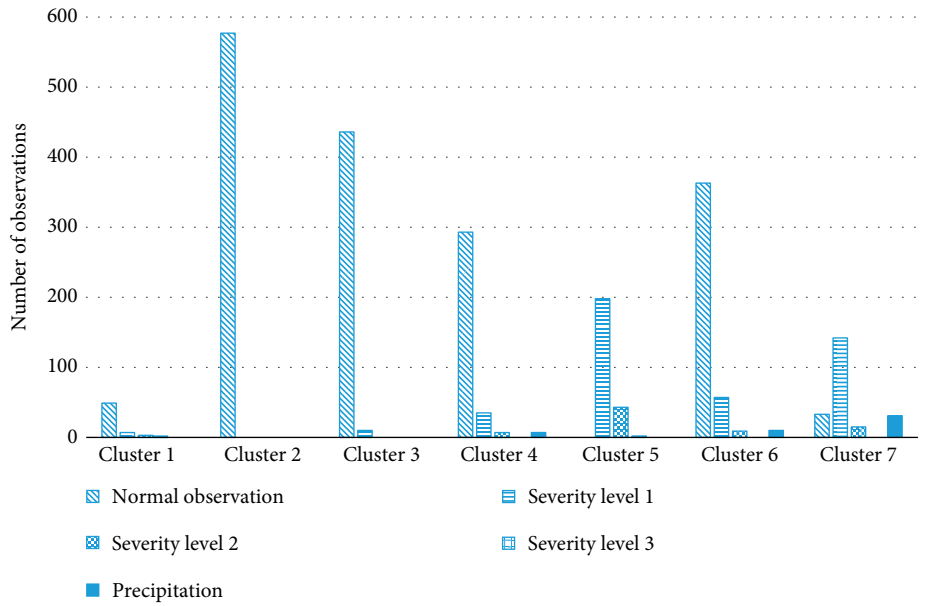

Figure 7: Observations with different levels of severity and precipitation in the clusters identified by the K-medoids method.

6.5. Incident Clusters. Observations with incidents were clustered in three clusters in the case of the K-prototypes method and two clusters in the case of the K-means with PCs method. To determine if it is justifiable to have three clusters vs. two clusters, the impacts of these incidents on traffic, the locations of incidents, and the time of occurrence of incidents within each cluster were examined. The box plots in Figure 11 show the 10^{th}, 50^{th}, and 90^{th} percentile values of the speed and occupancy for each incident cluster identified by each of the two methods. Incident statistics of each cluster reveals that one cluster produced by each method has high incident impacts with more incidents occurring in the middle segment of the facility, which is the most congested segment, and between 7:30 am and 9:00 am, which is the peak congestion interval. The other identified clusters by the two methods have lower incident impacts and have less observations in the peak intervals. It appears that in terms of speed and occupancy, two of the three K-prototypes incident clusters are very similar. Therefore, it seems that the two clusters identified by the K-means with PCs are sufficient to represent incident impacts.

6.6. Normal Clusters. The K-prototypes produced three distinctive clusters representing normal observations. These observations were grouped into only two clusters by the K-means with PC. The box plots in Figure 12 show the 10^{th}, 50^{th}, and 90^{th} percentile values of the speed and occupancy for each of the normal cluster identified by the two methods. The K-prototypes seems to focus more on the quantitative traffic flow parameter variations and this is the reason that it splits the normal observations into more clusters based on

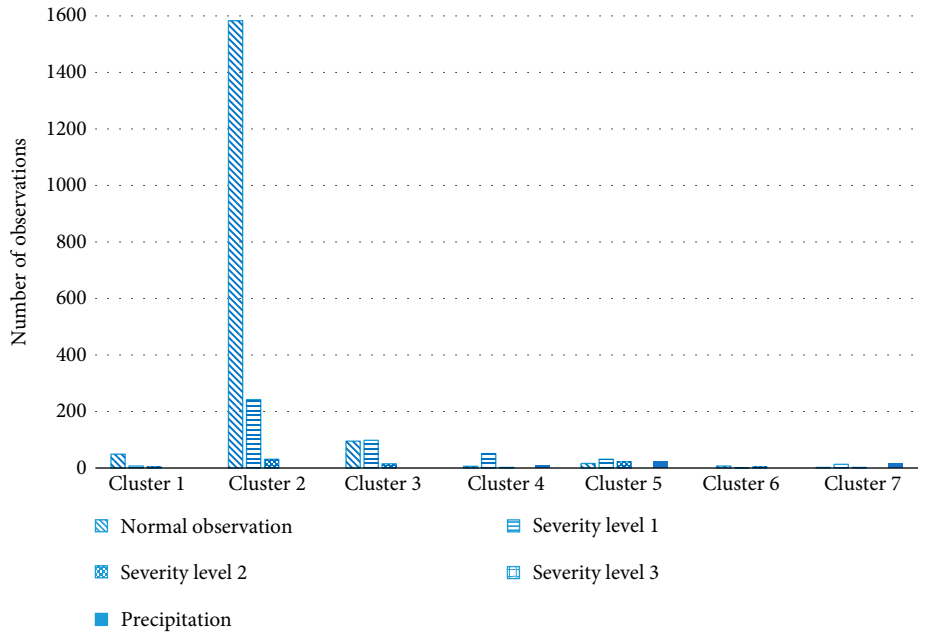

Figure 8: Observations in each cluster when using hierarchical clustering with the complete linkage.

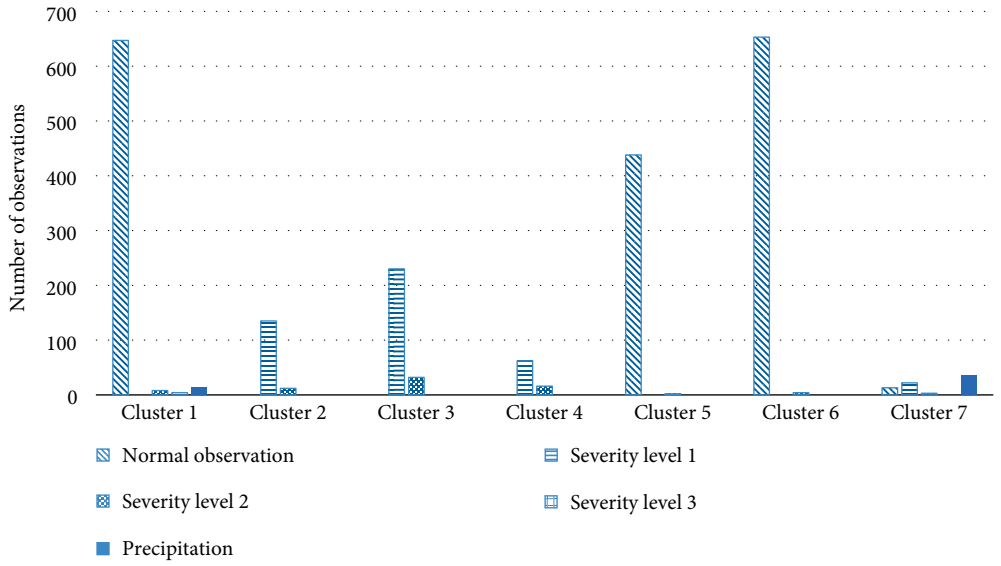

Figure 9: Observations with different level of severity and precipitation in the clusters identified by the K-prototype method.

these parameters. As shown in this figure, the three clusters identified by the K-prototypes divide the normal days into three levels—with a median occupancy of 9.5%, 11.5%, and 15.5%, respectively and a median speed of 44 mph, 57 mph, and 65 mph. The medians of the two clusters K-means with PC were 10% and 14% for occupancy and 50 mph and 64 mph. It appears that the two clusters are sufficient to cover the normal day conditions. However, if further clustering is needed for normal observations, a second level of clustering can be conducted on the normal observations only based on the quantitative traffic flow parameters (volume, speed, and occupancy).

6.7. Selection of Operational Scenarios. Based on examining the results, five operational scenarios were identified for the inclusion in the modeling as of follows:

(1) Normal traffic pattern with high volume, high speed, and low occupancy.

(2) Normal traffic pattern with high volume, low speed, and high occupancy.

(3) Minor incident traffic pattern with high volume, moderate to high speed, and low to medium occupancy.

(4) Major incident traffic pattern with low volume, low speed, and high occupancy.

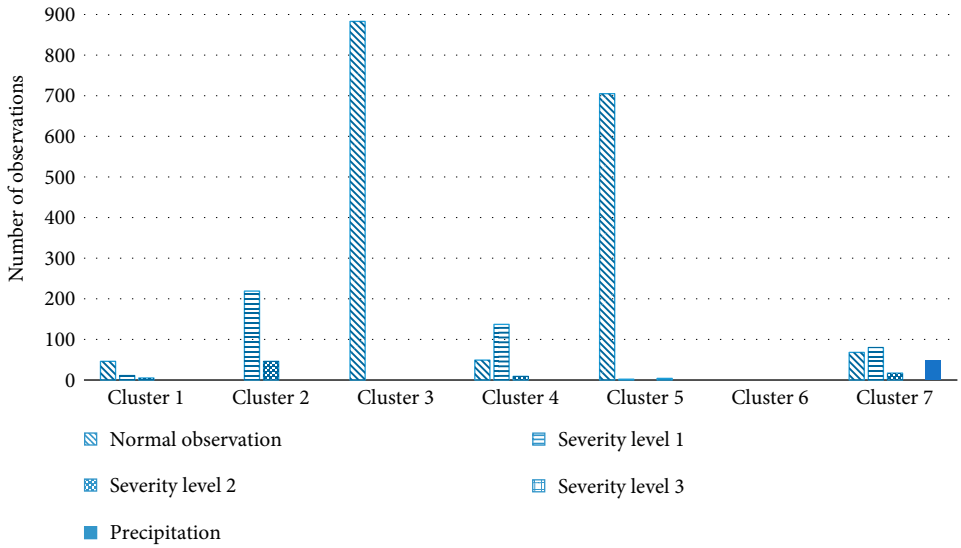

Figure 10: Observations with different levels of severity and precipitation in the clusters identified by the K-means with PCs method.

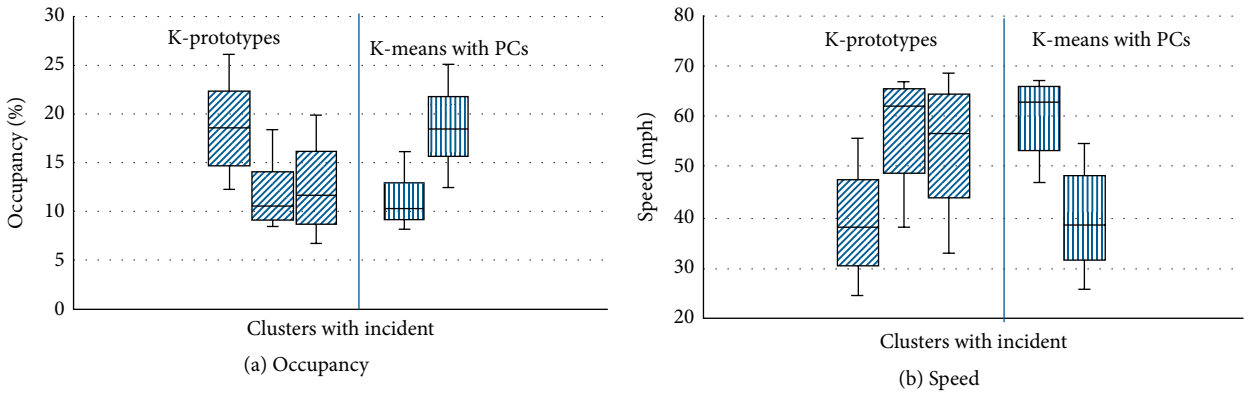

Figure 11: Variation of occupancy and speed within the clusters identified by the clustering methods for incident days.

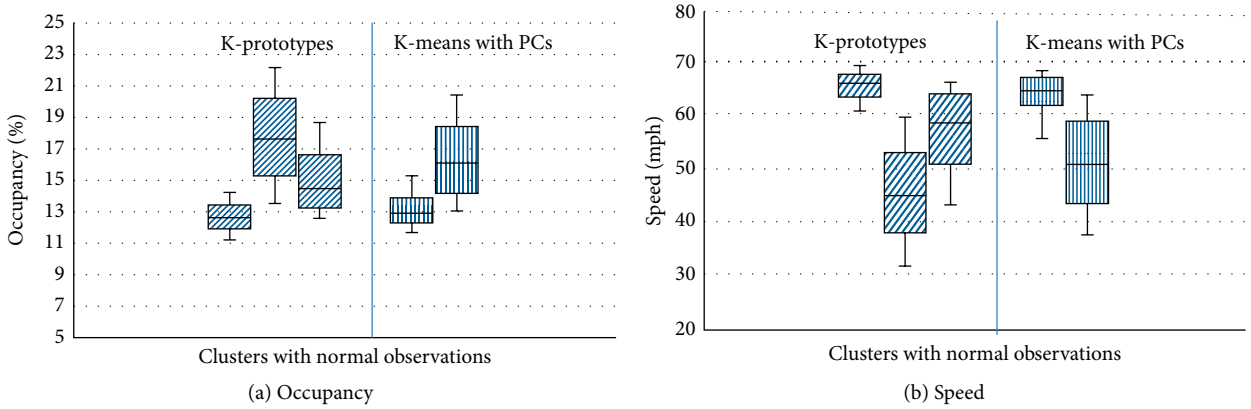

Figure 12: Variation of occupancy and speed within the clusters identified by the clustering methods for normal days.

(5) Precipitation traffic pattern with high volume, low speed, and high occupancy.

7. Conclusion and Recommendation

This paper has investigated and demonstrated the use of a number of existing clustering methods for traffic pattern identifications, considering the above-mentioned issues. These methods include the K-prototypes, K-medoids, four variations of the Hierarchical method, and the combination of Principal Component Analysis for mixed data (PCAmix) with K-means. The K-means method by itself was determined to be not suitable for use in clustering based on categorical variables and thus was dropped from the further comparison.

It was found that the K-medoids was not able to discern between normal observations and incident and rain observations, apparently because the dissimilarity metric used with the K-medoids is dominated by the quantitative variables (volume, speed, and occupancy) on the expense of the categorical variables (incident and precipitation attributes). Hierarchical clustering has the problem of putting most of the data points into a single cluster. The internal measure comparison based on the Silhouette coefficient and connectivity further confirm the inferior performance of the K-medoids and Hierarchical clustering.

The K-prototypes and K-means with PCs produced the best results when utilizing both internal and external measures in the comparison. However, the K-prototypes clustering was not able to distinguish special patterns like the days during hurricane preparation, Black Friday, and full lane closures. On the other hand, it splits the incidents observations and normal observations into three patterns each instead of two each when using the K-means with PCs.

In all cases, the analyst should examine the clustering results to determine if further clustering and/or merging of clusters is needed. Such decisions will have to be based on the purpose of the study. The selection of an observation from each cluster for use in AMS will also have to be carefully done. If there is a large variation in one or more pattern attributes, more than one representative observations may need to the from each cluster. In some cases, when there is a large variation in the attributes within each cluster, two-level clustering for finer traffic pattern identification is recommended.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Funding

The presented work in this paper is a part of the research project funded by the Florida Department of Transportation under grant no. [BDV29 977-38]. The opinions, findings, and conclusions expressed in this publication are those of the author(s) and not necessarily those of the Florida Department of Transportation or the U.S. Department of Transportation.

References

[1] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* , Springer, Stanford, California, 2nd edition, 2017.

[3] J. Xia and M. Chen, "Defining traffic flow phases using intelligent transportation system-generated data," *Journal of Intelligent Transportation Systems*, vol. 11, no. 1, pp. 15–24, 2007.

[4] J. Xia and M. Chen, "A nested clustering technique for freeway operating condition classification," *Computer-Aided Civil and Infrastructure Engineering*, vol. 22, no. 6, pp. 430–437, 2007.

[5] B. Park, "Hybrid neuro-fuzzy application in short-term freeway traffic volume forecasting," *Transportation Research Record: Journal of Transportation Research Board*, vol. 1802, no. 1, pp. 190–196, 2002.Paper No. 02-2921

[6] H. Chen, C. Yang, and X. Xu, "Clustering vehicle temporal and spatial travel behavior using license plate recognition data," *Journal of Advanced Transportation*, vol. 2017, pp. 1–14, 2017.

[7] K. Zhang, D. Sun, S. Shen, and Y. Zhu, "Analyzing spatiotemporal congestion pattern on urban roads based on taxi GPS data," *Journal of Transport and Land Use*, vol. 10, no. 1, pp. 675–694, 2017.

[8] S. Yang, J. Wu, G. Qi, and K. Tian, "Analysis of traffic state variation patterns for urban road network based on spectral clustering," *Advances in Mechanical Engineering*, vol. 9, no. 9, pp. 1–11, 2017.

[9] C. Oh, A. Tok, and S. G. Ritchie, "Real-time freeway level of service using inductive-signature-based vehicle reidentification system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 138–146, 2005.

[10] P. Alvarez and M. Hadi, "Time-variant travel time distributions and reliability metrics and their utility in reliability assessments," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2315, no. 1, pp. 81–88, 2012.

[11] FHWA, *Analysis, Modeling, and Simulation (AMS) Testbed Requirements for Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs*, Federal Highway Administration, U.S. Department of Transportation, 2013

[12] FHWA, *Analysis, Modeling, and Simulation (AMS) Testbed Framework for Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs*, Federal Highway Administration, U.S. Department of Transportation, 2013

[13] M. Vasudevan and K. Wunderlich, *Analysis, Modeling, and Simulation (AMS) Testbed Preliminary Evaluation Plan for Active Transportation and Demand Management (ATDM) Program*, FHWA, US, 2013.

[14] B. Yelchuru, K. Abdelghany, I. Zohdy, S. Singuluri, and R. Kamalanathsharma, *Analysis, Modeling, and Simulation (AMS) Testbed Development and Evaluation to Support Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs-Dallas Testbed Analysis Plan*, FHWA-JPO, US, pp. 16–373, 2016.

[15] B. Yelchuru, M. Juckes, I. Zohdy, and R. Kamalanathsharma, *Analysis, Modeling, and Simulation (AMS) Testbed Development and Evaluation to Support Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management*

(ATDM) Programs-San Diego Testbed Analysis Plan, FHWA, US, 2016.

[16] B. Yelchuru, T. Bauer, D. Roden, and J. Z. Ma, *Analysis, Modeling, and Simulation (AMS) Testbed Development and Evaluation to Support Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs — Pasadena Testbed Analysis Plan*, FHWA, US, 2016.

[17] B. Yelchuru, X. Zhou, P. Mirchandani, P. Li, I. Zohdy, and R. Kamalanathsharma, *Analysis, Modeling, and Simulation (AMS) Testbed Development and Evaluation to Support Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs-Phoenix Testbed Analysis Plan*, FHWA, US, 2016.

[18] B. Yelchuru, H. Mahmassani, and I. Zohdy, *Analysis, Modeling, and Simulation (AMS) Testbed Development and Evaluation to Support Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs-Chicago Testbed Analysis Plan*, FHWA, US, 2016.

[19] B. Yelchuru, B. Nevers, D. Richard, and Z. Ismail, *Analysis, Modeling, and Simulation (AMS) Testbed Development and Evaluation to Support Dynamic Mobility Applications (DMA) and Active Transportation and Demand Management (ATDM) Programs-San Mateo Testbed Analysis Plan*, FHWA, US, 2016.

[20] D. Sarkar, R. Bali, and T. Sharma, *Practical Machine Learning with Python. A Problem-Solvers Guide to Building Real-World Intelligent Systems*, Apress, Berkeley, 2018.

[21] A. K. Jain and R. C. Dubes, *Algorithm for Clustering Data. Prentice-Hall*, ACM, USA, 1988.

[22] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[23] J. A. Hartigan, *Clustering Algorithms*, Wiley, New York, NY, 1975.

[24] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, no. 4, pp. 857–871, 1971.

[25] J. Podani, "Extending Gower's general coefficient of similarity to ordinal characters," *Taxon*, vol. 48, no. 2, pp. 331–340, 1999.

[26] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, US, 2009.

[27] G. H. Dunteman, *Principal Components Analysis*, Sage, US, 1989.

[28] S. Meng, Y. Fu, T. Liu, and Y. Li, "Principal Component Analysis for Clustering Temporomandibular Joint Data," in *Computational Intelligence and Design (ISCID), 8th International Symposium*, IEEE, pp. 422–425, US, 2015.

[29] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the Twenty-First International Conference on Machine Learning*, ACM29, US, 2004.

[30] M. O. Hill and A. J. Smith, "Principal component analysis of taxonomic data with multi-state discrete characters," *Taxon*, vol. 25, no. 2-3, pp. 249–255, 1976.

[31] M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco, "Multivariate analysis of mixed data: the R package PCAmixdata," 2017.

[32] S. Marinai, S. Faini, E. Marino, and G. Soda, "Efficient word retrieval by means of SOM clustering and PCA," *Document Analysis Systems VII*, vol. 3872, pp. 336–347, 2006.

[33] C. Alzate and J. A. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 335–347, 2010.

[34] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[35] Office of highway policy information, "Most travelled urban highways average annual daily traffic," FHWA Department of Transportation, U.S., 2017.

[36] University of Maryland CATT , "Labtransportation system status. Retrieved from RITIS," 2018, http://www.ritis.org/.

[37] NOAA., "National centers for environmental information. retrieved from national oceanic and atmospheric administration," 2018, http://www.ncei.noaa.gov/.

[38] D. J. Ketchen and C. L. Shook, "The application of cluster analysis in strategic management research: an analysis and critique," *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, 1996.

[39] J. Lever, M. Krzywinski, and N. Altman, "Principal component analysis," *Nature Methods*, vol. 14, no. 7, pp. 641–642, 2017.

[40] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for K-means clustering," *Knowledge Discovery and Data Mining KDD*, ACM, US, pp. 877–886, 2009.

[41] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1st edition, 2005.

[42] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, pp. 53–65, 1987.

[43] G. Brock, V. Pihur, and S. Datta, "clValid: an R package for cluster validation," *Journal of Statistical Software*, vol. 25, no. 4, 31740 pages, 2008.

[44] D. Steinley, "K-means clustering: a half-century synthesis," *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 1–34, 2006.