A Nested Clustering Technique for Freeway Operating Condition Classification

Jingxin Xia & Mei Chen

Department of Civil Engineering, University of Kentucky, Lexington, KY 40506-0281, USA

Abstract: This article introduces a nested clustering technique and its application to the analysis of freeway operating condition. A clustering model is developed using the traffic data (flow, speed, occupancy) collected by the detectors and aggregated to 5-minute increments. An optimum fit of the statistical characteristics of the data set is provided by the model based on the Bayesian Information Criterion and the ratio of changes in dispersion measurement. This technique is flexible in determining the number of clusters based on the statistical characteristics of the data. Tests on multiple sites with varying operating conditions have attested to its effectiveness as a data mining tool for the analysis of freeway operating condition.

1 INTRODUCTION

In practice, freeway operating condition is usually qualitatively categorized into six levels of service (A–F), as defined in the Highway Capacity Manual (TRB, 2000). For any given freeway segment, adjustments are made to the baseline condition to account for the impact of a pre-defined list of site-specific geometric and traffic characteristics. Considered as the parameter that best describes the operating quality, density is used as the primary factor to determine the level of service on a freeway.

Some researchers have studied the characteristics of traffic parameters (e.g., flow rate, speed, and occupancy) at various operating conditions and attempted to find ways to distinguish traffic flow phases from each other (Hall et al., 1992; Banks, 1999; Kerner and Rehborn, 1996). Most arguments were made through observing the shape of the speed-flow curve as well as the distribution of the slope values of the time sequence plot in a flow-density domain. Adeli and Ghosh-Dastidar (2004) used a mesoscopic-wavelet model to extract congestion characteristics from traffic data obtained at a freeway work zone.

Xia and Chen (2004) analyzed the traffic flow phases from a macroscopic perspective. A K-means clustering algorithm was used to categorize freeway operating conditions using continuous traffic monitoring data. The algorithm starts with a predetermined number (K) of random clusters and then moves data records among clusters with the objective of minimizing variability within the clusters and maximizing variability between the clusters. This method requires the number of clusters be set before executing the algorithm. One needs to rely on empirical knowledge on individual applications and data sets to determine the number of clusters. For instance, Xia and Chen (2004) categorized freeway operating conditions into five different states based on the discussions in Hall et al. (1992), Banks (1999), and Kerner and Rehborn (1996). Nevertheless, operating conditions may vary significantly at different sites along a freeway; some sites may never experience congestion, and therefore, their data sets may not contain all five flow phases. This greatly limits the efficient application of the K-means clustering algorithm.

This article presents a nested clustering technique for classifying freeway operating condition into different flow phases. The method is advantageous in that it does not require a prior knowledge of the number of clusters. Starting with the set of traffic data obtained from a given site, the number of clusters (i.e., subsets of data with distinct traffic characteristics) that these data points should be grouped into is automatically determined based on the statistical characteristics of the data. This clustering process is then repeated on all subclusters until the dissimilarity between the data points is not significant enough to warrant further grouping. The characteristics
of the resulting clusters and their implications on freeway operating condition analysis are then discussed.

2 METHODOLOGY

The basic idea of the data-clustering model is to group data samples with relatively homogeneous characteristics by essentially minimizing the variance or spread across defining variables of interest within the groups (i.e., clusters) while, at the same time, maximizing the difference between clusters. It is a widely used technique in data mining for uncovering potential patterns in the underlying data. Over the past decades, data clustering has been used in many fields as discussed by Huang (1998), Bacher (2000), Bender et al. (2001), etc. The transportation-related applications include highway traffic volume forecasting (Park, 1998, 2002), crash modeling (Ma and Kockelman, 2005; Sheu, 2002), signal timing design by time-of-day (Hauser et al., 2000), development of real-time level of service criteria (Cheol et al., 2005), and flow phase definition (Xia and Chen, 2004). Most of these studies involved the $K$-means clustering algorithm.

The nested clustering technique developed in this study applies an agglomerative clustering algorithm and uses the Bayesian Information Criterion (BIC) (Schwarz, 1978) to determine the optimum number of clusters. If the attributes of the data used in the analysis are good indicators of the freeway operating condition, the clusters obtained through this procedure shall be statistically distinct from each other (to various extents) in terms of traffic characteristics.

2.1 Agglomerative clustering algorithm

The agglomerative clustering algorithm forms clusters in a bottom-up manner and produces a binary tree with single articles as leaf nodes and a root node containing all of the articles (i.e., the entire data set) (Weng, 1997). The algorithm is initialized by treating each article (e.g., traffic data record) as a separate cluster. The second step is to calculate the distance between any two clusters and obtain a distance matrix. In this study, we use the Euclidean distance as the distance measurement. If $d_{ij}(i, j)$ denotes the Euclidean distance between clusters $i$ and $j$, it can be defined as $d_{ij}(i, j) = (\sum_{a=1}^{A} (c_a(i) - c_a(j))^2)^{1/2}$, where $A$ denotes the total number of attributes, and $c_a(i)$ and $c_a(j)$ denote the centers of the $a$th attribute for clusters $i$ and $j$, respectively. If $n_j$ denotes the number of articles in cluster $j$ and $x_{a,m}(j)$ denotes the numerical value of the $a$th attribute of the $m$th article in cluster $j$, then $c_a(j) = \frac{1}{n_j} \sum_{m=1}^{n_j} x_{a,m}(j)$. Note that the attributes should be standardized prior to the distance calculation to eliminate the effects of the different units if the attributes do not have the same commensurability. In the third step, one should identify, from the matrix, the two clusters with the shortest distance between them among all pairs and merge them. The center of the new cluster needs to be recalculated using the above equation.

The agglomerative clustering algorithm then repeats Step 2 and Step 3 until all articles have been merged into one final cluster, which is the root node of the binary cluster tree.

2.2 Determining the number of clusters

The binary cluster tree provides a range of cluster solutions. Different criteria can be applied to determine which solution is optimum. Solutions range from non-parametric methods such as cross-validation to parametric methods such as the Bayesian Information Criterion (Schwarz, 1978). This article adopts the widely used BIC criterion and a two-stage process presented by Chiu et al. (2001) for this task.

BIC is a likelihood criterion for model comparison that penalizes models with additional complexity (parameters). The BIC for the $J$-cluster solution is defined as

$$BIC(J) = -2 \sum_{j=1}^{J} \xi(j) + 2J \log(N),$$

where $N$ denotes the total number of articles in the data set, $\xi(j)$ is a measurement of the estimated variance of the attribute variables for cluster $j$ defined as

$$\xi(j) = -n_j \left( \sum_{a=1}^{A} \frac{1}{2} \log(\hat{\sigma}_a^2 + \hat{\sigma}_{ja}^2) \right),$$

in which $n_j$ denotes the number of articles in cluster $j$, $\hat{\sigma}_a^2$ denotes the estimated variance of the $a$th attribute for all articles in the data set, and $\hat{\sigma}_{ja}^2$ denotes the estimated variance of the $a$th attribute for those articles in cluster $j$. Theoretically, the smaller the BIC, the better the fit of the clustering model to the data set.

The two-stage process (Chiu et al., 2001) first examines the BIC for all potential clustering solutions as defined by the binary cluster tree. The goal is to find the smallest number of clusters that has the lowest BIC, because the BIC decreases first and then increases as the number of clusters increases. Starting from the root of the cluster tree (when $J = 1$), the BIC for each $J$ (clustering solution) is calculated. Beginning from $J = 1$, the first $J = J$ value that satisfies $BIC(J) < BIC(J + 1)$ is chosen as a coarse estimate of the number of clusters. Particularly, if $BIC(1) < BIC(2)$, the process stops here, and the optimum number of clusters ($J^*$) is set as one.
In the second stage, the ratio of changes in dispersion measurement is used to determine the optimum number of clusters based on the coarse estimate obtained in the first stage. The ratio of changes in dispersion measurement is defined as \( R(J) = \frac{s_{J-1}}{s_{J}} \) for \( J = 2, \ldots, \hat{J} \), in which \( s_{J-1} \) denotes the change in dispersion measurement if \( J \) clusters are merged into \( J - 1 \) clusters. The parameter \( s_{J} \) can be computed as \( s_{J} = I_{J-1} - I_{J} \), in which \( I_{J} = \sum_{j=1}^{J} \xi(j) \). This second stage is based on the understanding that a significant increase in \( R \) will be observed when two clusters that should not be merged are merged. The \( R(J) \) value for each \( J (J = 2, \ldots, \hat{J}) \) is calculated, and the two largest \( R(J) \) values are identified as \( J = J_1 \) (the largest) and \( J = J_2 \) (the second largest). Bacher et al. (2004) suggested an empirical threshold value of \( R(J_1)/R(J_2) = 1.15 \); that is, if \( R(J_1)/R(J_2) > 1.15 \), \( J^* \) is set to \( J_1 \); otherwise, \( J^* \) is set to \( \max(J_1, J_2) \).

Once the optimum number of clusters is determined, the \( K \)-means clustering algorithm is then implemented. The algorithm is initialized with \( J^* \) clusters and the center of each cluster is calculated. An article is assigned to a cluster if the Euclidean distance between the article and the center of the cluster is the shortest among those between the article and centers of all other clusters. The cluster centers are re-calculated at each iteration until the total Euclidean distance between the articles and their corresponding cluster centers converges.

### 3 IMPLEMENTATION

The development of a traffic operation and control strategy often requires an understanding of the operating condition based on the data collected on the roadway. The method described above can determine the optimum clustering model that fits the statistical characteristics of the data. Using the attributes (variables) that capture the most important characteristics of the freeway operating conditions, the algorithm can be used to distinguish freeway operating conditions (flow phase) from one another.

#### 3.1 Data description

Data used in this study was obtained through the California PeMS (Freeway Performance Measurement System). The PeMS archives traffic data collected by thousands of loop detectors in six districts of CalTrans. Traffic data (i.e., flow rate and occupancy) is downloaded from the field detectors every 30 seconds and archived at 5-minute intervals. Additional measures such as average speed are estimated at the 5-minute levels.

Prior to the application of the clustering algorithm, data quality screening was conducted to identify and eliminate suspicious or erroneous data. The set of quality assurance criteria was developed by Lomax et al. (2004) for the mobility monitoring program. These criteria include threshold values of maximum volume and occupancy, multivariate consistency checks, etc. In addition, incident information was obtained through the California Traffic Incident Information System. Data records that might be associated with those incidents were excluded, because highways may operate at reduced capacity before an accident is cleared. Consequently, the data records obtained under such scenarios would not accurately reflect normal operating conditions.

The detector located at mile point 293.42 on northbound SR-99 was chosen as the study site, as shown in Figure 1. There are three lanes going in this direction. The data collected between 15:00 and 23:59 on weekdays from Monday, October 4, 2004, through Friday, October 15, 2004 is used. A total of 1,080 data records were retrieved from the archive for this detector station, during this period. Among these, 214 were identified as suspicious or erroneous, and 25 data records were determined as accident-related through time matching with the incident log. Consequently, a total of 841 data records were used to develop the clustering model for flow phase classification.

#### 3.2 Variable selection

The variables selected for clustering analysis should be those best representing the characteristics of freeway operating conditions. In practice, traffic engineers often use density as the criterion to define the level of service for freeways, as recommended by the Highway Capacity Manual (TRB, 2000); while travelers tend to use speed as a measure of congestion. In this study, all three primary variables—flow rate, occupancy, and speed—are used as the input attributes to the agglomerative clustering

---

**Fig. 1.** Loop detector station (source: http://pems.eecs.berkeley.edu/public/).
A nested clustering technique for freeway operating condition classification

Table 1  
BIC and $R$ values for each clustering model

<table>
<thead>
<tr>
<th>Number of clusters ($J$)</th>
<th>BIC</th>
<th>$R(J)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1785.631</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>941.932</td>
<td>1.842</td>
</tr>
<tr>
<td>3</td>
<td>774.187</td>
<td>1.121</td>
</tr>
<tr>
<td>4</td>
<td>594.260</td>
<td>1.099</td>
</tr>
<tr>
<td>5</td>
<td>615.214</td>
<td>1.464</td>
</tr>
<tr>
<td>6</td>
<td>620.188</td>
<td>1.062</td>
</tr>
<tr>
<td>7</td>
<td>652.614</td>
<td>1.263</td>
</tr>
<tr>
<td>8</td>
<td>689.377</td>
<td>1.394</td>
</tr>
</tbody>
</table>

algorithm. This should account for the characteristics of operating condition as described by all three primary traffic parameters.

3.3 Initial clustering

The clustering algorithm is first implemented on the whole data set. Table 1 lists the BIC and the ratio of changes in dispersion measurement for each number of clusters ranging from $J = 1$ through $J = 8$. One can observe that the BIC declines from $J = 1$ through $J = 4$ and starts to increase when $J = 5$. Therefore, a coarse estimate of the number of clusters should be four, i.e., $\hat{J} = 4$. Next, the two largest ratios of changes in dispersion measurement for $J = 2, \ldots, \hat{J}$ are $R(J_1) = 1.842$ when $J_1 = 2$ and $R(J_2) = 1.121$ when $J_2 = 3$, respectively. As $R(J_1)/R(J_2) = 1.842/1.121 = 1.643 > 1.15$, then the optimum number of clusters ($J^*$) is set to $J^* = J_1 = 2$. Given the cluster number of two, the $K$-means algorithm is then performed to classify the data records into one of the two clusters. The results are presented on the speed-flow and flow-occupancy diagrams as shown in Figure 2.

Cluster 1 represents the uncongested operating condition symbolized by high speeds (with an average of 64 mph), low occupancy rates (less than 0.15%), and a wide range of flow rates. Cluster 2 contains data records under congested conditions, marked by a mostly high flow rate (equivalent to 1,800–2,000 vehicles per hour per lane) with moderate speeds, and a high occupancy rate.

Table 2 shows the centers of the two clusters. They can be interpreted as the average values for each of the attributes (flow rate, speed, and occupancy). Also listed in the table is the $F$-statistics (standardized values) for each of them, which measures the contribution of each attribute to the clustering result. The higher the $F$-statistics for a variable, the more significant it is to the clustering result. It can be observed that in this clustering stage, occupancy is the most significant among the three attributes.

Although the data set is divided into two clusters, each representing a statistically distinct operating condition, it is a common understanding among researchers and practitioners that traffic may display different characteristics within either condition (Banks, 1999; Kerner and Rehborn, 1996). However, such differences in traffic characteristics might not be highly significant to the clustering algorithm when data from both uncongested and congested conditions are present. To test this theory, the potential to form subclusters within each of the two main clusters using the same clustering concept is further investigated.

3.4 Subclustering

This time, the agglomerative clustering algorithm is applied to the set of data records in Cluster 1 only.

Table 2  
Cluster centers and $F$ statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>$F$-Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow rate (vehicles/5 min)</td>
<td>1208.203</td>
</tr>
<tr>
<td>Average speed (mph)</td>
<td>489.325</td>
</tr>
<tr>
<td>Occupancy (%)</td>
<td>2368.653</td>
</tr>
</tbody>
</table>

Fig. 2. Speed-flow and flow-occupancy plots after the initial clustering.
Table 3 lists the BIC and the ratio of changes in dispersion measurement for each clustering solution ranging from \( J = 1 \) through \( J = 5 \). Using the same rule described earlier, we set \( \hat{J} = 3 \). Thus, for \( J = 2, \ldots, \hat{J} \), \( R(J_1) = 1.422 \) when \( J_1 = 2 \) and \( R(J_2) = 1.206 \) when \( J_2 = 3 \). As \( R(J_1)/R(J_2) = 1.179 > 1.15 \), we set \( J^* = J_1 = 2 \). In other words, the data records under uncongested conditions can further be grouped into two subclusters with statistically significant differences. The \( K \)-means algorithm is then performed to classify the data records into one of the two clusters—Cluster 1-1 and Cluster 1-2.

Next, we further explore the potential of additional divisions under each of the clusters (1-1 and 1-2). Using Cluster 1-2 as an example, the above procedure is repeated on the data points in this cluster. With \( BIC(1) = 473.587 \) and \( BIC(2) = 498.097 \), the algorithm suggests that \( J^* = 1 \). In other words, the statistical differences between the data records are not significant enough to justify further division of the traffic flow phase under Cluster 1-2. A similar conclusion can be drawn for Cluster 1-1 as well.

For Cluster 2, the algorithm suggests that two subclusters (2-1 and 2-2) be formed. Furthermore, it indicates that two subclusters are justified under Cluster 2-2, as Cluster 2-2-1 and Cluster 2-2-2. No additional division is deemed necessary for Clusters 2-1, 2-2-1, and 2-2-2.

3.5 Final cluster tree

The nested clustering technique produces a final cluster tree shown in Figure 3. Each leaf node contains data records with statistically homogeneous attributes (flow rate, speed, and occupancy), and at the same time, demonstrates heterogeneity from other leaf nodes. Thus, each can be considered as representing a distinct operating condition or flow phase. Figure 3 also indicates the designation of flow phase ID in relation to the cluster ID. Table 4 lists the center for each leaf node cluster.
A nested clustering technique for freeway operating condition classification

Table 4
Centers of the clusters

<table>
<thead>
<tr>
<th>Variable</th>
<th>Flow phase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1  2  3   4   5</td>
</tr>
<tr>
<td>Flow rate (vehicles/5 min)</td>
<td>239 373 458 448 403</td>
</tr>
<tr>
<td>Average speed (mph)</td>
<td>60.1 66.9 56.3 44.5 35.0</td>
</tr>
<tr>
<td>Occupancy (%)</td>
<td>9  12  18  22  25</td>
</tr>
</tbody>
</table>

(flow phase) in which the transition of operating conditions can also be observed. The clustering result is also plotted in a flow-speed-occupancy diagram as illustrated in Figure 4. Each leaf node cluster is shown in different marker and shade and is labeled by the flow phase ID.

One can observe that the operating conditions for Phases 1 and 2 are marked by high speed, low occupancy, and low flow rate. With the increase of occupancy, the traffic flow enters Phase 3 with flow rate reaching the highest level and average speed dropping to a moderate level (at 56.3 mph). It can be observed from Figure 4 that the transition from Phase 2 to Phase 3 occurs when occupancy exceeds approximately 15% at this site. When occupancy further increases to above 20%, the flow enters Phase 4. Average speed drops to 44.5 mph with a small change in average flow rate. The worst congestion recorded by the data set is depicted by Phase 5 featuring high occupancy and low speed.

It should be noted that most data points (about 95%) in Phase 1 were recorded during the nighttime, while the data points recorded during the daytime mostly represent the relatively heavy flow condition. This could justify the lower average speed in Phase 1 compared to that in Phase 2.

4 RESULT DISCUSSIONS

The nested clustering technique produces a cluster tree based on the characteristics of the attributes. The hierarchical structure of the tree basically matches the common understanding that there are five phases in the traffic flow. However, the dissimilarities among all five phases are not equally significant at the same time. For example, the difference in operating conditions between Phase 1 and Phase 2 are not recognized by the algorithm during the 1st tier clustering; it is only significant enough in the 2nd tier clustering. Similarly, the difference between Phase 4 and Phase 5 is only significant in the 3rd tier clustering.

4.1 Significance of the attributes

At each partition of the data set, the $F$-statistics are also generated to illustrate the significance of the attributes to the clustering result. They are also displayed in Figure 3. For the 1st tier clustering, occupancy is the dominant factor with a considerably higher $F$-value than flow rate and speed. At the 2nd tier, flow rate becomes the predominant contributor during the division of Cluster 1 for Subclusters 1-1 (Phase 1) and 1-2 (Phase 2). When dividing
Cluster 2, occupancy and flow are far more significant than speed while having an almost equal contribution between them. The 3rd tier clustering involves only the division of Cluster 2-2 into Subclusters 2-2-1 and 2-2-2. During this process, occupancy turns out to be the most significant contributor to the clustering result.

The varying significance of a particular attribute at different clustering tiers suggests that one should probably explore the impact of each attribute to the freeway operating condition classification rather than depending on a particular one (such as density or speed) throughout the process.

4.2 Applicability of the technique

The nested clustering technique is also implemented on traffic data for other sites. This is to verify the applicability of the technique on various sets of traffic data, which could describe very different traffic scenarios. For example, the nested clustering technique also identifies five flow phases based on the data collected at mile point 25.056 on I-80 westbound in California between Tuesday, October 5 and Friday, October 15, 2004. Figure 5 presents the distinct phases as classified by the nested clustering technique in a speed-flow diagram.

Furthermore, sites with relatively homogeneous operating conditions are selected to test the performance of the nested clustering technique. For illustration purposes, we use the traffic data collected by a single-loop detector located at mile point 64.08 on I-580 westbound in California between Tuesday, October 5 and Friday, October 15, 2004. Figure 6 shows the final clusters for this site on a speed-flow diagram. It is determined by the nested clustering algorithm that only two flow phases existed at this site during the period.

These tests indicate that the nested clustering technique is able to adapt to the variation in operating condition at different sites and to provide reasonable clustering results. This feature is particularly important in the analysis of freeway operating conditions because every site is different.

5 CONCLUSIONS

In this study, a nested clustering technique is developed for analyzing freeway operating conditions. The optimum number of clusters is determined by applying the BIC criterion on the binary tree generated by the agglomerative clustering algorithm. This procedure is repeated on all clusters and subclusters until further clustering becomes unnecessary, as deemed by the algorithm.

The attributes used in the nested clustering technique are flow rate, speed, and occupancy, because they are the three most important parameters in describing operating condition. Each leaf node cluster generated by the nested clustering technique contains a set of data points with statistically homogeneous traffic characteristics, while
the dissimilarity between the clusters is statistically significant. Analysis shows that each leaf node cluster can be considered as a traffic flow phase with distinct characteristics.

For a site with a full range of operating conditions, the significance of attributes varies with clustering tiers as well as operating conditions. It appears that occupancy and flow rate are the two variables that are overall significant to the clustering results. When data under all operating conditions is present, occupancy is the most significant contributor to the division of operating condition into uncongested and congested states. It is also the most significant variable in the nested clustering process under congested conditions. Flow rate appears to be the most significant variable for further division of uncongested conditions. This observation implies the importance of considering more than one variable during the analysis of freeway operating condition.

A major advantage of this technique is its flexibility in determining the number of clusters based on the statistical characteristics of the data. The nested clustering technique is tested on traffic data sets obtained from multiple sites with varying operating conditions. The clustering results for each test site seem reasonable.

This technique appears to be an effective and flexible tool for data mining in freeway flow phase analysis. Preset values as the boundaries between flow phases are not required. Instead, it differentiates flow phases by grouping data records according to their statistical characteristics. This technique can be applied to a broad range of roadways from which continuous traffic data are collected.

REFERENCES


