This is the final version (unedited) of the paper published in IEEE Transactions of Intelligent Transportation Systems.

Optimizing Freeway Traffic Sensor Locations by Clustering Global Positioning System Derived Speed Patterns

Jalil Kianfar and Praveen Edara

Abstract—This paper presents a new clustering-based methodology for sensor placements on freeways for estimating travel times. The proposed methodology is applicable to both freeways without any existing deployment and freeways with existing sensor deployments where it identifies the critical sensors that need to be regularly maintained. The freeway sections are clustered based on speed data—neighboring sections with identical speed profiles being grouped into a cluster. A new approach of estimating freeway travel times using the final clusters, called optimal placement method, is also proposed. A family of K-means clustering algorithms and a hierarchical algorithm are then explored using real-world case studies of three freeway segments in Virginia. Speed and travel time data are obtained using Global Positioning System (GPS) equipped probe vehicles. The clustering results indicated that the hierarchical and K-means with a priori knowledge algorithms produced the best clusters. The tradeoff plots of travel time measures (e.g., error) versus number of freeway sensors were generated for two travel time estimation methods—optimal placement method and midpoint placement method. The optimal placement method consistently produced better travel time estimates than the midpoint placement method for all three case studies. The travel times were also estimated using three other methods found in the literature: 1) zone of influence method, 2) instantaneous method, and 3) linear method. The results showed that the optimal placement method outperformed these methods in all three case studies.

Index Terms—Clustering methods, Sensors, Traveler information systems

I. INTRODUCTION

STATE transportation agencies make significant investments in installing and maintaining point sensors (such as inductive loops or microwave sensors) on highways to monitor traffic conditions. Traditional design practice has called for the sensors to be located at uniform spacings, most commonly ½ kilometer (or ½ mile). The origin of this practice is based largely on analyses that have concluded that incident detection algorithms can provide acceptable time-to-detection with this spacing interval. However, given the proliferation of cellular phones, incident detection algorithms seldom play a key role in modern traffic management centers [1]. In recent times, transportation agencies have found travel time data to be of particular importance in managing traffic and providing traveler information. Many agencies derive travel time estimates using speed data reported by traffic sensors.

This new focus on travel times as opposed to incident detection demands a new methodology for sensor deployment (spacing) on highways. Such a methodology would identify the locations that are critical to accurate travel time estimation. Significant cost savings will be achieved by discontinuing maintenance of sensors at non-critical locations on highways with an existing dense deployment. For highways without existing sensor deployments installation costs will also be saved by installing sensors only at the critical locations (as opposed to every ½ kilometer or ½ mile). In this paper, we present a clustering-based methodology that identifies the optimal sensor locations to accurately estimate travel times on highways. In addition, a new travel time estimation procedure is proposed to improve the travel time estimation accuracy. The proposed methodology is then implemented on three freeway segments in Virginia using real traffic data collected using Global Positioning System (GPS) equipped probe vehicles.

The remaining parts of this paper are organized as follows: the next section reviews previous research literature; third section briefly describes the clustering algorithms used in this research; fourth section describes the proposed methodology for freeway sensor placement; fifth section presents the case studies and clustering results; and the last section concludes with a discussion of the study findings.

II. BACKGROUND

The focus of this paper is on travel time estimation using point sensors such as inductive loop detectors, remote traffic microwave sensors (RTMS), and magnetic sensors. The point sensors are in widespread use in state transportation agencies in the United States. Travel time measurement technologies such as automatic vehicle identification (AVI), vehicle signature analysis [2], vehicle re-identification [3], etc, are fundamentally different at obtaining travel times (by tracking vehicle trajectories) as compared to the estimation methods
commonly used with point sensors. Also, the problem of travel time prediction [4]-[7] is different from estimation. Both travel time measurement and prediction methods are beyond the scope of this paper. In the literature, optimization techniques have been used in determining the placement of 1) sensors for estimating origin-destination trip matrices [8]-[10], link traffic flows [11]-[12], 2) AVI readers for travel time estimation [13]-[14], 3) sensors for minimizing the travel time variance and social costs [15].

There is limited past research focusing on the placement of point sensors for the purposes of travel time estimation. Initial research has primarily focused on the travel estimation on urban arterial streets using traffic simulation [16]-[17]. Research on the sensor placement problem for freeway travel time estimation has recently begun.

Fujito et al. [18] studied the effect of sensor spacing on travel time index (a measure reflecting travel time) using field data from Cincinnati, Ohio, and Atlanta, Georgia. Their analysis concluded that the actual placement of sensors was critical in accurately estimating the traffic congestion levels on a freeway segment. Liu et al. [19] used simulation to examine some widely used travel time estimation methods for different sensor spacings including a constant speed based (CSB) algorithm, a piecewise constant speed (PCSB) algorithm, and a piecewise linear speed based (PLSB) algorithm. They built a simulation model for the I-70 corridor in Maryland and found that for free flow conditions it was sufficient, for both monitoring and travel time estimation, to have sensor stations placed at both ends of the segment as long as the sensor data are reliable. In addition, the study found that for congested segments, more sensors certainly provide a better estimate of travel time variation. Based on these findings, they proposed rules and an iterative procedure for locating a limited number of sensors.

Li et al. [20] compared the performance of four different models – an instantaneous model, a time slice model, a dynamic time slice model, and a linear model. All models compute freeway segment travel times by aggregating the travel times of constituent sections (with sensors at the beginning and end of a section). They are, however, different in the way they estimate the section travel times. The instantaneous model uses speeds reported by sensors at an instant of time \( t \). The time-slice and dynamic time-slice models use speed values at the time points when the vehicle is expected to travel on each section. The linear model interpolates speeds within a section instead of averaging the speeds of beginning and ending sensors. The purpose of this study was not to determine the optimal sensor locations, but instead to determine the best method to estimate freeway segment travel times. They conducted case studies of two freeway segments in Melbourne, Australia. Results showed that the estimated travel time errors for the four models were similar and that all four models underpredicted travel times.

Bartin et al. [21] considered the sensor placement problem as a space discretization problem irrespective of the travel time estimation method used. It was demonstrated that sensor placement can be formulated as a clustering problem. Space-time vehicle trajectories were used to identify the optimal locations of sensors. The proposed method was tested using simulated data of a 20 km roadway segment. The vehicle trajectories were collected in 20 m intervals. Global K-means clustering algorithm was applied to the study segment and was shown to produce better travel time estimates than the equidistant approach.

The current research presented in this paper addresses the key shortcomings of previous researches and makes the following contributions:

1) In this paper, we hypothesize that the deployment of traffic sensors on freeways can be optimized by clustering freeway sections based on typical speed profiles. We thoroughly examine the applicability of hierarchical clustering and several variations of K-means clustering algorithms for grouping freeway sections. The interdependency of the sensor placement problem and the travel time estimation problem is investigated to show that the two problems are not only interrelated but that more accurate travel times can be obtained by modeling the two problems jointly instead of treating them independently. In applying the clustering algorithm, we present a novel approach of defining sections and estimating travel times. The developed estimation method was applied to three real-world case studies in two different regions in Virginia and compared with estimation methods found in the literature.

2) Previous studies [19]-[21] have used freeway segment travel time error as the key performance measure while comparing different travel time estimation methods. This is understandable since their objective was to determine the best travel time estimation method for the entire segment of the freeway. However, for sensor placement purposes this measure is not sufficient to capture the error propagation from one section to another within a freeway segment. A certain sensor placement strategy that produces the best travel time estimate for the entire freeway segment may actually overestimate the travel times in some sections and underestimate in some other sections. The positive and negative errors, resulting from the overestimation and underestimation, may cancel each other and hence result in a lower overall error. In this research, we propose a new performance measure called the error uniformity index (EUI) which specifically addresses this issue. This measure is unique and has not been used in previous research.

3) The data used in this research is obtained from actual Global Positioning System (GPS) equipped probe vehicles driven by the researchers on freeway study segments in Virginia. To the best of our knowledge, the use of such detailed field data for the sensor placement application (for travel time estimation) is unprecedented. Previous research has exclusively relied on simulated data. Li et al [20] have used field data for comparing different travel time estimation algorithms but not for determining sensor placements.

III. CLUSTERING TECHNIQUES STUDIED IN THIS RESEARCH

The objective of clustering algorithms is to classify a set of...
data into groups or clusters. Data elements are grouped such that similar elements are assigned to a same group [22]. The objects are clustered in such a way that similarity between members of a cluster is maximized and similarity between members belonging to different clusters is minimized [23]. Dunham [24] defines a cluster as a group of data points for which the distance of a data point from any member within its cluster is less than its distance from any member belonging to other clusters. A pattern is defined as a unique set of attributes. Each cluster represents a pattern in data and all data elements belonging to one cluster share this pattern which cannot be found in other clusters.

In this research we employ two clustering algorithms: hierarchical clustering and K-means clustering. The next two sections describe these algorithms.

A. Hierarchical Clustering

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. Consider a set of N data points \( Y = \{y_1, y_2, \ldots, y_N\} \) where \( y_i = (x_1, x_2, \ldots, x_i) \). In this paper, vector quantities are denoted in boldface and scalar quantities are denoted in normal font. Hierarchical clustering algorithm is described as follows [22]:

Step 1: Assign each data point to a separate cluster. The number of clusters would be equal to N.

Step 2: Find the two most similar clusters \( \{y_i\} \) and \( \{y_m\} \) and merge the two to form a new cluster \( \{y_i, y_m\} \). Decrease the total number of clusters by one.

Step 3: Repeat Step 2 until data is grouped into the required number of clusters (predefined).

There are several methods for computing the similarity between clusters. Single linkage method is one such method which computes similarity between two clusters based on the distance between the two closest cluster members. Consider two clusters J and K; equation (1) shows similarity between J and K based on the single linkage method.

\[
\|J - K\| = \min_{q=K} \|p - q\| = \sum_{i=1}^{m} \|p_i - q_i\|
\]

where, \( p = (p_1, p_2, \ldots, p_t) \) and \( q = (q_1, q_2, \ldots, q_t) \) belong to clusters J and K respectively.

B. K-means Clustering

K-means clustering algorithm has as an input, a predefined number of clusters, \( k \). Means stands for the average location of all the members of a particular cluster. K-means algorithm is an iterative procedure involving the computation of cluster centroids. Centroid is an artificial point in space which represents the average location of the particular cluster. The steps of the K-means algorithm are given below:

Step 1: Randomly choose \( k \) points as the seeds for the centroids of \( k \) clusters.

Step 2: Assign each data point to the closest centroid, hence forming \( k \) distinct clusters.

Step 3: Update the location of each centroid. The attributes of centroid are average of all attribute values of the data points belonging to the same cluster.

Step 4: Check if the cluster centroids have changed their location (coordinates). If yes, start again from step 2. If not, the algorithm has yielded the best \( k \) clusters.

The convergence of this algorithm usually needs only a few iterations. The distance metric that is commonly used for this algorithm is the Euclidean metric defined in (2). The Euclidean distance between two points, \( p = (p_1, p_2, \ldots, p_t) \) and \( q = (q_1, q_2, \ldots, q_t) \) is defined as:

\[
d = \sqrt{\sum_{i=1}^{t} (p_i - q_i)^2}
\]

The number of clusters \( k \) should be chosen so as to match the natural structure of the data (which is not a trivial task) in order to obtain good results. To do this, a trial and error procedure is performed with different values of \( k \). In theory, the best \( k \) value will exhibit the smallest intra-cluster distances and largest cluster to cluster distances.

C. Silhouette Measure

Silhouette value is a performance measure for evaluating cluster structures. It considers the impact of each cluster member on cohesion or separation of cluster. For each data point (cluster member) silhouette value is computed in the following way [25]:

Step 1: Consider data point \( (p) \) in cluster \( (J) \). Calculate average distance between data point \( (p) \) and other data points in cluster \( (J) \), this value is named \( a_p \).

Step 2: Calculate average distance between data point \( (p) \) and other data points in other cluster \( (K) \), repeat this for all clusters \( K \neq J \), find minimum average distance and name it \( b_p \).

Step 3: The silhouette value for data point \( (p) \) is:

\[
s_p = \frac{(b_p - a_p)}{\max(a_p, b_p)}
\]

The average silhouette value of all data points is called silhouette coefficient. 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 [26]. A clustering method with silhouette coefficient greater than 0.5 is considered to produce reasonable clusters. Next section describes the proposed methodology for sensor placement and travel time estimation.
IV. METHODOLOGY

The developed methodology groups locations within a freeway segment that have similar speed patterns into a predefined number of clusters. To do this, the freeway is first discretized into smaller cells. Let us assume the freeway segment shown in Fig. 1 is divided into \( N \) cells that are of equal length \( \delta l \). For example, \( \delta l \) could be equal to \( \frac{1}{2} \) km. The relationship between \( \delta l \) and the freeway segment length \( L \) is given by,

\[
\delta l = \frac{L}{N} \tag{4}
\]

\[\text{Fig. 1. Discretization of the Freeway Section}\]

A speed profile represents the speed variations across different cells of a freeway segment whereas a speed pattern represents the variation of speeds within a cell over time (i.e., across different speed profiles). Therefore, speed patterns are derived from speed profiles. Let \( V_j \) represent the speed pattern of cell \( j \). A speed pattern is a vector of speeds measured at the center of each cell during the data collection period (at different times). \( V_j = [v_{j1}, v_{j2}, \ldots, v_{jr}] \) where \( v_{ji} \) is the speed at cell \( j \) in \( i^{th} \) speed profile and \( r \) is the total number of speed profiles. \( V_j \) can represent speed during different times of day and days of week and it is not necessarily continuous over time. For example, in the case study that will be presented later real speed profiles were obtained using Global Positioning System (GPS) equipped probe vehicles that traversed the study freeway segment in Virginia several times during the peak period on three weekdays. The collected data is also used to calculate the ground truth travel time (GTTT). For each probe vehicle, the entering and exiting times for the freeway segment are compared to obtain the GTTT specific to that vehicle and its speed profile.

In the second step, a clustering algorithm is used to identify and group similar cells into clusters. Speed patterns of all (\( N \)) roadway cells \( \{V_1, V_2, \ldots, V_N\} \) are used for clustering. The number of clusters is varied from 2 to \( N \), the total number of cells. The next step is to convert the clusters into detection sections. The clustering algorithm assigns freeway cells to different clusters. Successive cells (cells that are physically next to each other) that belong to the same cluster are merged together to form a detection section (See Fig. 1). One sensor is allocated to each detection segment. Equation (5) shows the length of a detection section,

\[
\delta L_i = (d_i - u_i + 1) \delta l \tag{5}
\]

\[\text{where, } \delta L_i : \text{length of detection section } i \]
\[u_i : \text{starting cell of section } i \]
\[d_i : \text{ending cell of section } i \]
\[\delta l : \text{length of a cell}\]

It is important to note that the number of detection segments (and number of sensors) is not necessarily equal to the number of clusters. The cells that belong to the same cluster may be physically separated from each other with other cells belonging to other clusters in between them. In such cases, successive cells that belong to the same cluster are merged together to form a detection section and the remaining cluster cells will form separate section(s). Therefore, the number of detection sections will be greater than or equal to the number of clusters.

The next step is to identify the locations within these sections to place the sensors. We consider two approaches for sensor placement within the detection sections: Midpoint placement and Optimal placement. These methods not only identify the sensor location but also determine how a section’s travel time is estimated. The freeway segment travel time is then estimated by summing the travel times of all constituent sections. The next two sections describe these methods.

A. Midpoint Placement Method

In the midpoint placement method a sensor is placed in the middle of each detection section. The travel time over the detection section is calculated by dividing section length by the speed at the center of section. When the section size (in terms of number of cells) is odd then the sensor is placed in the middle cell but when the number of cells in the section is even, sensor is placed randomly at one of the two middle cells. Equations (6) and (7) show these travel time calculations.

\[
\text{sett}_i = \frac{\delta L_i}{v_i'} \tag{6}
\]

\[
\text{ETT} = \sum_{i=1}^{m} \text{sett}_i \tag{7}
\]

\[\text{where, } \text{sett}_i : \text{section } i \text{'s estimated travel time} \]
\[\delta L_i : \text{length of section } i \]
\[v_i' : \text{speed at midpoint of section } i \]
\[\text{ETT}: \text{estimated freeway segment travel time} \]
\[m: \text{number of sections i.e. number of sensors}\]

B. Optimal Placement Method

Although the midpoint placement method is intuitive, the cell in the center of a section might not be the best exemplar (representative) of speed in that section. To overcome this shortcoming an optimal placement method is proposed. Instead of placing the sensors at midpoints of every detection section, they can be placed at locations that produce the optimal travel time estimates for the detection sections. The
objective is to identify sensor locations that will minimize the average relative absolute error for each detection section. Sections travel times are estimated as shown in (8) to (12) that follow.

\[
sett_i' = \frac{\delta L_i}{v_{j_k}}
\]  
(8)

\[
\epsilon_i' = \sum_{j=1}^{r} \left[ \left| sett_i' - sgtt_i' \right| \right] / \left( sgtt_i' \times r \right)
\]  
(9)

where, \( sett_i' \): estimated travel time for section \( i \) when the sensor is placed at cell \( j \) for the \( k \)th speed profile \( \delta L_i \): length of the \( i \)th detection section \( v_{j_k} \): speed at cell \( j \) for the \( k \)th speed profile \( \epsilon_i' \): average absolute relative travel time estimation error for section \( i \) when a sensor is placed at cell \( j \) \( sgtt_i' \): ground truth travel time for section \( i \) for the \( k \)th speed profile \( r \): total number of speed profiles

The cell placement that results in the minimum error is obtained using (9).

\[
\epsilon_{\text{opt}} = \min \left( \epsilon_1', \epsilon_2', \ldots, \epsilon_i' \right)
\]  
(10)

where, \( \epsilon_{\text{opt}} \): section \( i \)'s minimum average absolute relative travel time estimation error \( \text{opt} \): optimal sensor location for section \( i \) \( u_i \): starting cell of section \( i \) \( d_i \): ending cell of section \( i \) \n
Estimated total travel time for the freeway segment (\( ETT \)) is then calculated by summing the travel times of all constituent sections as shown in (12).

\[
sett' = \delta L / v_{\text{opt}}
\]  
(11)

\[
ETT = \sum_{i=1}^{m} sett'
\]  
(12)

where, \( sett' \): estimated travel time for the section \( i \) \( \delta L_i \): length of the section \( i \) \( v_{\text{opt}} \): speed at section \( i \) \( ETT \): estimated freeway segment travel time \( m \): number of sections i.e. number of sensors

C. Performance Measures

Two different measures are defined to evaluate the freeway segment travel times estimated from different sensor deployments. First, the average absolute error (\( AAE \)) is the sum of absolute values of the error between the estimated travel time and the ground truth travel time over all speed profiles measured during data collection. It is defined as follows:

\[
AAE = \left( \sum_{j=1}^{r} \left| ETT_j - GTTT_j \right| \right) / r
\]  
(13)

where, \( ETT_j \): estimated travel time for speed profile \( j \) \( GTTT_j \): ground truth travel time for speed profile \( j \) \( r \): total number of speed profiles

\( AAE \) is a “segment-level” measure reflecting the error in the estimated travel time over the whole freeway segment. It only considers times when a driver enters and exits the freeway segment. The measure does not reflect the errors in the travel times of the sections contained within a segment. A certain sensor placement strategy that produces the best travel time estimate (minimum \( AAE \)) for the whole freeway segment may actually overestimate the travel times in some sections and underestimate in some other sections. The positive and negative errors, resulting from the overestimation and underestimation, respectively may cancel each other and hence result in a lower \( AAE \) value. To measure the degree of over and under estimation of travel times over the constituent sections, a second performance measure - Error Uniformity Index (\( EUI \)) is defined as follows:

\[
EUI = \frac{1}{L \times r} \sum_{j=1}^{r} \sum_{i=1}^{m} \delta L_i \left( \left| sett_{i, j} - sgtt_{i, j} \right| \right) / sgtt_{i, j}
\]  
(14)

where, \( EUI \): Error Uniformity Index \( L \): length of freeway segment \( r \): total number of speed profiles \( m \): number of freeway detection sections \( \delta L_i \): length of section \( i \) \( sett_{i, j} \): section \( i \) estimated travel time for the \( j \)th speed profile \( sgtt_{i, j} \): section \( i \) ground truth travel time for the \( j \)th speed profile

V. APPLICATION OF THE METHODOLOGY TO CASE STUDIES

The proposed clustering-based methodology was applied to three freeway segments in Virginia: 1) a 16 km (10mi) segment of eastbound Interstate 66 (I-66 EB) in Northern Virginia, 2) an 11 km (7 mi) segment of southbound Interstate 95 (I-95 SB) in Richmond, and 3) an 11 km (7 mi) segment of northbound Interstate 95 (I-95 NB) in Richmond. For the I-66 segment, three probe vehicles equipped with GPS devices were driven during the morning peak period on three weekdays to obtain a total of 33 different speed profiles. For the southbound and northbound segments of I-95, two probe vehicles were driven both during morning and evening peak periods to obtain 53 and 49 speed profiles, respectively. The two regions represent the two extremes of sensor deployment commonly found in traffic management centers. Northern Virginia already has extensive deployment of closely spaced sensors, while Richmond has very few sensors deployed. In these cases, Northern Virginia would benefit from knowing which sensors are critical to accurate travel time estimation.
and which sensors could be eliminated from future maintenance expenditures. On the other hand, in Richmond, where there are few sensors currently in the field, additional deployment could be undertaken with accurate travel time estimates as a goal and the realities of available maintenance funding as a constraint. Conceptually, this data collection method is straightforward. A GPS device is installed in a vehicle and then a driver drives the vehicle according to the “flow of traffic” throughout the study segment. While the vehicle is running, the GPS device automatically logs latitude/longitude points and times. In order to obtain reliable data and a sufficient sample size, multiple probe vehicles were deployed for data collection with headway of 5 minutes between subsequent vehicles. Speeds obtained from probe vehicles are compared with 1-minute loop sensors speeds to verify the accuracy of GPS collected data. Comparison is performed on a segment of I-66 EB which already has installed loop sensors. Only regularly maintained detectors were chosen for comparison. A two-tailed t-test with unequal variance showed that the GPS speeds were not statistically different from the loop sensor speeds the 5% significance level (Table I).

<table>
<thead>
<tr>
<th>Probe Vehicle Number</th>
<th>GPS Speed (mph) Mean</th>
<th>Std. Dev.</th>
<th>Speed from Loops (mph) Mean</th>
<th>Std. Dev.</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55.9</td>
<td>12.7</td>
<td>55.0</td>
<td>12.5</td>
<td>0.63</td>
</tr>
<tr>
<td>2</td>
<td>56.9</td>
<td>12.6</td>
<td>54.9</td>
<td>13.3</td>
<td>0.40</td>
</tr>
<tr>
<td>3</td>
<td>53.6</td>
<td>13.7</td>
<td>52.7</td>
<td>13.5</td>
<td>0.64</td>
</tr>
</tbody>
</table>

For more details about the data collection effort the reader is referred to [27]. For applying the developed methodology, the freeway segment was divided into cells of equal length as previously explained in section IV. Speed pattern at the center of each cell was extracted from the GPS data within a 91 m (300 ft) vicinity of the center. The length of each cell was set to 480 m (0.3 mile), the minimum size possible. Lengths less than 480 m could not be used due to errors in the speeds derived from GPS-reported data. Fig. 2 shows a sample speed profile for I-66 EB showing the speed of a probe vehicle at the center of 35 freeway cells during one passage. There are 33 speed profiles with GTTT between 9 and 21 minutes. Fig. 3 shows a sample speed pattern for I-66 EB. There are 35 speed patterns (equal to the number of cells) and the next section describes how similar speed patterns were identified using the clustering techniques defined earlier.

A. Application of Clustering Techniques
Clustering techniques were employed to cluster the freeway cells based on the speed patterns obtained from GPS data. After determining the clusters, sensor location within a cell cluster (i.e., a section) was determined using the midpoint method or the optimal placement method. Travel times for the whole freeway segment were estimated after placing sensors.

1) Hierarchical Clustering
A hierarchical clustering algorithm was applied to cluster cells based on the speed patterns. The pair-wise distance between the speed patterns of cells was computed using the Euclidean distance metric and the closest pairs were identified using the single linkage method.

2) Basic K-means
Initially, the basic K-means algorithm was investigated for clustering cells based on speed patterns. The initial cluster centroid locations were randomly generated by the algorithm.

3) K-means with replicates
For a given number of clusters, the basic K-means algorithm was applied several times with different sets (replicates) of initial centroid locations. The best among the resulting final clusters from all replicates was selected based on the least intra-cluster distance (data points are within close proximity) and the greatest inter-cluster distance. This procedure was implemented for 250 replicates.

4) K-means with a priori knowledge
A variation of the basic K-means algorithm, the K-means with a priori knowledge was also studied. In this algorithm, the initial centroid locations for n clusters were established based on the final centroids of n-1 clusters determined a priori. First, the n-1 centroid locations were set and the n° centroid location was randomly assigned. Similar to the previous algorithm of
replicates, several sets of initial centroid locations were obtained by randomly varying the location of the $n^{th}$ centroid. Speed patterns were then clustered with these initial centroid locations and the best set of clusters were determined using the intra and inter-cluster distances as previously described. Freeway sections were then defined based on the final cluster compositions.

B. Results of Clustering Applications

In order to obtain the tradeoff between the number of sensors and the travel time accuracy the number of clusters were varied from the minimum to the maximum number of clusters possible: 2 to 35 for I-66 EB, 2 to 24 for I-95 SB, and 2 to 24 for I-95 NB segments. The maximum possible number of clusters occurs when there is one sensor in each cell resulting in a full coverage of the freeway segment. Intuitively, it can be hypothesized that the travel time accuracy increases with the increase in the number of clusters due to the increased information of the freeway traffic conditions. The four clustering techniques were compared based on the Silhouette measure described earlier in Section III-C. The Silhouette coefficients are shown in Table II. The greater the Silhouette coefficient the better is the clustering of data points. From Table I it is evident that the performance of all four techniques was close to each other with hierarchical and k-means with a priori knowledge performing the best. Therefore, these two techniques were considered for further analysis.

<table>
<thead>
<tr>
<th>Clustering Technique</th>
<th>I-66 EB</th>
<th>I-95 SB</th>
<th>I-95 NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hierarchical</td>
<td>0.57</td>
<td>0.62</td>
<td>0.68</td>
</tr>
<tr>
<td>Basic K-means</td>
<td>0.54</td>
<td>0.56</td>
<td>0.61</td>
</tr>
<tr>
<td>K-means with Replicates</td>
<td>0.58</td>
<td>0.59</td>
<td>0.63</td>
</tr>
<tr>
<td>K-means with a priori knowledge</td>
<td>0.59</td>
<td>0.59</td>
<td>0.64</td>
</tr>
</tbody>
</table>

1) Tradeoff plots of performance measures

The tradeoff plots for the travel time performance measures, AAE and EUI, were generated for the three case studies. The plots are shown in Fig. 4, Fig. 5 and Fig. 6 for I-66 EB, I-95 SB, and I-95 NB, respectively. Each of these figures consist of four subplots – (a) and (b) subplots show the AAE values for hierarchical and k-means with a priori knowledge, whereas (c) and (d) subplots show the respective EUI values. In each of the subplots, the results of optimal placement method are shown using a solid line and the results of midpoint placement method are shown using a dotted line.

There are some common trends evident in Fig. 4, 5, and 6. First, the optimal placement method seems to outperform the midpoint placement method in terms of both AAE and EUI measures for all three case studies. The difference in the performance decreases with increase in the number of sensors. This occurs since both AAE and EUI values decrease as the sensor deployment on the study corridor increases. In addition, the increase in sensor deployment means fewer feasible solutions for optimizing the sensor placement within a short section. Second, the plot of optimal placement method is continuously decreasing both in terms of AAE and EUI in most cases (with the one exception in Fig. 6(b) which will be explained later) whereas the midpoint plot has few oscillations.
indicating a non-monotonic decrease. The reason this happens is that the midpoint placement method is susceptible to error cancellations; the travel time estimate errors (over and under estimations) over a segment’s constituent sections cancel out each other and result in an accurate segment travel time. Therefore, in the midpoint placement method the AAE values could increase with an increase in the number of sensors (for example, in Fig 5(a) the AAE for 2 sensors is 50 seconds whereas the AAE for 3 sensors is 62 seconds). Third, the marginal travel time improvement that can be achieved by deploying additional sensors decreases as the number of sensors increase.

There is one exception to the second trend noted earlier. In Fig. 6(b) the AAE value increases as the deployment increases from 2 to 4 sensors for the optimal placement method which means that there could be cancellation of travel time errors. This is further confirmed by reviewing the trend in the EUI plot shown in Fig. 6(d) which shows that the EUI value decreases as the number of sensors increase from 2 to 4. In other words, the solution with 2 sensors may result in an accurate segment travel time than the solution with 4 sensors however; the latter solution is more robust at the section level.

In order to compare the plots of hierarchical and k-means with a priori knowledge a statistical analysis was conducted and the results are presented in the next section.

2) Statistical comparison of hierarchical and k-means with a priori knowledge

The optimal placements for a given number of sensors are obtained using both clustering techniques. Based on these sensor placements, the travel times for every sensor deployment are computed for all speed profiles. A t-test was then conducted, with \( \alpha = 0.05 \), to verify if the travel time estimates obtained from sensor placements resulting from two clustering techniques were the same. The results of the t-test for all sensor deployments were then used to compute the number of instances where the two techniques produced same travel times (could not reject the null hypothesis) and the number of instances where one technique outperformed the other. For the I-66 EB and I-95 SB case studies, it was found that the performance of the hierarchical algorithm was not statistically different from the performance of the K-means with a priori knowledge. For the I-95 NB case study, the performance of the two algorithms was the same in 75% of the sensor deployments; the hierarchical performed better than the k-means with a priori knowledge in 8% of deployments; and the k-means with a priori knowledge performed better than the hierarchical algorithm in 17% of deployments.

The computational times of both algorithms were also compared. The computational times of K-means with a priori knowledge was 31, 17, and 18 seconds respectively for I-66 EB, I-95 SB, and I-95 NB segments compared to only 1 second each for the hierarchical algorithm when the programs were run on a 3.0 GHz CPU with 4 GB memory. This clearly indicates that the computational time is not an issue for both the clustering procedures.

3) Comparison of optimal placement method with other methods

The travel times were also estimated using three other methods found in the literature: 1) zone of influence method [18] that is commonly used by state transportation agencies in the United States, 2) instantaneous method [20], and 3) linear method [20]. For the first method, travel time for the freeway segment is estimated from the travel times of constituent sensor zones of influence. The zone of influence (ZOI) of a sensor is defined as half the distance upstream and downstream to the neighboring sensor. The instantaneous method assumes that speed for each section is equal to the average of speeds at upstream and downstream sensors. Section travel time is calculated by dividing section length over section speed. The linear method assumes a linear variation of the speed from upstream to downstream sensor of a section.

The performance of the optimal placement and the three methods were compared using two measures - mean squared error (MSE) and mean absolute error (MAE). The MAE is same as the AAE measure defined earlier. The variation of MAE and MSE versus the number of sensors is shown in Fig. 7 for I-66 EB, I-95 SB, and I-95 NB, respectively. From these figures, it can be inferred that the travel time errors for the optimal placement method are the lowest. This was expected due to the fact that the ZOI, instantaneous, and linear methods are not sensitive to the observed speed patterns. On the other hand, the optimal placement method specifically groups (clusters) the freeway sections based on speed patterns and is therefore able to produce accurate travel time estimates.
Another observation is also made from these plots. The MAE values follow a steady decreasing trend for the optimal placement method (with the exception of I-95 NB plots as explained in Section V-B-2). However, an oscillatory trend was observed for the other three methods sometimes exhibiting error increases with increase in sensor coverage. This illustrates that the three methods are more susceptible to error cancellations; the travel time estimate errors (over and under estimations) of sections in a segment cancel out each other and result in a smaller segment error value.

VI. CONCLUSION

This paper presented a new clustering-based methodology for sensor placements on freeways for estimating travel times. The proposed methodology is applicable to both, freeways with existing sensor deployments where it identifies the critical sensors that need to be regularly maintained and to freeways without any existing deployment.

The freeway sections are clustered based on speed data – neighboring sections with identical speed profiles being grouped into a cluster. A new way of estimating the freeway travel times using the final clusters was also proposed. A family of K-means clustering algorithms and a hierarchical algorithm were then explored using real-world case studies of three different freeway segments. The case study locations were chosen to represent two extremes of sensor deployment. Speed and travel time data were obtained using global positioning system equipped probe vehicles. Tradeoff plots of travel time performance measures versus number of freeway sensors were plotted. A new performance measure called the error uniformity index was developed to address the localized errors of travel time estimation methods in addition to the average absolute error measure that addresses the accuracy of the estimates for the whole segment (global).

The results indicate that the optimal placement method consistently outperformed the midpoint placement method in terms of both AAE and EUI measures for all three case studies. The difference in the performance decreases with increase in sensor coverage. The silhouette coefficient measure that indicates goodness of clustering indicated that all four clustering algorithms produced reasonable clusters including the basic K-means algorithm. Out of the four algorithms, the hierarchical and K-means with a priori knowledge produced the best clusters.

Statistical analysis of the travel time estimates obtained using hierarchical and K-means with a priori knowledge algorithms showed that the performance of the two algorithms was not statistically different from each other in I-66 EB and I-95 SB case studies. For the third case study (I-95 NB), the estimates were statistically different in 25% of the instances. The results also showed that the optimal placement method produced more accurate travel time estimates as compared to the zone of influence, instantaneous, and linear methods for all three case studies. Both the MAE and MSE values were the lowest for the optimal placement method.

REFERENCES


Jalil Kianfar received his Bachelor’s degree in Civil engineering from IAU-Najaf Abad, Esfahan, Iran and his Master’s in transportation engineering from IAU-Tehran South, Tehran, Iran. He is working towards his PhD degree in University of Missouri Columbia. His research interests include traffic modeling, intelligent transportation systems, artificial intelligence and operations research.

Praveen Edara received his PhD. degree in transportation systems from Virginia Tech. He worked as a research scientist for the Virginia Transportation Research Council conducting research in traffic operations and evacuation modeling. Since 2007, he is working as an Assistant Professor at University of Missouri, Columbia. His current research interests include traffic modeling, traffic simulation, and intelligent transportation systems.