Statistical Methods for Detecting Spatial Configuration Errors in Traffic Surveillance Sensors

Jaimyoung Kwon*
Department of Statistics and Institute of Transportation Studies
University of California,
367 Evans Hall
Berkeley, CA 94720-3860
Tel: (510) 642-2781, Fax: (510) 642-7892
kwon@stat.berkeley.edu

Chao Chen
Department of Electrical Engineering and Computer Science
University of California, Berkeley CA 94720
Tel: (510) 643-5894
chaos@eecs.berkeley.edu

Pravin Varaiya
Department of Electrical Engineering and Computer Science
University of California, Berkeley CA 94720
Tel: (510) 642-5270, Fax: (510) 642-7615
varaiya@eecs.berkeley.edu

For Presentation and Publication
83rd Annual Meeting
Transportation Research Board
January 2004
Washington, D.C.

June 18, 2003

# words: 3,292 (excluding Figure and Table captions)
Plus 1 Table and 9 Figures (2,500)
TOTAL: 5,792

*Corresponding Author
ABSTRACT

With large-scale deployment of traffic surveillance sensors becoming commonplace, it becomes critical to maintain correct information about the spatial configuration of the sensors. The problem is burdensome when hundreds or thousands of sensors are deployed. One common configuration error is the switching of directions of highway loop detectors that share the same cabinet. We propose semi-automatic and automatic methods for detecting such errors, based on the strong correlations between measurements made by spatially close sensors. The semi-automatic method uses a multidimensional scaling (MDS) map of sensors, which visually displays the similarity between sensor measurements, and enables one to easily identify sensor mislabeling. The automatic method uses a scoring scheme that computes the probability of sensor mislabeling from the pairwise distance or similarity matrix. The algorithm, tested on data from a four-lane freeway consisting of 64 sensor locations, 10 of which had switched locations, successfully detects all errors with 5.6% false detection rate, even with poor data quality. The MDS map can also be used for other applications such as detecting sensor malfunctions.

Keywords: loop detectors; advanced traffic management systems; artificial intelligence; multidimensional scaling (MDS)
1. INTRODUCTION

Accurate traffic data acquisition is essential for effective traffic surveillance, which is the backbone of advanced transportation management and information systems (ATMIS). Inductive loop detectors, video image processing and laser-based systems are used to measure fundamental traffic parameters: speed, volume, and occupancy.

These sensors are often deployed on a large scale, and a single operating unit may employ hundreds or even thousands of sensors. For example, Los Angeles County (Caltrans District 7) lists 1,323 mainline loop detector stations in 18 freeways, according to the Freeway Performance Measurement Project (PeMS) website (1).

Such a large number of sensors poses challenges of data quality and management. Individual sensors can fail permanently or intermittently. Thus, the measurements from individual sensors need to be automatically diagnosed for errors and corrected, if possible. This problem has been studied widely; a recent work is the PeMS algorithm developed by Chao et al (2).

A different problem arises in the maintenance of the correct geographical and spatial configuration information of the sensors. This is a simple task in a network with few sensors. But once the number of sensors exceeds a few dozen, even simple bookkeeping is error-prone, as sensor repair, installation, rewiring, etc, require continuous updating of configuration information. We recently observed that traffic management centers (TMC) might have incorrect configuration information about the location of loop detector stations. In one example, which is used in our study, 20 out of 128 loop detector stations (15.6%) had incorrectly assigned labels.

The most common error for loop detectors is switching of direction; for example, detectors in the eastbound direction are labeled westbound. This error can occur because loop detectors in the two freeway directions share the same cabinet. Such an error can be detected by careful investigation of the wiring/labeling of detector cabinets in the field, but such an investigation is costly and prone to human error.

Considerable attention has been devoted to understanding individual sensor data errors, and automated algorithms like (2) are available to detect and correct them. By contrast, little effort has focused on detecting sensor configuration errors, even though they, too, can invalidate subsequent analysis. For example, if a ramp-metering algorithm were to be executed based on switched detector information, the result would be devastating.

This paper presents an automatic and a semi-automatic method for detecting configuration errors. The method uses archived data from the sensors themselves to detect these errors. The method relies on the fact that flow, occupancy, or speed measurements between a pair of loops that are spatially close show higher correlations than when they are far apart. The data are used to construct a similarity matrix or distance matrix. A scoring scheme applied to this matrix then quantifies the chances of sensor mislabeling. Alternatively, one can use the visual representation method called multidimensional scaling (MDS) to identify suspicious sensors, leading to a semi-automatic procedure.
In an operational context, the method offers a low-cost quality control check. Such a check could be performed periodically, and used to dispatch field personnel to check the suspect configurations revealed by the method.

Details of the algorithms are presented in Section 2. In Section 3, the algorithm is applied to a stretch of freeway in the San Francisco Bay Area. Section 4 summarizes the results and concludes the paper.

2. METHODS

Denote the time series of measurements from sensor $a = 1, \ldots, A$ by

$$x_a(t), \ t = 1, \ldots, T,$$

in which $t$ indicates the time of measurement. The measurement could be flow, occupancy, or speed, or it could be a vector of these measurements. If two locations $a$ and $a'$ are close, and there are no major disturbances like freeway interchanges between them, $x_a(t)$ and $x_a'(t)$ should show a similar temporal pattern. The further apart they are, the less similar the patterns will be. We quantify the similarity between two locations by a measure

$$S(a,a'), \ a, a' = 1, \ldots, A.$$  

The measure could be the correlation coefficient

$$r(a,a') = \frac{\sum_{t}(x_a(t) - \bar{x}_a)(x_a'(t) - \bar{x}_{a'})}{\left[\sum_{t}(x_a(t) - \bar{x}_a)^2 \sum_{t}(x_a'(t) - \bar{x}_{a'})^2\right]^{1/2}}, \text{ with } \bar{x}_a = \frac{1}{T} \sum_{t} x_a(t),$$

or the negative of Euclidean distance

$$-\left[\sum_{t}(x_a(t) - x_a'(t))^2\right]^{1/2}.$$  

Distances can be derived as negative or inverse similarities. Thus, we expect

$$d(a,a') \sim d_r(a,a'),$$

in which $d_r(a,a')$ is the ‘traffic distance’ between two sensors. We define the traffic distance loosely as the difference in traffic patterns at two locations. We expect it to be proportional to the physical distance if sensors are located on the same section of a freeway. But this is not always the case: Measurements at sensors located at the same postmile of freeway but in opposing directions would show little or no similarity.

**Multidimensional Scaling**

If the distances $d(a,a')$ are exactly proportional to the physical distances between spatial coordinates, it is possible to reproduce the original configuration solely from the distance matrix $D = (d(a,a'), a,a' = 1, \ldots, A)$ by a method called multidimensional scaling (MDS). MDS is a statistical methodology that produces a configuration of points from the matrix of pairwise distances between these points. Basically, one tries to find 2-dimensional coordinates $y_a, a = 1, \ldots, A$ by minimizing the “stress” function
\[ \sum_{a,a'} (d(a,a') - \|y_a - y_{a'}\|^2), \]  

or some variant of it, by a stochastic gradient algorithm or regular eigenanalysis (3). The result of this embedding is also called MDS map.

Clearly one can reproduce coordinates only up to translation and rotation even with the exact distance matrix. Furthermore, since the relationship (6) is approximate, we cannot exactly reproduce the spatial coordinates of sensors from the distance matrix. Nonetheless, the output provides an effective visual representation of the sensor correlations, which enables one to identify configuration errors.

To summarize, an MDS map is constructed in three steps as follows:

1. For sensors \( a = 1, \ldots, A \), collect a time series of measurements \( x_a(t), \ t = 1, \ldots, T \), forming a \( T \times A \) data matrix.
2. Construct the \( A \times A \) traffic similarity matrix \( S(a,a') \) (2), or an equivalent distance matrix, from the data matrix, using (3) or (4).
3. Apply MDS to the similarity or distance matrix to find the MDS map \( y_a, a = 1, \ldots, A \).

For step 3 and also for visualization of the MDS map, public domain statistical packages such as XGvis (3) or R (4) are available for various computer platforms.

**Scoring scheme**

Potential sensor configuration errors can be automatically identified as follows. We choose the size of the neighborhood \( K \). We quantify the configuration accuracy score of a sensor \( a \) by the number of \( K \) nearest (in terms of distances according to the potentially incorrect configuration table) sensors that are indeed among \( K \) nearest (in terms of the similarity or distance) sensors to \( a \), divided by \( K \). If we let

\[ A = \text{the set of } K \text{ sensors that are closest in the configuration table}, \]
\[ B = \text{the set of } K \text{ sensors that have highest correlation}, \]

the score is

\[ s_a = \frac{|A \cap B|}{K}, \]

in which \( |A \cap B| \) is the number of sensors in \( A \cap B \). The score ranges between 0 and 1. Incorrectly labeled sensors will have scores close to 0 and those with correct labels will have higher scores.

**3. ANALYSIS**

We apply the algorithm to 128 loop detector locations along I-80 in the San Francisco Bay Area. Each location has sensors for either or both eastbound and westbound traffic. At each location, there can be up to four sensors (one sensor per lane) for each direction, with lane 1 being a time-activated high occupancy vehicle (HOV) lane.
An independent study had revealed that the TMC had incorrect detector configuration information. In particular, the directions were switched for some locations. The study concerned estimation of travel times on I-80. Anomalous results led to a close investigation of the data and the conjecture that some detectors were switched. Caltrans personnel verified the conjecture.

We extracted 288 5-minute volume and occupancy data for May 24, 2003. Figure 1 shows the space-time distribution of flow and occupancy with the incorrect labels. From the plots, especially the one for occupancy, one suspects errors in the loop configuration information, but identifying the source of errors from these plots is not easy.

To apply our methods, we first calculate the similarity between two detectors by their correlation coefficient for 5-minute volume or occupancy. The ‘traffic distance’ between two detectors is computed as one minus the correlation, so the pairwise distance matrix $D$ is

$$D = 1 - R,$$

in which $R$ is the pairwise correlation matrix, and 1 is the matrix all of whose elements are 1.

We then use classical metric MDS (5) to embed the detectors in a 2-dimensional plane, with the “cmdscale” command of R (4). We will call the MDS based on correlation of occupancy measurements occupancy-MDS and the one based on volume measurements as volume-MDS.

The two MDS maps are shown in Figure 2, 3 and 4. We observe two clusters, separated by boundaries near the vertical line ($x = 0$) for volume-MDS and a straight line for occupancy-MDS. We expect eastbound traffic at different detectors to behave similarly, and westbound traffic to behave similarly. But westbound and eastbound traffic should behave differently from each other, because traffic in the westbound direction (towards San Francisco) has a strong AM peak and eastbound traffic has a strong PM peak. Therefore we expect eastbound detectors to lie on one side of the line and westbound detectors to lie on the other side.

But for both MDS several detectors are on the ‘wrong’ side: Many eastbound detectors (black squares) occupy the right half of the plane and westbound detectors (white squares) occupy the left half. The group of points in the outer shell in the volume-MDS correspond to detectors in the HOV lane (lane 1), as is clear from the Figure 2. They are far from the non-HOV lane detectors because their flow characteristics are quite different.

The lines joining sensors in different lanes (Figure 3) at the same location lie on one or the other side of the vertical line ($x = 0$), which suggests that the mislabeling occurs for all lanes. Volume-MDS and occupancy-MDS both show good clustering. We can identify mislabeling of directions visually from the figures. Indeed by exchanging the black and white squares for several lines, we end up with most white squares on the westbound, right side and black squares on the eastbound, left side (see Figure 7 for an example of such correction). The detectors visually identified in this way are judged mislabeled and this amounts to a semi-automatic method for detecting configuration errors.

We now turn to automatic methods for detecting configuration errors. As a naïve classifier in this MDS-embedded plane, we use the straight lines $x = 0.085$ and $y = -0.1 + 2.1x$ as the cluster boundary for volume-MDS and occupancy-MDS respectively (Figure 5). If two or more lanes at
a particular location belong to the wrong side of the boundary, the location is declared mislabeled. The result from this naïve ad-hoc classifier is shown in Table 1.

We also apply our scoring algorithm, with a neighborhood size of \( K = 40 \). Figure 6 shows the clustering of scores and a threshold of 0.3 can be used for classification. Smaller scores mean higher chances of mislabeling. We observe that flow-based score is more conservative than the occupancy-based score.

So far, we have four separate classification schemes, each a combination of naïve or scoring classification and volume-based or occupancy-based. We take locations that are judged mislabeled by all four schemes as our final decision. See Table 1.

Independent inspection shows that most decisions are ‘true positives’, i.e. directions of these loops are indeed switched. There are ten locations 22,23,26,35,39,41,46,51,56, and 57. There are three false positives at locations 1, 28 and 63. Those false positive errors of the algorithm are due to 1) poor data quality, and 2) a large number of locations with switched directions. There are no false negatives. To summarize, the detection rate is 100% and the false detection rate is 5.6% (three false negatives divided by the total number of correctly labeled detectors, which is 54).

Figures 7 and 8 show the result of MDS embedding after the labels are corrected by the proposed automatic method. Loops for the same traffic direction cluster together in these plots. The space-time distribution of flow and occupancy shown in Figure 9 also confirms that the algorithm corrected mislabeling fairly well.

**Other types of mislabeling**

There are other types of detector mislabeling, including wrong lane and wrong station. We first consider mislabeling of lanes, which can happen at the detector cabinet, just like direction switching. The good separation of HOV clusters in the MDS map of Figure 7 lets us infer from our data that lane 1 (the HOV lane) is nowhere mislabeled. For non-HOV lanes (lane 2, 3, 4), it is not clear if there is a lane switching. Recall that inspection of the MDS map can identify mislabeling if (a) detectors in the same group share similar traffic characteristics and (b) detectors in different groups have different traffic characteristics. Those conditions are met for HOV vs. non-HOV detectors. But it is unlikely that such conditions are met for, say, lane 2 vs. lane 3 detectors. That is, at the same station, traffic flow characteristics in lane 2 and 3 are quite similar, indeed more similar to each other than lane 2 traffic characteristics at two nearby stations. Therefore, to identify mislabeled non-HOV lanes, one should develop a more targeted algorithm that exploits such signals as higher volume and smaller speed for outer lanes for example.

Similar suggestions can be made concerning identification of mislabeled locations, which is less likely to occur than mislabeled lanes or direction. Labels switched for two stations that are very far apart could be identified from the MDS map, since their traffic flow characteristic would be very different from their putative neighbors. But if the switching occurred for two very close stations, it would be hard to identify it by any algorithm including MDS.
To summarize, identification of configuration errors from mislabeling by inspection of MDS requires the error induced by the mislabeling to be substantial. Mislabeling of HOV lane or switching of directions is very easy to identify from MDS map, but other types of mislabeling like switching of non-HOV lanes or switching of nearby locations may not be readily visible in MDS map. Practitioners need to be aware of these limitations when applying MDS to field data.

4. CONCLUSION

To identify potential sensor mislabeling, we presented the MDS method for embedding sensors in a plane, in which distances between two sensors reflect the similarity or correlation of their measurements. Visual inspection of the embedded sensor map clearly reveals the error in sensor labeling, which leads to a semi-automatic process of detecting and correcting sensor mislabeling. To correct switched sensors, one could form an ad-hoc classification line in the embedded sensor map, which we called the naïve classification rule. We also proposed an automatic scoring scheme that identifies sensor mislabeling from the pairwise distance or similarity matrix.

The algorithm, when applied to I-80 data with many location-switching errors, successfully detects all errors with 5.6% false detection rate. MDS maps also provide rich visual information about more abstract configuration properties of sensor measurements, such as separation of HOV lanes from non-HOV lanes. This implies that the proposed method is an accurate and effective tool for routine integrity checks of sensor configuration information.

The algorithm is potentially useful for other applications. One of them is sensor malfunction detection. Many sensor malfunction detection algorithms are based on high correlation between neighboring sensors (2), and that is exactly the feature that is visualized in MDS map. If a sensor malfunctions for a significant period of time, the resulting divergence of its measurements from those of its spatial neighbors will be apparent in MDS map, suggesting that a MDS map can be a useful visual tool for periodic monitoring of sensor data quality.

5. ACKNOWLEDGEMENT

This study is part of the PeMS project, which is supported by grants from Caltrans to the California PATH Program. We are very grateful to Caltrans Traffic Operations engineers for their support.

The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views of or policy of the California Department of Transportation. This paper does not constitute a standard, specification or regulation.

REFERENCES


LIST OF TABLES

TABLE 1 Locations with Switched Directions Detected by Various Algorithms ................. 10

LIST OF FIGURES

FIGURE 1 Space-time distribution of median-over-lanes flow and occupancy. Left half corresponds to the eastbound traffic moving from bottom to top, and the right half to the westbound traffic, which moves from top to bottom. ............................................................... 11

FIGURE 2 Volume MDS embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic. For each direction, points corresponding to lane 1 (HOV lane) are connected from the beginning to the end of the freeway section by dashed lines, and those corresponding to lane 2 by solid lines. ................................................................. 12

FIGURE 3 Volume MDS embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic. This is the same plot as FIGURE 2, except that lines connect points corresponding to loop detectors at the same station ........................................ 13

FIGURE 4 Occupancy MDS Embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic. Either points corresponding to the same lane are connected (left) or those corresponding to the same station are connected (right) .............. 14

FIGURE 5 Suggested ad-hoc classification boundaries (dashed straight lines) for volume-MDS (top) and occupancy-MDS (bottom) ....................................................................................... 15

FIGURE 6 Relationship between volume-based (over-the-lane) median scores against occupancy-based medians scores. Locations are indicated by numbers in black and gray colors, corresponding to eastbound and westbound direction .............................................. 16

FIGURE 7 Volume MDS Embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic, after directions are corrected .................................................. 17

FIGURE 8 Occupancy MDS Embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic, after directions are corrected ........................................ 18

FIGURE 9 Space-time distributions of median-over-lanes flow and occupancy after sensor mislabeling is corrected by the proposed algorithm. Left half corresponds to the eastbound traffic and the right half to the westbound traffic. ................................................................. 19
TABLE 1 Locations with Switched Directions Detected by Various Algorithms

<table>
<thead>
<tr>
<th>Method</th>
<th>Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve classifier based on volume-MDS</td>
<td>1 11 22 23 26 28 33 35 39 41 46 51 56 57 63</td>
</tr>
<tr>
<td>Naïve classifier based on occupancy-MDS</td>
<td>1 8 9 22 23 25 26 28 32 33 35 39 41 46 51 52 56 57 63</td>
</tr>
<tr>
<td>Volume-based scoring with K=40</td>
<td>1 10 11 22 23 26 28 35 39 41 46 51 52 56 57 63</td>
</tr>
<tr>
<td>Occupancy-based scoring with K=40</td>
<td>1 8 9 10 11 12 13 22 23 26 28 32 33 35 36 39 41 46 48 51 52 56 57 63 64</td>
</tr>
<tr>
<td>The intersection of the four algorithms above</td>
<td>1* 22 23 26 28* 35 39 41 46 51 56 57 63*</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>22 23 26 35 39 41 46 51 56 57</td>
</tr>
</tbody>
</table>

* - false positive
FIGURE 1 Space-time distribution of median-over-lanes flow and occupancy. Left half corresponds to the eastbound traffic moving from bottom to top, and the right half to the westbound traffic, which moves from top to bottom.
FIGURE 2 Volume MDS embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic. For each direction, points corresponding to lane 1 (HOV lane) are connected from the beginning to the end of the freeway section by dashed lines, and those corresponding to lane 2 by solid lines.
FIGURE 3 Volume MDS embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic. This is the same plot as FIGURE 2, except that lines connect points corresponding to loop detectors at the same station.
FIGURE 4 Occupancy MDS Embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic. Either points corresponding to the same lane are connected (left) or those corresponding to the same station are connected (right).
FIGURE 5 Suggested ad-hoc classification boundaries (dashed straight lines) for volume-MDS (top) and occupancy-MDS (bottom).
FIGURE 6 Relationship between volume-based (over-the-lane) median scores against occupancy-based medians scores. Locations are indicated by numbers in black and gray colors, corresponding to eastbound and westbound direction.
FIGURE 7 Volume MDS Embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic, after directions are corrected.
FIGURE 8 Occupancy MDS Embedding of loop detectors in eastbound (black squares) and westbound (white squares) traffic, after directions are corrected.
FIGURE 9 Space-time distributions of median-over-lanes flow and occupancy after sensor mislabeling is corrected by the proposed algorithm. Left half corresponds to the eastbound traffic and the right half to the westbound traffic.