Toward Management and Imputation of Unavailable Data in Online Advanced Traveler Information Systems

Rafael J. Fernández-Moctezuma, Kristin A. Tufte, David Maier, and Robert L. Bertini

Abstract—Advanced Traveler Information Systems (ATIS) deliver information derived from Intelligent Transportation Systems data. ATIS continue to emerge and grow in popularity, as some of these systems are public and freely available. ATIS performance and perceived usefulness depends, among several design factors, on the reliability and quality of its data sources, especially when the systems are reporting current conditions. We present a sample scenario in which an inductive loop detector station along a freeway stops transmitting data, and an ATIS service presents an estimated state of the system instead of showing that no data is available. The estimation is based on sources of available and correlated data. Our first approach to implement an estimation framework consists of deriving a parametric description of the correlation from archived historical data. An example application is provided as well as suggestions for future research.

I. INTRODUCTION

ADVANCED Traveler Information Systems (ATIS) are important components of Intelligent Transportation Systems (ITS) strategies. ATIS present consumers with information measured from various sources, such as probe vehicles, AVL equipped platoons, and the transportation infrastructure, usually adding value in processing and interpreting raw measurements from ITS and disseminating content through diverse channels, such as the Web, PDAs, or ultimately in-vehicle navigation systems [1]. Presentation products such as web-accessible speed maps provide travelers with a representation of the current (or forecasted) state of the system; such products may be provided by companies or public agencies. Examples of these products include the speed map available at TripCheck.com, sponsored by the Oregon Department of Transportation (ODOT) [2], Google Maps’ traffic information [3], the PORTAL system from Portland State University [4], and Traffic.com [5]. Images from Google Maps and Traffic.com are shown in Fig 1; notice how congestion information for the Portland Oregon Metropolitan Area is displayed in different ways. In general, ATIS obtain data from one or more sources, including online measurements, and update the representation of the state of the system as new information becomes available; such state updates typically occur at predefined intervals. In speed map applications, color schemes provide a quick and easy way to interpret the state of the system. For example, if a sensor reports that the average speed has dropped from 30 mph to 20 mph, the sensor’s area of influence is re-colored from yellow to red. The color schemes and cut-off points vary among different systems and providers, as seen in Fig 1.

Inductive loop detectors remain the most common sensor...
deployed for traffic system state measurement [1]. While widespread, they are still subject to several conditions which affect the quality and availability of data [3]. Because any ATIS relies on measurements from these sensors as its foundation, the quality of the information presented to the user is important along two dimensions. First, the reliability of the measurements should be understood and somehow communicated to the user. Also, data may be missing due to sensor or communications malfunctions. Furthermore these data gaps vary from day to day. For example, some areas of a speed map might show “no data” on Tuesday while they contained valid data on Monday. We illustrate this scenario with the TripCheck speed map in Fig 2.

The motivation for our work is to develop a technique that will enable ATIS providers to deliver as much useful information to the users as possible in the presence of data gaps. In particular, our work focuses on failures in the freeway sensor infrastructure in Portland Oregon. In Section II we state the problem of data loss in ATIS, and address it in terms of predicting the expected values of unavailable data based on measurements from nearby sensors, exploiting the benefits of having access to an Archived Data User Service (ADUS). We approach the task of expressing relationships between stations based on regression analysis of archived data, since this allows us to focus on choosing a functional form and fitting its parameters, without having to rely on further theoretical explanations of freeway dynamics. In Section III we provide preliminary results for a simulated data loss. We further discuss the experimental results and indicate future work in Section IV.

<table>
<thead>
<tr>
<th></th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>August</th>
<th>September</th>
<th>October</th>
<th>November</th>
<th>December</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Stations</td>
<td>31%</td>
<td>30%</td>
<td>27%</td>
<td>27%</td>
<td>28%</td>
<td>27%</td>
<td>34%</td>
<td>32%</td>
<td>44%</td>
<td>38%</td>
<td>31%</td>
<td>33%</td>
</tr>
<tr>
<td>US 26 E</td>
<td>2%</td>
<td>4%</td>
<td>1%</td>
<td>4%</td>
<td>19%</td>
<td>1%</td>
<td>8%</td>
<td>1%</td>
<td>10%</td>
<td>5%</td>
<td>6%</td>
<td>13%</td>
</tr>
</tbody>
</table>

II. PROBLEM STATEMENT

Data loss due to sensor unavailability or malfunction can be observed in our freeway systems. For example, in the Portland Oregon system, not one month during 2006 had complete measurements as shown in Table 1. One can see how data loss directly impacts final products such as speed maps. If one assumes that the measured speed at one point (representing a freeway segment by extrapolation) is related in particular ways to the measured speeds on neighboring segments, it should be possible to find an expression for this relationship based on historical information. We know from traffic theory that this assumption is true, however, a challenge arises in identifying relationships between well correlated information. While contiguous stations may measure similar dynamics with predictable relationships, stations far apart from one other may not necessarily be related—in fact, a number of events may occur between two distant points on a freeway, such as an active bottleneck in between [7] or a busy off ramp. It is also possible that some events affecting the dynamics will never be measured due to inadequate sensor spacing, making approaches such as simple interpolation questionable.

In the case of unavailable or unreliable data it is necessary to perform imputation in order to provide an estimate of the current conditions. A naïve approach for dealing with unavailable data consists of continuously reporting the last known good measurement. However, if a sensor goes offline before a change in regime (e.g., from free flow to congestion), the ATIS will report an invalid state during the

![Fig. 2. Two screenshots of speed maps of the Portland Metropolitan Area Freeway System presented by TripCheck, showing differences in data availability for various segments (circled).](image-url)
Regression is a method of deriving a mathematical relationship between a set of inputs; in our work, regression is used to derive a function relating speeds of nearby stations. Assuming that the measured speeds of nearby stations are related, regression can be used to find a function \( f \) to estimate the missing quantity \( \hat{S}_B \) looking at the available measured speeds from nearby stations \( A \) and \( C \), namely:

\[
\hat{S}_B = f(S_A, S_C)
\]  

As mentioned before, archived historical measurements from the Portland Oregon Freeway System exist in the PORTAL ADUS [4]. We consider using regression techniques on archived data to parameterize a model function \( f \), which will enable the production of on-line estimates. Evaluating a function on-line is considerably cheaper than going through entries in an archived data. For complete system deployment, a regressor function such as (1) must be defined for every station. The architectural change in the final software product would require a simple decision mechanism: when online data for a particular station was not available, the regression function associated with that station would be evaluated. This approach limits stations at the beginning and end of a given freeway to look only at one neighboring station. In general, the relationship between stations may be linear, or contain un-captured events, which would suggest a nonlinear relationship. We explore both possibilities with two methods:

A. Linear Regression

Under the assumption that the relationship between stations can be expressed as a linear function, a joint Gaussian probability distribution can model the relationship between the speeds at three locations. A multidimensional Gaussian distribution is parameterized with two statistics: the mean vector \( \mu \), and the covariance matrix \( \Sigma \), both of which can be estimated from historical data. If we represent the inputs (stations \( A \) and \( B \)) with the variable \( x \) and the target with the variable \( y \), the parameters can be written in block form as

\[
\mu = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}
\]

(2)

\[
\Sigma = \begin{bmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx}^T & \Sigma_x \end{bmatrix}
\]

(3)

where \( \sigma_y^2 \) is the variance of the random variable \( y \), \( \sigma_{yx} \) is the covariance of the random variables \( y \) and \( x \), and \( \Sigma_x \) is the covariance matrix of the inputs \( x \). One can obtain a conditional probability density from Bayes theorem, which will also be Gaussian. Since the expected value of a random variable distributed as Gaussian is the mean, we can evaluate the conditional mean for a given input \( x = x_i \). This is a simple linear regressor, expressed as

\[
\hat{\mu}_{y|x} = \mu_y + \sigma_{yx} \Sigma_x^{-1} (x_i - \mu_x)
\]

(4)

B. Nonlinear regression

Since inadequate spacing may result in a more complicated relationship between the stations, we also consider modeling this relationship as nonlinear. Artificial
Neural Networks can be a good tool for nonlinear regression [8].

Our proposed architecture for regression consists of three layers: input, hidden, and output. Just as with the linear regressor, the inputs are the measured speeds at nearby stations A and C and the output is the estimated speed at station B. The hidden layer contains several nodes, each with a nonlinear activation function. The outputs of the hidden functions are linearly combined to produce the output, as illustrated in Fig. 4.

In contrast with the linear regressor, this approach requires additional architectural decisions, such as determining an appropriate number of hidden nodes $n$, as well as setting the weights for each choice of $n$. Algorithms for fitting these parameters are well documented in the Machine Learning community [8, 9] and generally available software, such as Netlab [10], already contain implementations of these techniques.

Setting the weights is achieved using the Backpropagation algorithm [8] for each choice of $n$. In order to select the number of hidden nodes, we split our available historical data in two sets: training and evaluation. We iterate through a range of candidate values for $n$, fit the parameters for each choice using the training set, and we measure the performance of the model on the evaluation set as Mean Square Error. We then select the choice of $n$ that yields the least error.

### III. Experimental Results

We obtained archived data from the Portland Regional Transportation Archive Listing (PORTAL) [4], in particular, one week of five minute aggregated data for a morning peak period (6:00 a.m.–10:30 a.m.) from three sensor locations across US 26 Eastbound, shown in Fig 5, with data from October 16–20, 2006. The stations selected are detailed in Table 2.

<table>
<thead>
<tr>
<th>Milepost</th>
<th>Intersection</th>
<th>Spacing from previous station (in miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70.9</td>
<td>SW Canyon Rd</td>
<td>1.59</td>
</tr>
<tr>
<td>71.37</td>
<td>Skyline Rd</td>
<td>0.47</td>
</tr>
<tr>
<td>73.62</td>
<td>I-405 Count Station</td>
<td>2.25</td>
</tr>
</tbody>
</table>

An analysis of the measured speed dynamics from station to station within the same time interval was conducted. It was determined that predicting conditions at Skyline Rd. from measurements from the I-405 and Canyon Rd. stations was possible. The relationship between the selected stations was examined with two tools: time series plots to find clear trends, and scatter plots to look for potential structure. Sample plots from the training set are shown in Fig 6 and Fig 7. In Fig 6 one can appreciate evidence of correlation, as the drops and increases of speed are maintained in these two stations. The scatter plots in Fig 7 offer visual cues for candidate regression features: one can identify visually the existence of a functional form.

<table>
<thead>
<tr>
<th>Canyon Rd, Station, Oct. 16, 2006</th>
<th>Skyline Rd, Station, Oct. 16, 2006</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Fig. 6. Speed time series for Oct. 16, 2006 for Canyon Rd. and Skyline Rd. stations." /></td>
<td><img src="image2" alt="Fig. 7. Scatter plots of measured speeds." /></td>
</tr>
</tbody>
</table>

Our proposed experiment assumed the existence of historical data from Monday through Thursday, and a total loss on the station at Skyline Rd. during Friday. It would not be an accurate measurement if we had mixed data points from the entire week and randomly selected points for evaluation—the parameters of the regression functions would have been set while exposed to potentially contiguous...
measurements, therefore inducing a more optimistic result. By simulating the data loss in the future, we maintain the initial assumption of only having been exposed to historical information while training. Fig 8 shows the speed time series for the true value and both predictions. Notice how both regressors correctly follow the measured trend in speed. Horizontal lines highlight the limits between speed categories in TripCheck, which correspond to high congestion (< 25 mph), moderate congestion (25–50 mph), and free flow (> 50 mph.) The motivation for predicting speed rather than “color” is to provide flexibility for future changes in the interface, such as different cutoffs for coloring.

For the current sample application we quantify the number of times a particular speed regime was colored correctly, and the number of times it was colored incorrectly. We show these results in Table 3 as a confusion matrix. Under this simulated data loss event, the linear predictor had an accuracy of 80%, while the nonlinear predictor had a higher accuracy of 89%. In addition to prediction accuracy as measured in the three speed categories considered in the speed map, we notice that for this simulation the errors made occurred within one category, i.e., we did not predict free flow (green) when the observed state was congested (red).

IV. DISCUSSION AND FUTURE WORK

The preliminary results reported in this work encourage us to study larger segments. It is our intention to direct attention to the quality of final products, and we believe dealing with data loss effectively will improve both user perception and quality of information presented by ATIS consumer products. Future work will explore incorporating additional inputs for prediction, such as time-delayed measurements, in addition to exploring more choices of functional forms for nonlinear regression. User surveys will indicate the usefulness of having estimated segment states as compared to missing data in speed maps.

The speed map in TripCheck currently updates its state approximately every 2 minutes. The lowest resolution available from ODOT through PORTAL is 20 seconds, and the next level of aggregation available through PORTAL is 5 minutes. It is known that aggregation attenuates the noisy nature of low resolution measurements. Future work should determine a lower level of aggregation that still provides good results but is closer to user needs.

<table>
<thead>
<tr>
<th></th>
<th>Linear</th>
<th>Nonlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>14 5</td>
<td>Red 17 2</td>
</tr>
<tr>
<td>Yellow</td>
<td>0 9 0</td>
<td>Yellow 2</td>
</tr>
<tr>
<td>Green</td>
<td>0 6 21</td>
<td>Green 0 2</td>
</tr>
</tbody>
</table>

TABLE 3

CONFUSION MATRICES

REFERENCES