EMPIRICAL ANALYSIS OF MISSING DATA ISSUES FOR ATIS APPLICATIONS: TRAVEL TIME PREDICTION

Jianwei Wang, Ph.D., wjwish@umd.edu
Nan Zou, Ph.D., nanzou@umd.edu
Gang-Len Chang, Ph.D., Professor, gang@umd.edu
Department of Civil Engineering
The University of Maryland, College Park, MD 20742
Tel: (301) 405-2638

Number of words: 4376
Number of tables and figures: 12
Total number of words: 12×250 + 4376 = 7376

Submitted to the 87th meeting of the Transportation Research Board for presentation and publication
Revised on Oct 2007

ABSTRACT

As reported in the literature for Intelligent Transportation System (ITS) applications with traffic detectors, various missing data patterns are frequently observed in such systems and may dramatically degrade their performance. This study presents two imputation approaches for contending with the missing data issues in travel time prediction. The first model is based on the concept of multiple imputation technique to directly predict the travel times under various missing data patterns. The second model that serves as the supplemental component is to estimate the missing detector values using neighboring detector data and historical traffic patterns. Both models have been incorporated with reliability indicators so as to assess the quality of imputed data and its applicability for use in prediction. The numerical example based on 10 roadside detectors on I-70 in Maryland has demonstrated that both developed models outperformed existing methods and offers the potential for field implementation.

KEYWORD:

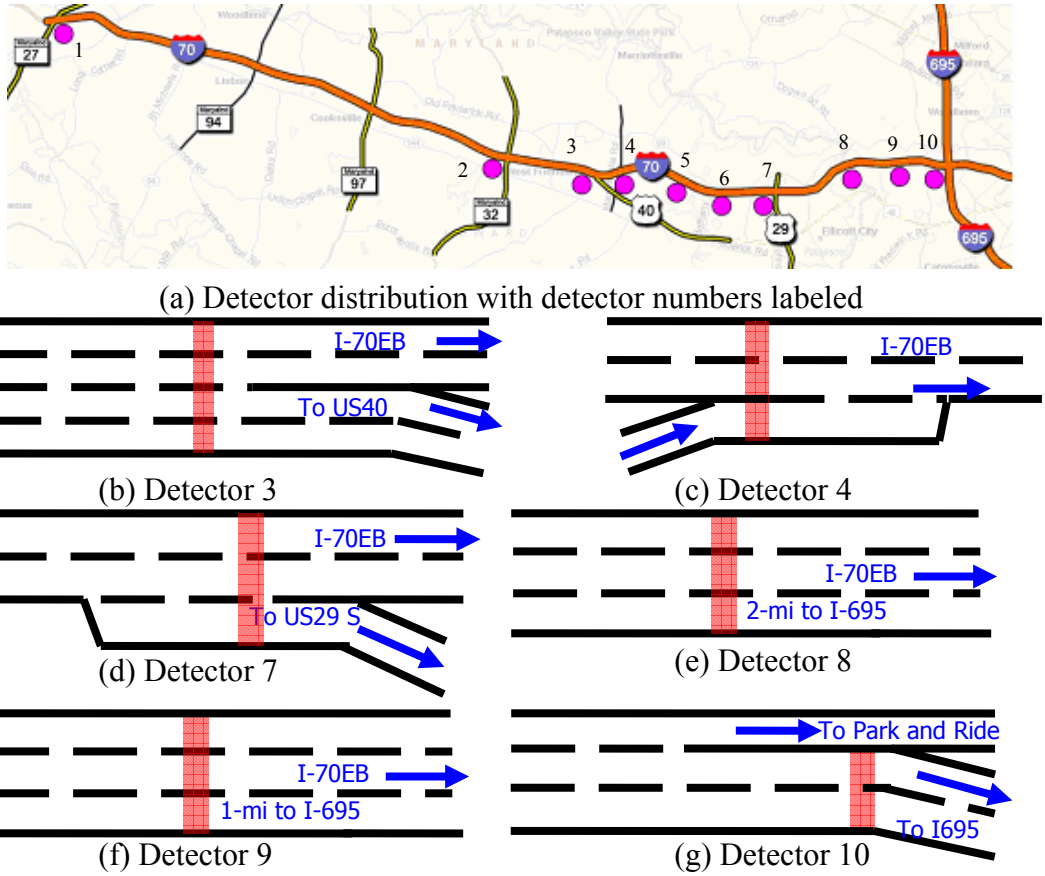
Missing Data, Multiple Imputation, Travel Time Prediction

INTRODUCTION

As is well recognized, traffic detectors for freeway monitoring or any Intelligent Transportation System (ITS) applications may not function as reliably as expected and often produce various patterns of missing data that consequently degrade the quality of control operations. The inevitable encountering of the missing data issue has also further complicated the challenging travel time prediction task, especially when only sparsely distributed detectors are available for collecting real time traffic conditions. In view of the quality of detectors in the existing market and their associated communication issues, it seems essential for any potentially deployed ITS system to have a function that can be effective in contending with the missing data.

In most travel time prediction systems, one can attribute most common data missing to either data delay or data loss. Most communication errors often contribute to short-term missing data, whereas device failures, such as the traffic detector or the data storage device, potentially result in a long-term missing data.

The dataset from 10 detectors illustrated hereafter was taken from the field demonstration project of a real-time travel time prediction system with widely spaced detectors (2 miles per detector) between February 9th and August 2nd, 2006, which contain various commonly seen patterns of missing data. The demonstration project, named Automated Real-Time Travel Time Prediction System (ARAMPS), is located on a 25-mile stretch of I-70 eastbound from MD27 to I-695, which includes 7 interchanges and 10 traffic detectors (Fig. 1).



(a) Detector distribution with detector numbers labeled

(b) Detector 3

(c) Detector 4

(d) Detector 7

(e) Detector 8

(f) Detector 9

(g) Detector 10

Figure 1 Distribution of detectors in ARAMPS

Example Long-Term Missing Data

Long-term missing data, which is defined as data with a missing rate of greater than 10%, has occurred quite often at each detector location during the project period of ARAMPS. Table 1 shows the distribution of the number of days when the daily missing data rate exceeded 10% (144 minutes) at each detector during the 116-day period of demonstration from February 9th to June 4th, 2006. In 39 days out of the total field demonstration period (i.e., 116 days), the ARAMPS prediction system experienced at least one detector that suffered a daily missing data rate exceeding 10%.

Table 1 A summary of missing data distribution by detector during the period experiencing more than 10% missing rate

Detector	1	2	3	4	5	6	7	8	9	10
Number of Days*1	3	2	2	3	9	8	14	2	11	10
Average Daily Availability (%)*2	30.0	25.5	25.5	57.7	47.6	63.7	73.4	0.0	20.3	26.9
Total Data Loss Duration (Hours) *3	50.4	36.0	36.0	31.2	112.8	69.6	88.8	48.0	211.2	175.2

*1: Total number of days in which the detector had a missing rate of more than 10%
*2: Average daily data availability during those days in which the detector had a missing rate of more than 10%
*3: Total duration of data loss during those days in which the detector had a missing rate of more than 10%

Example Short-Term Missing Data

Figure 2 illustrates the distribution of timestamps of the traffic data versus the arrival timestamps in the database from Detector 5 collected between 14:21 and 15:42 on June 2nd, 2006. There were several segments of data loss during this period. It clearly shows that the order of data recovery was "first missing first recovered" for all 3 short periods of data loss. Note that the order of data arrivals may vary with the system setup and configurations.

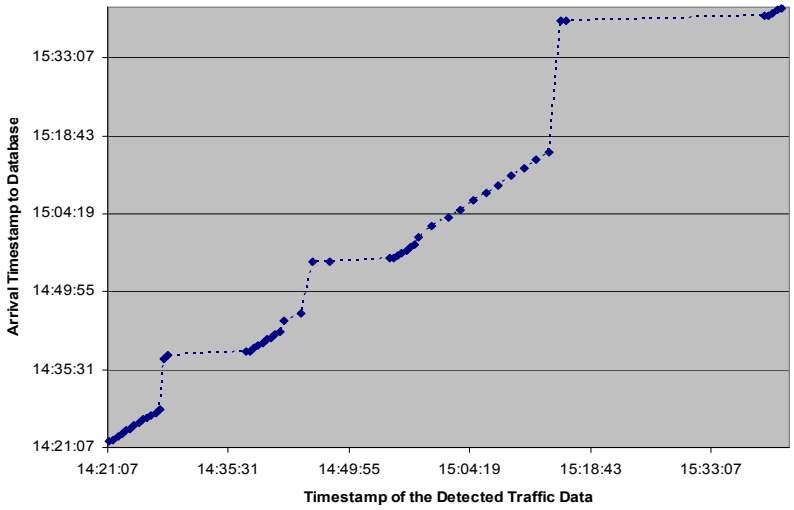


Figure 2 Timestamps of the detected traffic data vs. timestamps of data arrival at the database

Example Impacts of the Missing Data on Travel Time Predictions

To highlight the need of developing an effective imputation model for missing detector data, Figure 3 presents the performance of a travel time prediction model under 3 levels of missing data from detector 10. In this example, the missing data was imputed with the most widely used imputation method, mean substitution (MS) (*1*), in transportation systems. The resulting prediction error over a 12.3-mile segment with 15-minute missing data can be up to 140% of the actual trip times that are between 10.7 minutes to 22.5 minutes, and more than 6 times higher than the predicted results if without missing data.

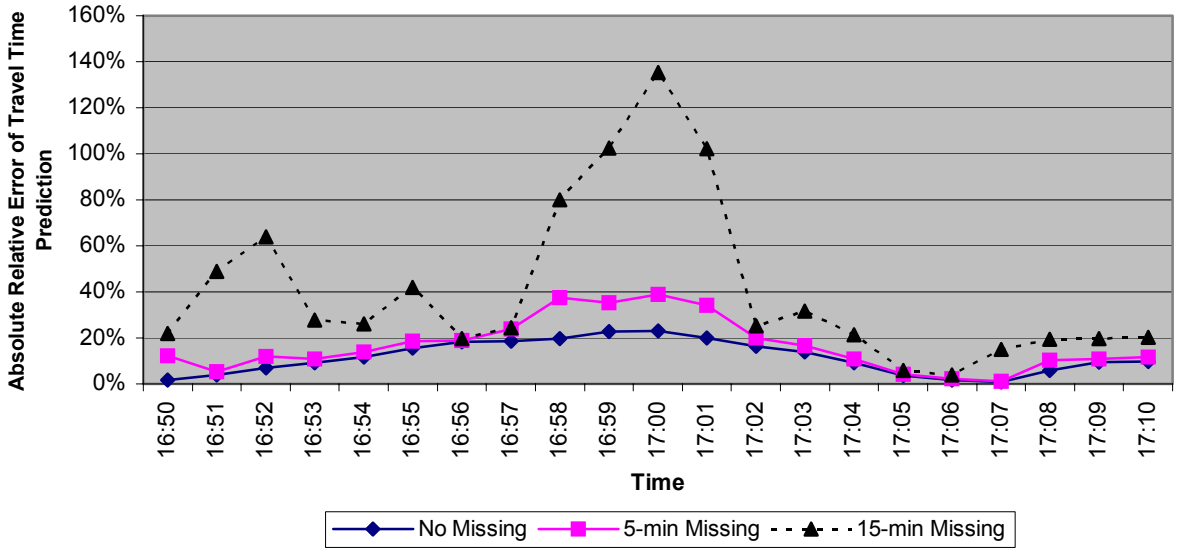


Figure 3 Absolute relative errors of travel time prediction with data missing at Detector 10 on June 20th, 2006

Hence, it is noticeable that a system for travel time prediction without an effective mechanism to deal with missing data may not fit the needs of real-world implementations. To contend with such an issue in travel time prediction especially with sparsely distributed detectors as in most highway networks, one needs to develop an imputation model that can have the following functions:

- Use site specific geometric features and traffic patterns to maximize its performance;
- Best use of all historical and related information to ensure the proper function of travel time prediction under various missing data scenarios; and
- Provide a reliability indicator so that the system can determine if the predicted travel time should not be displayed to avoid encountering unacceptable errors caused by the large mount of missing data.

This paper will first review available methods for missing data imputation in the literature, followed by the developments of two multiple imputation approaches for travel time prediction. The performance evaluation of the proposed models along with those in the literature using the dataset collected from ARAMPS will contribute the core of Section 4. Concluding comments are summarized in the last section.

LITERATURE REVIEW

Over the past few decades, many researchers in different technical fields — including econometrics, social sciences, biostatistics and transportation have devoted significant efforts in solving the missing data issue. Some early studies in contending with the missing data simply employed primitive approaches, such as case deletion and mean substitution (1, 2). Since 1970s, more researchers have recognized the complexity of the missing data nature and its impacts on the resulting performance of all employed models (1, 3-5). In fact, effective methods for dealing with the missing data issue may also vary with its pattern, the target applications, and the intended primary methods for prediction (1).

This study has categorized existing methods for handling the missing data estimation into two groups: single imputation, and multiple imputation models. Most recent studies in the literature indicated that multiple imputation approaches generally outperform the single-imputation methods that are widely used due to their convenience of implementation.

Single Imputation Methods

Most single imputation (SI) methods mainly impute the missing data from the means and distributions of the observable dataset. The Mean substitution (MS) is effective for these types of studies with emphasis on the mean of the data (1, 6). Hot-deck method (7) typically replaces the missing value with an estimation imputed from one or several similar data records using certain searching criteria. The expectation-maximization (EM) algorithm (1, 4) is an iterative estimation method that focuses on both the mean and the variance. The potential deficiency of such methods lies in the exercise of only one imputation that may not take full advantage of all available information embedded in the dataset.

Multiple Imputation Methods

In view of the deficiencies of most single imputation methods, some researchers have developed the Multiple Imputation (MI) techniques that aim to improve the imputation quality by incorporating the uncertainty of the missing data (1, 5). The core logic of multiple imputation methods is to estimate the same missing value m times ($m > 1$) with a simulated process (e.g., a Markov chain Monte Carlo simulation) to generate m complete datasets, and then analyzes the mean and variance of the estimators in these datasets to produce the final estimate. More specifically, let $Q = Q(X, Y)$ be denoted as the quality of the imputation, where X is the set of complete variables and Y contains the variables with missing data, the posterior distribution of Y_{mis} can then be determined by Eq. 1 (5):

$$\Pr(Q \mid X, Y_{obs}, R) \quad (1)$$

where Y_{mis} is missing values;

R is a $N \times p$ matrix with binary values indicating missing of Y; and

$Y = (Y_{obs}, Y_{mis})$

Rubin (5) showed that Eq. 2 and Eq. 3 can properly estimate the mean, $\hat{Q}$, and the variance, U, of the posterior distribution of completed data. Therefore, the simulation procedure incorporated in the multiple imputation framework is valid to estimate the posterior mean and variance of the missing values.

$$\hat{Q} = E(Q \mid X, Y_{obs}, Y_{mis}, R) \quad (2)$$

$$U = V(Q \mid X, Y_{obs}, Y_{mis}, R) \quad (3)$$

MI has been widely applied to the social sciences (8), biostatistics (9, 10), and transportation, and reported to generally outperform the single imputation methods in most data missing scenarios. Another advantage of the MI methods is their ability to evaluate the variance of the final imputed value.

Applications of the Data Imputation Methods in the Transportation Study

The use of missing data techniques has received an increasing attention in transportation over the past decade. Several methods for missing data treatment have already been widely used by practitioners and researchers in transportation applications. These methods include conditional mean substitution, regression models (such as interpolation), and time-series models (11). These transportation studies have focused mainly on replacing the missing values (flow, occupancy and/or speed) with imputed values so as to construct a complete set of traffic data (11-19). Most of these studies applied single imputation techniques, and only a few employed the multiple imputation methods (19). Some research also proposed the use of advanced prediction models, such as ARIMA, local weighted regression, and Neural Network models for missing data estimation in order to capture the temporal and spatial distributions of the detector data (15, 16). However, as reported in the literature, much remains to be done in terms of developing generalized and effective methods for imputing missing detector data.

NEW IMPUTATION APPROACHES FOR TRAVEL TIME PREDICTION

Grounded on the existing theories for missing data estimation, this study proposes two multiple imputation models, named M-1 and M-2, to supplement a real-time travel time prediction systems developed for use in a sparsely-distributed detection environment (20). The key feature of Model M-1 is to integrate the missing data imputation with the travel time prediction, and directly estimate the missing travel time with available information. In contrast, Model M-2 is focused on restoring the missing detector data used by the prediction models over the target highway segment.

Model M-1: An Integrated Model for Travel Time Prediction under Missing Data

As seen in the literature, it is difficult to measure the reliability of the predicted travel times under the missing-data impacts. The integrated imputation approach (Model M-1) developed in this study views the travel time to be predicted during the data missing interval as a missing variable. Therefore, one can apply the core logic of existing multiple-imputation techniques but with some enhancements to estimate both the missing traffic data from detectors and the travel times from historical travel times. Eq. 5 describes these relations in the proposed M-1 Model.

$$Y_{mis}(t) = (Y_{mis}^{de.t}(t), \tau(t)) \quad (5)$$

where $Y_{mis}^{de.t}(t)$ is the missing detector data at time t; and

$\tau(t)$ is the travel time to be predicted at time t.

Determining Posterior Distribution

Same as those in the literature, one of the most critical tasks for implementing the model M-1 is to properly estimate the posterior distribution of the missing variables from their observed values. To do so, this study has carefully studied the dynamic nature of the detector data and their complex interactions with travel times. This is due to the fact that some variables do not directly contribute to the prediction of the travel time, excessive inclusion of such information may actually degrade the estimation results. For example, as shown in Figure 4, volume distributions varied significantly between two through lanes, due to drivers' knowledge of potential congestion on the right through lane caused by the daily queue spillback from the off-ramp to US29 southbound and/or the presence of queue on the off-ramp which blocks the usage of the right through lane at the detector location.

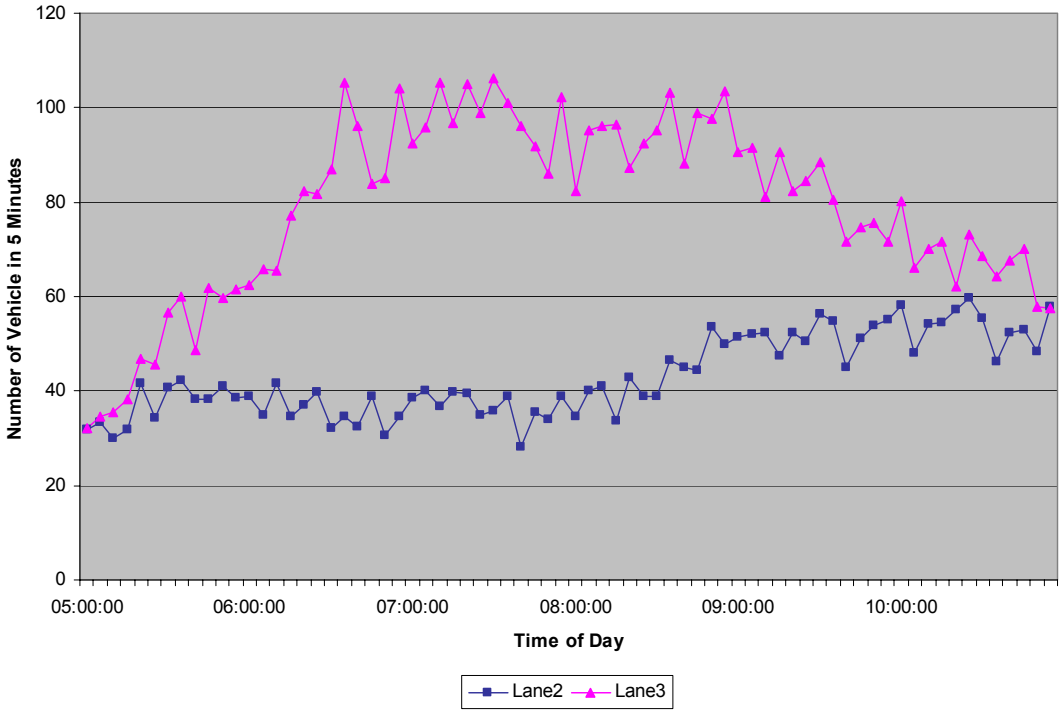


Figure 4 Average vehicle counts in 5-minute intervals on four Thursdays in July, 2006 at Exit 87A on I-70 (Detector 7)

To eliminate the impacts of unstable variables that do not contribute to the travel times for through traffic, the proposed imputation approach will first identify the critical lanes for through vehicles under various congestion patterns based on the following procedures:

- Categorize traffic scenarios based on the distributions of congestion patterns. In most cases, congestions can be categorized into morning peak, evening peak and off-peak hours. However, other traffic scenarios with different patterns may exist at the same locations and may result in selection of different critical lanes which are used by most through vehicles.
- Conduct field observations of congestion patterns so as to determine the critical lanes, which may include both the mainline and ramp lanes, for each scenario p.

- In the presentation hereafter, let $\mathbf{CLT}_d(p)$ be defined as the set of all critical through lanes at Detector *d*, which significantly contribute to computing of the average through traffic condition under scenario *p*, and $\mathbf{CLR}_d(p)$ as the critical ramp lanes at Detector *d* under the same scenario.

To better identify the posterior distribution of the missing data, this study has incorporated an enhanced *k*-Nearest Neighbor model (20) to search for similar historical cases. To take into account the traffic characteristics, one shall first categorizes traffic conditions with detected occupancy information, and then use the following equations to define the traffic conditions of free-flow, heavy congestion, and moderate congestion.

$$TC_d^{l_a}(t,t+\Delta t)=\begin{cases}-1 & \text{, when } o_d^{l_a}(t,t+\Delta t)\le OF_d^{l_a}\\ 1 & \text{, when } o_d^{l_a}(t,t+\Delta t)\ge OC_d^{l_a}\\ 0 & \text{, otherwise}\end{cases} \quad (6)$$

Where $TC_d^{l_a}(t,t+\Delta t)$ is the traffic type in lane l_a at detector *d* from time *t* to *t*+Δ*t*,
$o_d^{l_a}(t,t+\Delta t)$ is the average occupancy in lane l_a at detector *d* from time *t* to *t*+Δ*t*, and,
$OF_d^{l_a}$ and $OC_d^{l_a}$ are the upper bound of free-flow occupancy and lower bound of heavy congestion occupancy, respectively, for lane l_a at detector *d*.

The searching model then defines the distance, *dis,* between the current case and the candidate historical case as following:

$$dis=\sqrt{\sum_{i=1}^{k} w_i(p_i^*-q_i^*)^2} \quad (7)$$

Where $p_i^*=\begin{cases}p_i & \text{, when } TC_d^{l_a}(t,t+\Delta t)=0\\ OC_d^{l_a} & \text{, when } TC_d^{l_a}(t,t+\Delta t)=1\\ OF_d^{l_a} & \text{, when } TC_d^{l_a}(t,t+\Delta t)=-1\end{cases}$

$q_i^*=\begin{cases}q_i & \text{, when } TC_d^{l_a}(t_h,t_h+\Delta t)=0\\ OC_d^{l_a} & \text{, when } TC_d^{l_a}(t_h,t_h+\Delta t)=1\\ OF_d^{l_a} & \text{, when } TC_d^{l_a}(t_h,t_h+\Delta t)=-1\end{cases}$

p_i is the value of the i^{th} variable in the historical record;
q_i is the value of the i^{th} variable in the current state;
t and t_h are the time of day of the current case and the historical case respectively; and
w_i is the nonuniform weighting factor.

Operational Procedures

Using the geometric features and traffic patterns detected from the critical lanes on the target segment and the traffic characteristics accounted by the enhanced search model, One can exercise the integrated multiple imputation model (M-1) with the following steps.

Step 1: Construct a dataset with the information from critical lanes that do not encounter missing data at current time *t*. The dataset shall also include data prior to the current time *t* from each critical lane.

Step 2: Search for *h* complete historical cases that have traffic conditions most similar to the critical lanes at the current time. If historical cases are not adequate, the model will report that no reliable prediction can be made under the current missing patterns.

Step 3: Set imputation index i=1.

Step 4: Construct a set of complete variables (VAR_{COM}) in critical lanes in all *h* complete historical data records.

Step 5: Determine the probability distribution of missing variables, given the available complete data records $p(VAR_{MIS} \mid VAR_{COM})$.

Step 6: Impute all missing values and the travel time prediction based on $p(VAR_{MIS} \mid VAR_{COM})$.

Step 7: Integrate the newly obtained values from Step 6 with VAR_{COM} to form VAR'_{COM}.

Step 8: Test if the i^{th} imputation converge based on the differences in both the mean and the variance between $p(VAR_{MIS} \mid VAR_{COM})$ and $p(VAR_{MIS} \mid VAR'_{COM})$. If it converges, then go to Step 9. Otherwise, let $VAR_{COM} = VAR'_{COM}$, and then go to 6.

Step 9: Record the imputation results, then let i=i+1. If $i \le m$, go to Step 5.

Step 10: Determine the mean and variance of each variable in the *m* imputed data records. If all estimated variances are less than the assigned thresholds, then Model M-1 will output the average value of *m* imputed travel times as a reliable prediction under the current missing data pattern. Otherwise, the model will inform the system that no reliable result can be produced.

Note that one can execute the above procedures *m* times to generate a set of *m* imputed values. Prior to implementing the model, it is essential to determine four important parameters: the number of imputations *m*, the number of similar historical cases *h*, the criteria to determine the convergence of each imputation, and the location- and time-dependent thresholds of the variances of missing values. The convergence criteria and thresholds are available from the literature (1, 5). The former are usually defined as a penalty term that equals the covariance of two imputed values, given the non-missing data and the estimated covariance matrix.

Model M-2: Multiple Imputation of the Missing Detector Data

In addition to the integrated multiple imputation model, this study also develops a traditional multiple imputation model that imputes the missing values only for the scenarios in which similar historical cases are insufficient for Model M-1 to perform a reliable prediction. Under the same framework, this model groups related variables into one set of search indicators to generate imputations from similar cases. Model M-2 only takes into account traffic patterns in critical lanes in the same identified subsegment where the detector is experienced the missing data.

To capture the lane-varying traffic conditions, the proposed Model M-2 needs to divide the target freeway segment into several sub-segments based on the following criteria:

Step 1: Identify traffic scenarios based on the recurrent congestion patterns, and then perform the following steps for each traffic scenario.

Step 2: Group adjacent detectors into one subsegment if there are no ramps between the detectors.

Step 3: Combine adjacent subsegments if the detector at the interface point has a very low volume in the current traffic scenario and all ramps in the newly combined subsegment are covered by detector stations.

Step 4: Repeat Step 3 until no further combination is possible.

With the predefined subsegments for the current traffic scenario, one can further apply the following step-by-step procedures to estimate the missing values:

Step 1: Divide all observed missing values into groups based on their locations in the predefined subsegments for the current traffic scenario.

Step 2: Search for *h* similar historical cases with complete data in the target subsegment. If historical cases are not adequate, the model will report that no reliable imputation can be done for this group of missing data.

Step 3: Set the imputation index *i*=1.

Step 4: Construct VAR_{COM} with variables in the critical lanes within the subsegment from those *h* historical cases.

Step 5: Go through the same Steps 5 to 10 for Model M-1 to generate the final imputation results for the target subsegment.

Step 6: Repeat Steps 2 to Step 5 for all subsegments that experience missing data.

The system will then replace the missing detector data with the imputed results, and construct a complete input dataset for use by the travel time prediction model. By taking into account the geometric features and traffic congestion patterns, Model M-2 can supplement Model M-1 when a direct estimate of travel time is not available.

System Flowchart

Figure 5 illustrates the framework of the proposed system, which combines two missing data imputation models: Model M-1 and Model M-2. When the system detects that some data missing incurs in the input dataset for prediction during real-time operations, it will first apply Model M-1 to directly impute the travel time $TT_{M1}(t)$. If the variance of the imputed result from Model M-1 is larger than the time-dependent threshold $TH_{M1}(t)$, the system will then switch to Model M-2 to impute the missing values (traffic flow and/or occupancy by lane) in the input dataset. The system will execute the models for travel time prediction if the imputed detector values are reliable when compared with the time- and location-dependent flow threshold, $TH^{v}_{M2}(d,la,t)$, and/or the occupancy threshold, $TH^{o}_{M2}(d,la,t)$. Otherwise, the system will stop the prediction for the target segment.

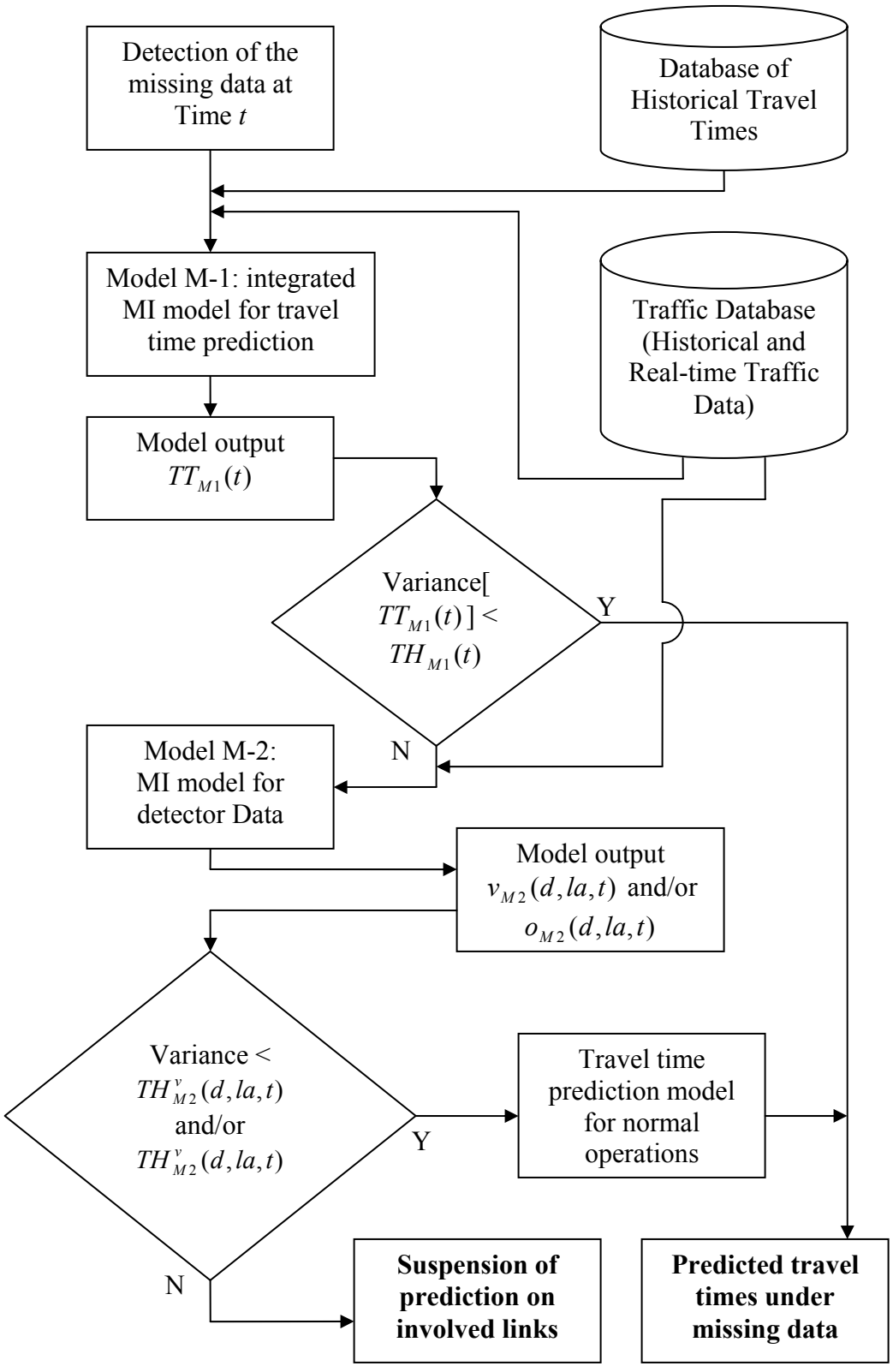


Figure 5 Flowchart of the system flowchart

NUMERICAL EXAMPLES

This section presents some numerical evaluation results for these two proposed missing data imputation models. The target highway used for the evaluation is the subsegment between Detector 2 and Detector 10 from ARAMPS, which is about 13.81 miles with a free-flow travel time of 694 seconds. The evaluation periods were between 15:00 and 19:00 on four of five consecutive weekdays from June 20th, 2006 (Tuesday) to June 26th, 2006 (Monday), excluding June 23rd, 2006 (Friday). The numerical examples intend to explore the following two issues:

- The performance of each imputation model under different missing rates (for detector 10); and
- The relation between the number of multiple imputations executed in each proposed model and its resulting performance.

Figure 6 shows the distribution of the estimated travel times between 15:00 to 19:00 on these four weekdays. The peak periods of these four experimental days had different starting times, but the same ending times. The estimated travel times that are based on the complete set of detector information prior to and after the peak period serve as the true values for the performance evaluation. The estimation model is from the study by Zou et al. (21).

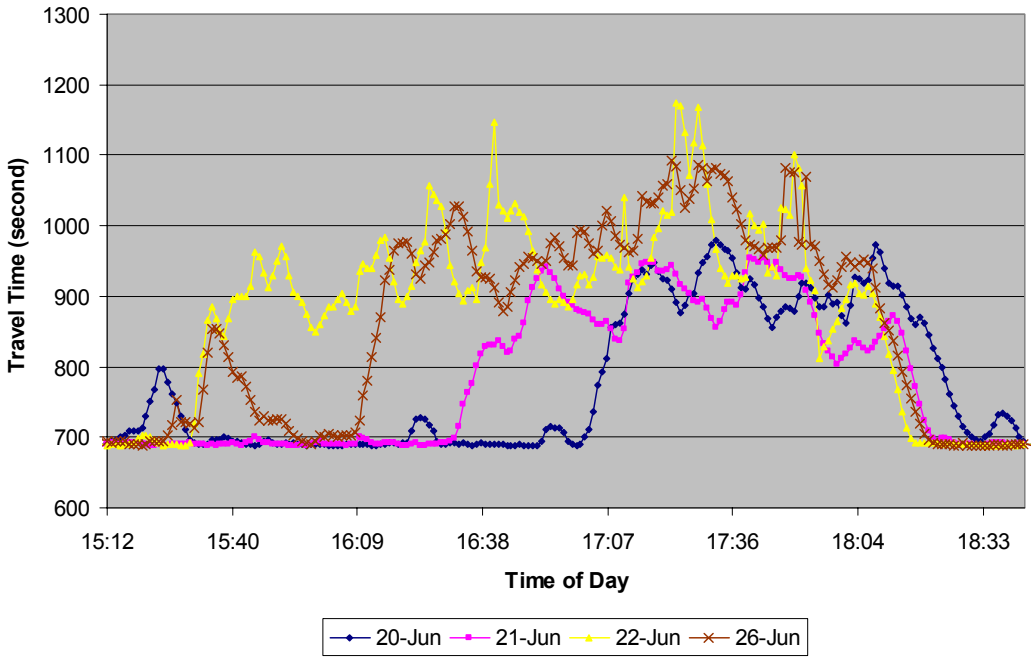


Figure 6 Distributions of Travel Times between 15:00PM and 19:00PM over selected experimental 4 days in 2006

The numerical analysis is focused on comparing the performances of the following five types of models for missing data estimation: mean substitution (MS), Bayesian forecast (BF), multiple imputation model (Model M-2), and the integrated multiple imputation approach (Model M-1). The travel time prediction model developed by Zou et al. (20) is used to generate the predicted travel times for MS, BF and M-2. A sensitivity analysis of the performance quality with respect to the required number of imputations (m = 5, 10, 20 and 50) for each

candidate imputation model has also been conducted. The experimental scenarios for evaluation include the data missing rates of 20%, 40%, 60%, and 100% incurred at Detector 10, which is a critical detector for both travel time estimation and prediction in all traffic scenarios.

Overall Performance over All Four Days

Figure 7 shows the distributions of average absolute relative errors (AARE) , as defined in Eq. 8, with each of those four methods over those four experimental days. In what follows, M-2-*m* and M-1-*m* denote Model M-2 and Model M-1 with *m* imputations, respectively. The results showed that Model M-1-50 has the best performance, compared to all other models when the data is missing at the rate of 20%, 40% and 60%, and its performance is very similar to Model M-2-50 when the data is missing at the rate of 100%. Model M-2-50 provided a similar performance to MS and BF at the missing rate of 20%, but exhibited better accuracy than all other three methods at the missing data rate of 100%.

$$\text{AAE} = \frac{1}{N}\sum_{n=1}^{N} | \tau_n - \hat{\tau}_n |$$

$$\text{AARE} = \frac{1}{N}\sum_{n=1}^{N} \frac{| \tau_n - \hat{\tau}_n |}{\tau_n} \qquad (8)$$

where AAE is the average absolute error,
AARE is the average absolute relative error,
N is the number of data samples available for comparison,
n is the index of the data sample,
τ_n is the n^{th} observed travel time, and
$\hat{\tau}_n$ is the travel time from the model.

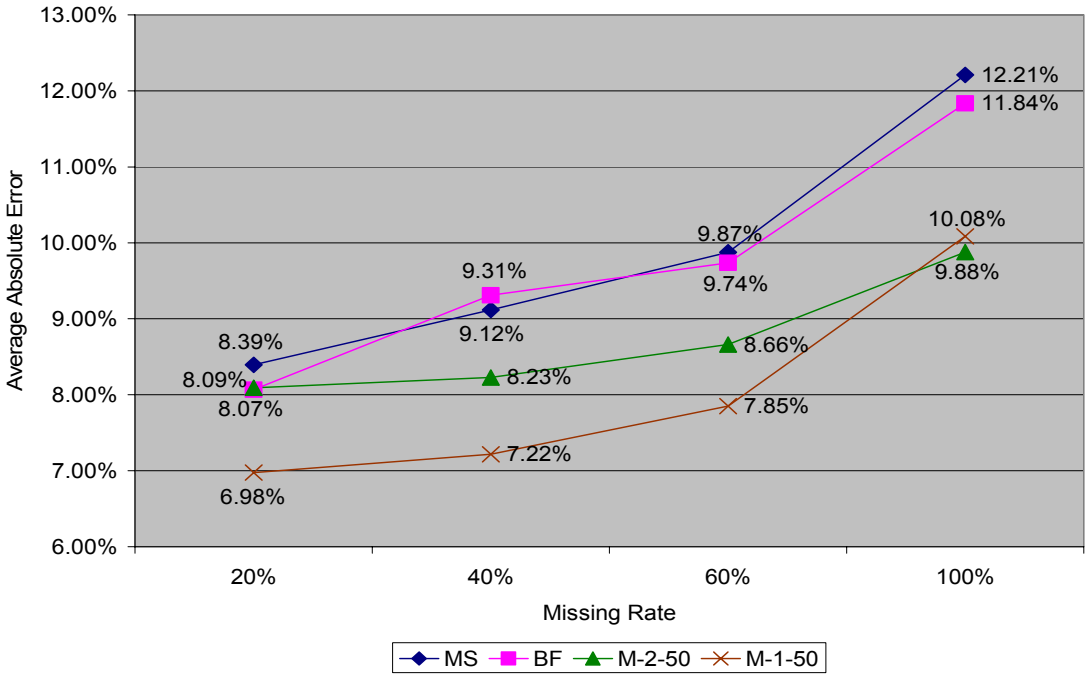


Figure 7 Average Absolute Relative Errors of All 4 Days under Different Missing Rates

Table 2 further compares the performance of all methods in different ranges of travel time, which include congestion-free conditions (travel time less than or equal to 700 seconds), moderate congestions (travel time between 700 and 900 seconds), and heavily congested conditions (travel time exceeds 900 seconds). It is noticeable that Model M-1-50 was the best among all models at the missing data rates of 20%, 40% and 60%, while Model M-2-50 outperformed the other three methods when Detector 10 could not function at all. Model-M1-50 and Model-M2-50 exhibited the same level of performance at the missing data rate of 100% in all three categories.

Table 2 Performance of Four Imputation Models in Different Traffic Conditions (Average Absolute Relative Error)

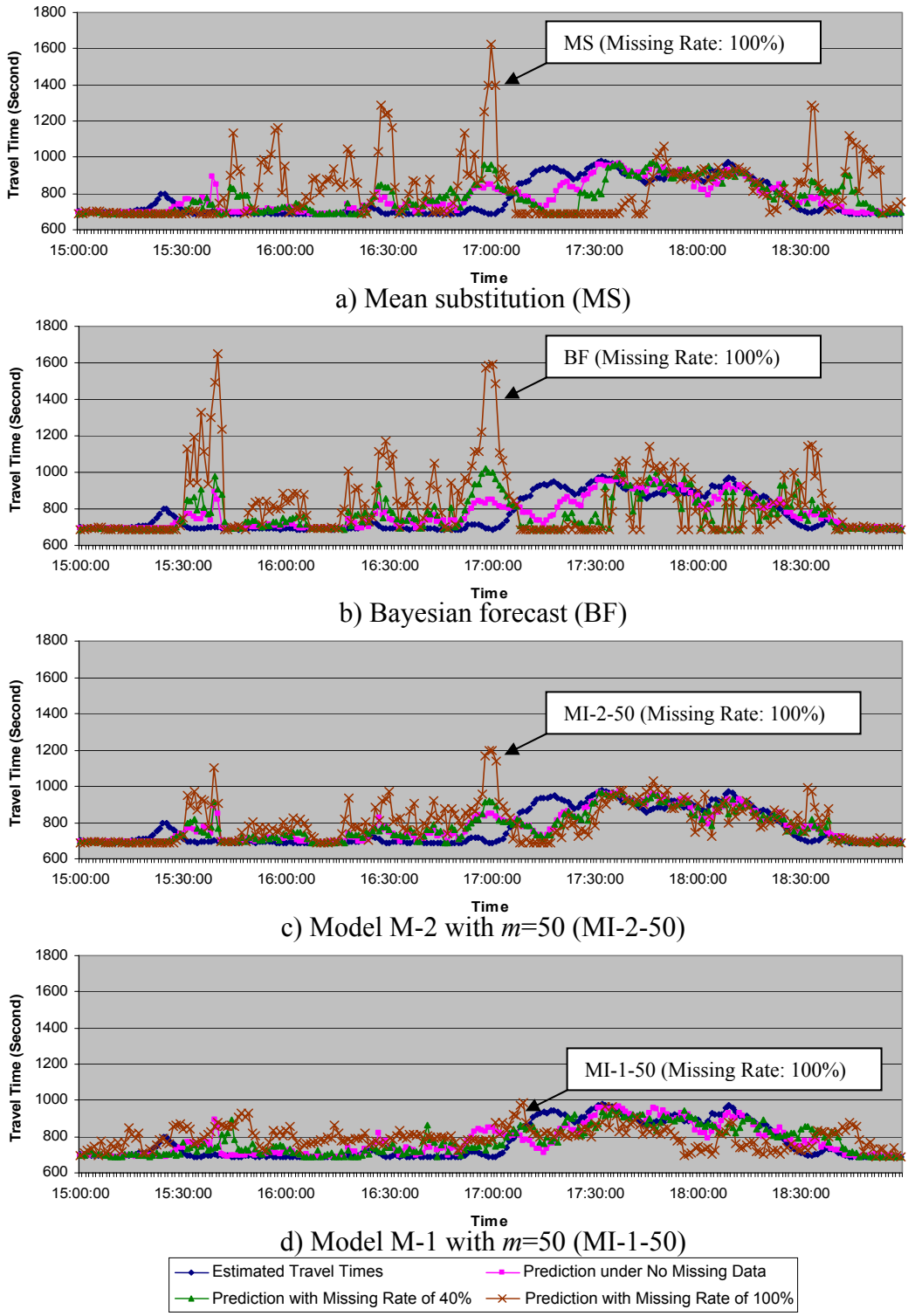
TT≤700		MS	BF	M-2-50	M-1-50
Missing Rate	20%	3.10%	2.78%	3.11%	2.54%
	40%	4.10%	3.63%	3.26%	2.80%
	60%	5.05%	4.47%	3.73%	3.07%
	100%	8.53%	7.47%	5.42%	6.37%
700<TT≤900		MS	BF	M-2-50	M-1-50
Missing Rate	20%	8.23%	7.35%	7.43%	6.65%
	40%	8.76%	8.56%	7.29%	6.43%
	60%	9.33%	8.64%	7.71%	6.95%
	100%	10.48%	10.26%	8.58%	8.66%
TT>900		MS	BF	M-2-50	M-1-50
Missing Rate	20%	13.46%	12.80%	12.36%	10.96%
	40%	13.82%	15.05%	12.57%	11.86%
	60%	14.29%	15.28%	12.99%	12.76%
	100%	16.12%	15.86%	13.55%	14.07%

TT: Travel time; MS: Mean substitute; BF: Bayesian forecast
M-2-50: Model M-2 with the number of imputation m=50
M-1-50: Model M-1 with the number of imputation m=50

Performance Comparison with Individual Day Data

This study has further evaluated the performance of each candidate models on a single day to evaluate the potential errors due to various congestion patterns. As shown in Figures 8(a) and 8(b), both MS and BF models, which are widely used in the existing traffic data warehouse systems, provided satisfactory results when the missing data rate was less than 40% in the evening peak hours, except during the transition periods between moderate and congested conditions. However, when the missing rate for Detector 10 was up to 100%, both models yielded unacceptable prediction results. Figures 8(c) and 8(d) show the prediction results from Model M-1-50 and Model M-2-50 under the same missing rates of 40% and 100% on June 20th, 2006. It is clear that travel time predictions with these two proposal multiple imputation models are more reliable and robust, especially during the transition periods. The integrated model M-1-50 is much more robust than MS, BF, and MI-2-50; its largest prediction error was less than 4 minutes (2%) when detector 10 is not functioning at all. Model M-1-50 and Model M-2-50

have similar average absolute relative errors of 12.15% and 12.06%, respectively, for the travel time of around 16 minutes over the entire evening peak on June 20th, 2006, compared to the prediction errors of 18.41% and 20.78%, respectively, for MS and BF.



a) Mean substitution (MS)

b) Bayesian forecast (BF)

c) Model M-2 with m=50 (MI-2-50)

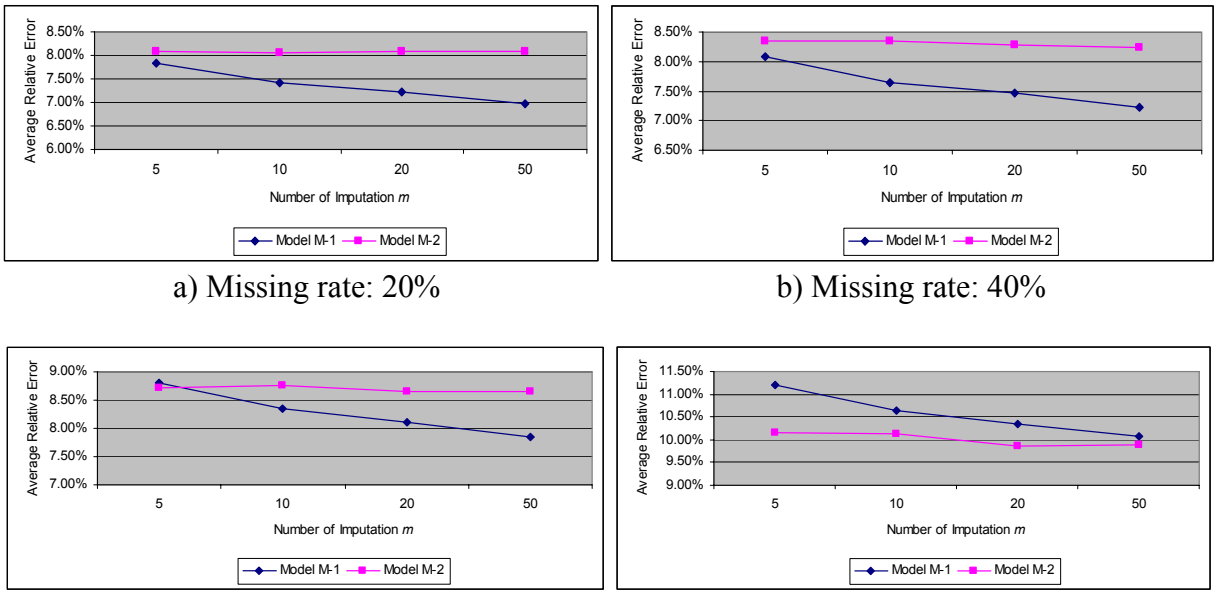
d) Model M-1 with m=50 (MI-1-50)

Figure 8 Performance comparisons of four imputation models at missing data rates of 40% and 100% at Detector 10 on June 20th, 2006

Sensitivity Analysis for Multiple Imputation Models

Since both Model M-1-50 and M-2-50 show better accuracy and reliability than other two commonly-used models on the data over these sample days, this section will further investigate their performance under different numbers of imputation.

Figures 9(a) to 9(d) illustrate the average absolute relative errors from Model M-1 and Model M-2 on all four sample days with different numbers of imputation and different missing data rates. As expected, the predicted accuracy of the integrated multiple imputation method, Model M-1, varies with the number of imputation being used. Its performance increased more than 10% when the number of imputation *was changed* from 5 to 50. However, an increase in the imputation number seems to have less significant impact on the performance of Model M-2 as shown in Figure 9. Its performance improvements are all less than 3% when *m* is increased from 5 to 50. The results of Model M-2 are consistent with those reported by Little and Rubin (1), which suggested that *m* should be between 3 and 10.



a) Missing rate: 20%

b) Missing rate: 40%

c) Missing rate: 60%

d) Missing rate: 100%

Figure 9 Average relative errors of models M-1 and M-2 on all four days with different *m* and missing data rates

Because Model M-1 was developed specifically for the proposed travel time prediction model, Little and Rubin's estimation of an efficient *m* does not fit this model. As shown in Figure 9 and Table 3, on average, the prediction error may increase about 5% when *m* decreases from 20 to 10, and increase about 3% when *m* is reduced from 50 to 20. For a prediction error of 4 minutes, an increase of 3% amounts to 7.2 seconds and an increase of 10% equals about 24 seconds. Hence, one can determine the number of *m* based on the required accuracy of the application. For example, an increased accuracy of 7 seconds may not be critical for the predicted travel times of more than 30 minutes for commuters.

Table 3 Performance improvements of Model M-1 under different imputation numbers

Increase of *m*	Missing Rate			
	20%	40%	60%	100%
From 5 to 10	5.38%	5.32%	5.32%	5.12%
From 10 to 20	2.60%	2.82%	2.82%	2.82%
From 20 to 50	3.31%	3.12%	3.12%	2.39%

* The numbers in the content of Table 5 represent the decrease rates of average absolute relative error for travel time prediction with the according increase of the number of imputation *m*.

Overall, the developed missing data imputation system, consisting of an integrated multiple imputation model for direct prediction of the travel time and a multiple imputation model for estimating the missing detector data, demonstrated its potential for use in practice, based on the experimental results with the field data (June 20th, 21st, 22nd and 26th, 2006). Both models outperform other widely-used imputation methods. With the number of imputations being set at 50, the integrated model can offer the acceptable accuracy and robustness for travel time prediction over those sample days.

CONCLUSIONS

This paper has developed two multiple imputation models, one integrated imputation model for the travel time prediction (Model M-1) and one multiple imputation model for estimating the missing detector data (Model M-2). Both models taking into account geometric features and traffic patterns over the target freeway segment can achieve better accuracy and robustness than those in the literature. Model M-1 can directly predict the missing travel time under some missing data scenarios. Model M-2 that classifies detector stations into groups is designed to take advantage of historical information in restoring missing set of detector data for the prediction model. Based on the data collected from 10 roadside detectors on a 25-mile stretch of I-70 eastbound in ARAMPS, the evaluation results indicate that both Models M-1 and M-2 outperformed two commonly-used imputation methods (mean substitution and Bayesian forecast) when missing data rates of 20%, 40%, 60% and 100% incurred at a critical detector. The average absolute relative errors of M-1-50 and M-2-50 under the missing rate of 100% (10.08% and 9.88% respectively) were noticeably lower than MS (11.84%) and BS(12.21%). A sensitivity test showed that the performance of Model M-1 may increase more than 10% when the number of imputation (*m*) increases from 5 (M-1-5) to 50 (M-1-50).

In brief, this study has presented two effective methods for contending with the data missing issues in real-time operations of travel time prediction. Since the impact of missing data on the prediction accuracy may vary with the methods used for prediction and the actual detector spacing, it should be recognized that extensive field calibration of the proposed methods will be needed for any real-world applications. Also note that the effectiveness of any imputation model is based on the assumption that all developed sensors have been rigorously calibrated and detected traffic data such as speed, volume are sufficiently for developing a reliable prediction model.

REFERENCE

1. Little, R., and Rubin, D., (1987) "Statistical Analysis with Missing Data", *Wiley*, New York.
2. Schafer, J., and Graham, W., (2002) "Missing Data: Our View of the State of the Art," *Psychological Methods*, Vol. 7(2), pp. 147-177.
3. Rubin, D., (1976) "Inference and Missing Data," *Biometrika*, 63, pp. 581-592.
4. Dempster, A., Laird, N., and Rubin, D., (1977) "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with Discussion)," *Journal of the Royal Statistical Society*, Series B, 39, pp. 1-38
5. Rubin, D., (1987) "Multiple Imputation for Nonresponse in Surveys," *Wiley*, New York.
6. Schafer, J., and Schenker, N., (2000) "Inference with Imputed Conditional Means," *Joruanl of the American Statistical Association*, Vol 95, pp. 144-154.
7. Madow, W. G.;Olkin, I., and Rubin, D. B., eds. (1983). Incomplete data in sample surveys, volume 2: Theory and bibliographies. New York: Academic Press
8. Saunders, J., Morrow-Howell, N., Spitznagel, E., Dork, P., Proctor, E., and Pescarino, R., (2006), "Imputing Missing Data: A Comparison of Methods for Social Work Researchers," *Social Work Research*, Vol. 30(1), pp. 19-31.
9. Souverein, O., Zwinderman, A., and Tanck, M., (2006), "Multiple Imputation of Missing Genotype Data for Unrelated Individuals," *Annals of Human Genetics*, Vol. 70, pp. 372-381.
10. Gan, X., Liew, A. and Yan, H., (2006) "Microarray missing data imputation based on a set theoretic framework and biological knowledge," *Nucleic Acids Research*, Vol. 34(5), pp. 1608-1619.
11. Nguyen, L., and Scherer, W., (2003) "Imputation Techniques to Account for Missing Data in Support of Intelligent Transportation Systems Applications," *Research Project Report for the Virginia Department of Transportation*, Research Report No. UVACTS-13-0-78.
12. Haj-Salem, H., and Lebacque, J., (2002) "Reconstruction of False and Missing Data with First-Order Traffic Flow Model," *Transportation Research Record* 1802, pp. 155-165
13. Chen, C., Kwon, J., Rice, J., Skabardonis, A., and Varaiya, P., (2003) "Detecting Errors and Imputing Missing Data for Single-Loop Surveillance Systems," *Transportation Research Record* 1855, pp. 160-167
14. Smith, B., Scherer, W., and Conklin, H., (2003) "Exploring Imputation Techniques for Missing Data in Transportation Management Systems," *Transportation Research Record* 1836, pp. 132-142
15. Zhong, M., Sharma, S., and Lingras, P., (2004) "Genetically Designed Models for Accurate Imputation of Missing Traffic Counts", *Transportation Research Record* 1879, pp. 71–79.
16. Al-Deek, H., and Chandra, C., (2004) "New Algorithms for Filtering and Imputation of Real-Time and Archived Dual-Loop Detector Data in I-4 Data Warehouse," *Transportation Research Record* 1867, pp. 116–126.
17. Al-Deek, H., Kerr, P., Ramachandran, B., Pooley, J., Chehab, A., Emam, E., Chandra, R.,

Zuo, Yo, Petty, K., and Swinson, I., (2004) "The Central Florida Data Warehouse (CFDW), Phase-2-, The Central ITS Office Funding," *Research Report to Florida Department of Transportation*.

18. Kwon, T., (2004) "TMC Traffic Data Automation for Mn/DOT's Traffic Monitoring Program," *Research Report to Minnesota Department of Transportation*, Report No. MN/RC-2004-29.
19. Ni, D., Leonard II, J., Guin, A., and Feng C., (2005) "Multiple Imputation Scheme for Overcoming the Missing Values and Variability Issues in ITS Data," *Journal of Transportation Engineering*, Vol. 131 (12), pp. 931-938.
20. Zou, N., Wang, J. and Chang, G. (2007) "A Reliable Hybrid Prediction Model for Real-time Travel Time Prediction with Widely Spaced Detectors" *Working Paper*
21. Zou, N., Wang, J. and Chang, G. (2008) "A Hybrid Model for Reliable Travel Time Estimation on a Freeway with Sparsely Distributed Detectors," *14th World Congress on ITS*, Beijing, China