Flexible and Robust Method for Missing Loop Detector Data Imputation

Kristian Henrickson
Graduate Research Assistant
101B More Hall
Department of Civil Engineering
University of Washington, Seattle, WA 98195-2700
Phone: (509) 217-8068
Email: henr2237@uw.edu

Yajie Zou, Ph.D.
Research Associate
133B More Hall, Box 352700
Department of Civil and Environmental Engineering
University of Washington, Seattle, WA 98195-2700
Phone: (936) 245-5628 Fax: (206) 543-5965
Email: zouyajie@gmail.com

Yin hai Wang1, Ph.D.
Professor
Department of Civil and Environmental Engineering
University of Washington, More Hall 133B
Phone: (206) 616-2696 Fax: 206/543-1543
Seattle, WA 98195
Email: yinhai@uw.edu

Paper submitted for presentation and publication at the 94th Annual Meeting of
The Transportation Research Board.

Word Count: 5735 (text) + 1500 (6 tables) + 250 (1 figure) = 7485
ABSTRACT

This work is primarily focused on missing traffic sensor data imputation for the purpose of improving the coverage and accuracy of traffic analysis and performance estimation. Missing data, whether attributable to hardware failure or error detection and removal, is a constant problem in loop and other traffic detector datasets. As the rate of missingness increases, the treatment of missing values quickly becomes the controlling factor in overall data quality. Previously, a number of imputation approaches have been developed for traffic data. However, few studies aim at handling the traffic data with large blocks of missing values for network-wide implementation. A proven predictive mean matching multiple imputation method is introduced and applied to loop detector volume data collected on Interstate 5 in Washington State. Using the iterative multiple imputation by chained equations approach, the spatial correlation between nearby detectors is considered for prediction and the presence of missing data in all predictors is effectively dealt with. The proposed methodology is shown to perform well on a range of missing data patterns including missing completely at random, missing days, and missing months. After applying the imputation method to 20-second data and performing post-imputation aggregation, the results in this study suggest that the proposed method can outperform elementary pairwise regression and produce reliable imputation estimates, even when entire days and months are missing from the dataset. Thus, the predictive mean matching multiple imputation method can be used as an effective approach for imputing missing traffic data in a range of challenging scenarios.
INTRODUCTION

This work is primarily focused on missing data imputation in transportation applications for the purpose of improving the coverage and accuracy of performance estimation. It is readily apparent that, as the rate of “missingness” increases due to detector malfunction or removal during quality control processing, the way that missing data is dealt with quickly becomes the controlling factor in overall data quality. While a great deal of work has been completed on identifying and correcting the various sources of error and imputation of missing values, there remains a substantial gap in terms of a) the relationship between missing data mechanisms/patterns and imputation accuracy and b) statistically principled methodologies that deal with missing transportation data in a way that is efficient both in terms of computational complexity and analyst time investment. This work addresses these needs by first applying an intuitive and computationally efficient multiple imputation approach to loop detector data. Second, the accuracy of the proposed imputation approach is investigated under multiple missing data scenarios including missing completely at random, missing day, and missing month. No previous work has identified which attempts to quantify the accuracy of any industry standard or novel missing data imputation methods on substantial continuous blocks of missing data, although such patterns are frequently observed in practice.

Background

Chen et al. (1) demonstrated an iterative single imputation process using pair-wise linear regression. Each missing value is then imputed using the median of the linear regression estimates for all neighboring loop detectors, restricted to only those reporting “good” values. Expanding on this concept, Al-Deek and Chandra (2) reported good results for dual loop detector data using a pair-wise second order models with speed, volume, and occupancy interaction terms. They also developed a selective median predictor selection approach, which improved results on the test scenario (which included only a single detector cabinet).

A number of imputation approaches have been developed which utilize a low dimensional representation of road network detector data in order to take advantage of key spatial and temporal correlation structures. Qu et al. (3) developed a method based on Probabilistic Principle Component Analysis (PPCA), a data mining approach which seeks to represent high dimensional data as a set of principle components or linear combinations of predictors. Li et al. (4) developed a Kernel Probabilistic Principle Component model, expanding on the basic PPCA approach to incorporate both spatial and temporal predictors. Asif et al. (5) developed two imputation methods for large, interconnected road networks using dimension reduction data mining techniques. Specifically, Fixed Point Continuation with Approximate Singular Value Decomposition and Canonical Polyadic decomposition were applied to project the spatial and temporal relationships between neighboring locations into a lower dimensional space. While these data mining approaches to missing data imputation were shown to be both accurate and computationally efficient, they were all developed and tested on 5-minute aggregation intervals where better spatial and temporal correlation is present, and are not likely to perform as well on shorter time intervals. Smith et al. (7) developed an imputation algorithm combining Expectation Maximization (EM) and a Data Augmentation (DA) for missing ITS data imputation. The algorithm was implemented on pre-aggregated 10-minute data which, as described previously, often results in better correlation between neighboring detectors and time periods as well as dramatically decreased noise. Ni et al. (8) developed a EM/DA multiple imputation methodology for missing ITS data, and applied it to video-based volume data collected on GA 400 near the Atlanta metropolitan area. In a multiple imputation framework, this methodology is similar to that in Smith et al 2003 except that the process is repeated $m$ times to create $m$ multiple imputed datasets, and the final results averaged to give a final imputed value. In a similar vein, Ni
& Leonard (9) developed a DA multiple imputation scheme using a Bayesian network to describe the model structure. Imputations were generated in this case under an ARIMA time series modeling framework, with DA used to iteratively draw Bayesian network and ARIMA parameters for predictions. As noted by the authors, the ARIMA modeling approach is not useful when many sequential observations are missing from the dataset. This work (8, 9) stresses the importance of imputation at the 20-second aggregation level, and applied their algorithm to video-based ITS data from the Atlanta area.

Several works have identified non-parametric modeling as a flexible approach to missing traffic data imputation. For example, Haworth and Cheng (10) developed a non-parametric scheme for online missing data imputation based on K-nearest neighbor (KNN). Multiple variations of the KNN algorithm as well as a kernel regression approach were developed and tested on 5-minute link travel time data in the London metropolitan area. Chang et al. (11), on the other hand, took a time-series approach to KNN imputation, defining each state vector as a set of observations (at 1 hour intervals) from the detector of interest over a block of time. The performance of this method was compared to a seasonal ARIMA model, and shown to perform comparably with reduced effort and computation time. As with other strictly time series approaches, this method does not take nearby detector records into account, and so is not useful when longer time periods of data are missing.

Tan et al. (6) developed an imputation method using a 4-way tensor model to represent the spatio-temporal correlation structures in 5-minute loop data. The algorithm was demonstrated on 5-minute loop data and required substantial tuning of the model complexity and parameter set. This project was unique in that imputation accuracy was presented for cases where an entire day(s) of data were missing. With the exception of (6), none of the above listed imputation approaches attempted to quantify the performance of the method in cases where longer time periods of data are missing (e.g. day, month). In addition, most recent methods were developed and tested on 5-minute or longer time interval data, for which better spatial and temporal correlation is present. As has been previously demonstrated (8, 9), better imputation can be achieved by applying a statistically principled imputation method at the 20-second level and aggregating after imputation.

Missing Data Patterns

There are a number of reasons why missing data is present in traffic sensor datasets, including various types of hardware malfunction, communications failure, and events related to traffic conditions. Though in some cases the occurrence of “missingness” is predictable, others are completely random or are related to unobserved predictors. Several studies in statistical literature have focused on identifying missing data patterns for which the causal mechanisms can be ignored in the imputation process. Most current work considers the occurrence of missing data under a probabilistic framework, with the pattern described by a statistical distribution (12). The mechanism driving the missing data pattern is assumed to be ignorable if data is Missing At Random (MAR), which is only true when the distribution of missingness is not dependent on the unobserved or missing values. That is, if the $n \times p$ dataset defined as $X$ is constituted of both observed and unobserved components ($X_{obs}$ and $X_{mis}$ respectively), and $R$ designates a corresponding $n \times p$ matrix with values 0 or 1 depending on whether a value is missing or observed, the probability distribution for $R$ depends only on $X_{obs}$ as shown in Eq (1) (13, 12)

$$Pr(R|X) = Pr(R|X_{obs})$$ (1)
Data is described as Missing Completely At Random (MCAR) when the occurrence of missingness is independent of both observed and unobserved values, as shown in Eq (2). This is considered a special case of MAR.

\[ \Pr(R|X) = \Pr(R) \]  

(2)

If both Eq (1) and Eq (2) are violated, the data is considered to be Missing Not At Random (MNAR), which means that the missing data mechanism is not ignorable. When the mechanism or distribution of missingness is not known (as is often the case), data is typically assumed to be MAR to simplify the imputation process. While this is often violated, the assumption is made more plausible by including additional predictors that may help to describe the distribution of missingness. Note that these definitions seem to allow substantial blocks of time to be missing from a time series dataset without violating the MAR assumption, because there is no causal relationship between the missing data pattern and the unobserved values. However, several publications (e.g. 12) assert that Eq (1) and Eq (2) do not imply a causal relationship. Thus, if blocks of time are missing from a time series, data is only MAR if the missing data follow a distribution identical to that of the observed data. Because the missing data is not observed, this cannot be assumed to be true, and so some bias will likely be introduced. In any case, as noted by Schafer (14), principled MAR-based missing data treatments are superior to ad hoc solutions, as the bias that can be explained by the observed values is removed, which is not true in general for ad hoc procedures.

**Missing Data Patterns in Loop Detector Data**

In developing an imputation scheme for loop detector data, it is important to consider the various causes of missingness in loop detector data with specific attention to violations of the MAR assumption inherent to the imputation framework. Some missingness mechanisms in loop detector data can result in truly MAR or even MCAR patterns. For example, some error types (e.g. segmentation error) can be detected using an occupancy or volume/occupancy threshold. Assuming no consistent underlying hardware issues, the measured value in such cases is an artifact from a random event (i.e. a vehicle crossing the detector at a particular time) bearing no relation to the true quantity, and can be safely removed without violating the MAR assumption. While it is true that the ability to detect such errors is somewhat dependent on traffic conditions, this mechanism can be described at least in part by neighboring detector observations. Other error types, such as sensitivity maladjustment, stuck on/off detectors, or other consistent hardware malfunctions, often result in a large number of sequential observations being removed, clearly violating the MAR assumption. Thus, it is of critical importance to estimate the sensitivity of any imputation method to such MNAR patterns, and the algorithm used to detect and eliminate erroneous values must have some mechanism for identifying the error type in order to make the distinction between random and not random missing patterns.

**IMPUTATION FRAMEWORK**

Ruben (15) introduced multiple imputation (MI) as a principled way to deal with non-response in survey and census data. MI is a Monte Carlo technique, in which each missing value is replaced by \( m > 1 \) replacement values. This results in \( m \) completed datasets, which are then analyzed using complete data methods and the results combined to give confidence limits incorporating the uncertainty in the imputed values.
Multiple Imputation by Chained Equations

Multiple Imputation by Chained Equations (MICE) is a MI approach introduced in (16), in which a predictive model is defined separately for each variable with missing data. First, the missing values are filled with some initial estimate, typically by randomly sampling from the observed values. Then, for each variable with missing values, a regression model is estimated using the observed portion of the other variables as predictors. The missing values are then replaced with random draws from the resulting posterior predictive distribution. This process is repeated for each variable with missing data, using the observed and most recently estimated imputation estimates as predictors. After completing this process for all variables, the cycle is repeated several times updating the imputation estimates as described. This then constitutes a single imputed data set, and the entire process is repeated \( m \) times to give \( m \) multiple imputed datasets.

Under a parametric modeling framework, the MICE procedure for each iteration described in (16, 17) is as follows: Given the following dataset with \( p \) incomplete variables \( X = X_1, X_2, \ldots X_p \), define the observed portion of \( X \) as \( X^{obs}_1, X^{obs}_2, \ldots X^{obs}_p \) and the missing portion of \( X \) as \( X^{miss}_1, X^{miss}_2, \ldots X^{miss}_p \). Assuming that the multivariate distribution of the complete dataset is specified by the unknown parameters set \( \theta \), the posterior distribution described by \( \theta \) is obtained implicitly by sampling in an iterative fashion form the conditional distributions defined for each variable as shown below in Eq. 3:

\[
P(X_1|X_{-1}, \theta_1) \]
\[
| \]
\[
P(X_p|X_{-p}, \theta_p) \]

Where
\[
\theta_1, \theta_2, \ldots \theta_p = \text{parameters describing the conditional densities of } X_1, X_2, \ldots X_p
\]
\[
X_{-1} = \text{dataset } X \text{ with } X_1 \text{ variable removed}
\]

Imputations are drawn iteratively for each incomplete variable in \( (i = 1,2, \ldots p) \) steps by first simulating a random draw for the parameters, and then simulating random draws for the missing values in the variable of interest. This can be interpreted as a Gibbs sampling approach, which is a simple Markov Chain Monte Carlo algorithm used when sampling directly from the full multivariate distribution is difficult. The procedure can be shown as follows (Eq. 4) (from (18)):

\[
\theta_1^{* (t)} \sim P(\theta_1|X^{obs}_1, X^{t-1}_2, \ldots X^{t-1}_p)
\]
\[
X_1^{* (t)} \sim P(X_1|X^{obs}_1, X^{t-1}_2, \ldots X^{t-1}_p, \theta_1^{* (t)})
\]
\[
| \]
\[
\theta_p^{* (t)} \sim P(\theta_p|X^{obs}_p, X^{t-1}_1, \ldots X^{t-1}_{p-1})
\]
\[
X_p^{* (t)} \sim P(X_p|X^{obs}_p, X^{t-1}_1, \ldots X^{t-1}_{p-1}, \theta_p^{* (t)})
\]

Where
\[
t \in (1,2, \ldots n) = \text{iteration index}
\[ \theta_1^{*\text{(t)}}, \theta_2^{*\text{(t)}}, \ldots, \theta_p^{*\text{(t)}} = \text{random draws for parameters } \theta_1, \theta_2, \ldots, \theta_p \text{ during iteration } t \]

\[ X_1^{*\text{(t)}}, X_2^{*\text{(t)}}, \ldots, X_p^{*\text{(t)}} = \text{imputations drawn for } X_1^{\text{miss}}, X_2^{\text{miss}}, \ldots, X_p^{\text{miss}} \text{ during iteration } t \]

This process is completed for \( n \) iterations, and then the entire process is repeated \( m \) times to give \( m \) complete datasets as previously described.

**Predictive Mean Matching**

As described previously, the multiple imputation process involves generating random draws from the posterior predictive distribution of the variable of interest (\( y \)). In predictive mean matching, this is completed as follows:

1. Fit regression model using only observations corresponding to observed values of \( y \)
2. Generate random draws from the joint posterior distribution of the resulting parameters (i.e. \( \sigma \) and \( \beta \))
3. Identify the \( k > 1 \) observed values with the closest predicted mean value to each missing observation (using the estimated parameters from of step 2)
4. Fill in the missing values with a random draw from the \( k \) values identified in step 3
5. Repeat for each variable with missing values to be imputed, using the observed and most recently imputed values of predictors
6. Cycle through steps 2 – 4 multiple times to achieve some measure of convergence

These steps are discussed in order in the following paragraphs, according to \((15)\) and \((18)\). To start, we assume a model for normally distributed variables of the form shown below in Eq. (5).

\[ p(y|X, \beta) \sim N(\beta X, \sigma^2) \quad (5) \]

Where
- \( y \) = a variable containing missing which we would like to impute (i.e. traffic volume)
- \( X \) = Predictor matrix
- \( \beta \) = parameter vector

For the purpose of this illustration, let us assume \( y \) is constituted of two components, \( y^{\text{miss}} \) and \( y^{\text{obs}} \), corresponding to the missing and observed data respectively. Define \( \hat{\beta} \) as the estimated parameter vector obtained by fitting the linear model using only observations corresponding to \( y^{\text{obs}} \), \( V \) as the estimated covariance matrix for \( \hat{\beta} \), and \( \hat{\sigma} \) as the estimated root mean squared error. We can then draw imputation parameters \( \sigma^{*} \) and \( \beta^{*} \) in sequence, first drawing \( \sigma^{*} \) as shown below in Eq (6).

\[ \sigma^{*} = \hat{\sigma} \sqrt{(n^{\text{obs}} - k)/g} \quad (6) \]

Where
- \( \sigma^{*} \) = imputation parameter (sample root mean squared error)
- \( \hat{\sigma} \) = estimated root mean squared error from fitted model
- \( g \) = random draw from \( \chi^2 \) with \( (n^{\text{obs}} - k) \) degrees of freedom
- \( n^{\text{obs}} \) = number of observations in \( y^{\text{obs}} \)
- \( k \) = number of predictors + 1
With the imputation parameter $\sigma^*$, we can then draw the imputation parameter $\beta^*$ as shown in Eq (7) from (15).

$$\beta^* = \hat{\beta} + \frac{\sigma^*}{\hat{\sigma}} u_1 V^{1/2}$$

Where

$u_1 =$ row vector containing $k$ independent random draws from standard normal distribution

$V^{1/2} =$ Cholesky decomposition of $V$

Using $\beta^*$ and $\sigma^*$, predictions are made for all values in the dataset (both $y_{\text{miss}}$ and $y_{\text{obs}}$). After identifying the $k > 1$ observed values with the closest predicted mean to each missing value, the missing value is filled with a random draw from these observed values. This process, from (18), differs somewhat from the method described by Ruben (15), most notably in that matching between missing and observed values is based on the predicted means (not the observed value).

By iteratively sampling parameters and imputations, assuming the models have been correctly specified, the true variability of the estimate is represented in the final $m$ datasets. The PMM approach insures that the imputed values are within the range of observed values, and may perform better than linear regression approaches under certain violations of normality (19, 20). While the PMM multiple imputation approach has been widely used in a number of fields, this work represents the first application of the methodology in traffic sensor data.

**APPROACH**

**Overview**

In this research, a predictive mean matching multiple imputation algorithms is implemented using R statistical computing software. Specifically, the Multiple Imputations by Chained Equations (mice) package is used in conjunction with several other R packages to perform the following steps:

- Query 20-second single loop detector data from a Microsoft SQL Server database using the RODBC package in R (21)
- Preprocess data for formatting and consistency, as well as to remove erroneous observations. Specifically, volume/occupancy thresholding and manual inspection were used to identify erroneous observations
- Create several different error patterns for testing including missing completely at random (i.e. uniform random missing pattern at multiple rates of missingness), missing a predetermined selection of days, and missing a predetermined selection of months
- Predictors are selected is from among detectors at the cabinet of interest as well as from the closest upstream and downstream cabinets in both travel directions. Selection is based on correlation, including only predictors with Pearson correlation coefficient $> 0.1$
- Conduct multiple imputation using mice package in R (18, 22), generating $m$ multiple imputed datasets. To reduce processing time, the doParallel (23) and foreach (24) R packages were used to parallelize the process.
- Conduct aggregation and calculate measures of imputation accuracy
Report and compare results

Data is queried in month blocks for the random and missing days patterns. For missing month, data is queried for an entire year in hour blocks, such that each model is built using a single hour of data from each of the 366 days in 2012. It is thought that, by including data from every available month in model development, the bias that could be introduced by seasonal differences between months can be minimized. For each test, results are reported for a single detector, selected to have a very low missing rate in the ground truth data. Varying missing rates of 5% - 30% are present in all predictors. Figure 1 (below) shows a flow chart of the data processing and imputation approach, with a conceptual visualization of the missing data patterns.

FIGURE 1. Data Processing Flow Chart

MICE Procedure
MICE imputation was performed using the mice package in R (18). The number of imputations and iterations are both set by default to 5. A range of values were tested, the default values appear to provide a good balance of accuracy, consistency, and computation time. The visitation sequence is set to monotone, which results in imputation being performed in order of increasing missing data rate as suggested in
previous work (16). The predictor matrix is developed as described in the previous subsection. The PMM method was used in imputation as described previously.

The results from the mice function include m complete datasets, each corresponding to a single imputation. These can then be analyzed individually and the results averaged to give a final result. For example, if the result of the analysis is to be aggregated 5-minute volume data, each complete dataset is aggregated into 5-minute intervals and the results averaged over all imputed datasets.

**Pairwise Linear Regression**

For reference, we compared the performance the proposed PMM against a conventional imputation method. The conventional approach is a commonly used pairwise linear model proposed by Chen et al. (1), which considers the correlation between neighbor loops. This approach can relate the measurements from adjacent loops and is described in the following equation (Eq. (8)):

\[ q_i(t) = \alpha_0(i, j) + \alpha_1(i, j)q_j(t) \]  

(8)

where \( q_i(t) \) is the traffic volume at loop detector i and time t; \( q_j(t) \) is the traffic volume at loop detector j and time t; \( \alpha_0(i, j) \) and \( \alpha_1(i, j) \) are coefficients.

For each pair of neighbors (i, j), the least squares estimation method (Eq. (9)) was used in (1) to determine the parameters \( \alpha_0(i, j) \) and \( \alpha_1(i, j) \) using 5 days of historical data. In this study, it was found that using a full week of historical data to estimate parameters improved imputation estimates slightly, and so this was done for all pairwise imputation with the exception of missing months. When an entire month of data is missing, regression estimation was performed using the month before or after the month of interest, selected to insure the lowest rate of missingness in the training data.

\[ \alpha_0(i, j)\alpha_1(i, j) = \arg \min_{\alpha_0, \alpha_1} \left\{ \frac{1}{n} \sum_{t=1}^{n} [q_i(t) - \alpha_0 - \alpha_1q_j(t)] \right\} \]  

(9)

Readers are referred to (1) for more details about the pairwise linear model. The final imputed value for loop i is taken as the median of the pairwise estimates from its neighbors. Note that, to enhance the estimation performance of the pairwise linear model, the selective median algorithm proposed in (2) was implemented in this study. This insures that only the detectors closest to the detector of interest are used for prediction if they report usable data during the interval and, if not, then all nearby detectors can be used.

**Aggregation Levels**

The majority of current research on the topic of imputing missing traffic sensor data relies on data aggregated to 1-minute, 5-minute, or longer time intervals. This reduces the impact of random noise, and results in better spatial and temporal correlation structures. In Washington State, much of the performance reporting is based on 5-minute intervals, which further strengthens the argument for imputing pre-aggregated data. However, this research follows the principle of imputation before aggregation for several reasons. First, for data that is measured at the 20 or 30 second level, some elementary imputation is implicitly applied in the aggregation step. Thus, many “complete” 5-minute intervals are based on incomplete data, and many 5-minute intervals are marked as missing even when not all of the contributing
20-second values are actually missing. Second, by applying a principled imputation method at the lowest available aggregation level and aggregating the complete dataset to the desired time intervals, better performance can result even if the per-observation “accuracy” of imputation is lower than could be achieved by imputing pre-aggregated data (8, 9).

Data Description
The Washington State Department of Transportation (WSDOT) manages the loop detectors on state highways and interstate freeways within Washington State. The University of Washington STAR Lab downloads and archives a great deal of data for research work using an online FTP site provided by WSDOT. Two sites from the Interstate 5 corridor between mileposts 150 and 170 from 2012 between the hours of 6:00AM – 10:00PM are used in testing. This corridor includes the Seattle metropolitan area, Seattle-Tacoma International Airport, and the Port of Seattle. It is a combination of urban, semi-industrial, and suburban land use, with frequent interchanges rolling topology.

For the testing dataset, detectors were selected which consistently produce reliable volume measurements. Various synthetic missing data patterns were applied to the data including missing completely at random, missing day, and missing month. The accuracy of the imputed values can then be determined in relation to the true measured data.

RESULTS
In this section, the performance of PMM is evaluated using the traffic volume data collected on two sites along the Interstate 5. Tables 1 and 2 show the results for PMM and pairwise regression imputation on varying missing data rates at 20-second and 5-minute aggregation levels. For both methods, it is clear that the 20-second errors are not sensitive to the rate of missing data, which is as expected given that the data is missing completely at random as parameter are unbiased. The 20-second MAE and MAPE are for the imputed values only, and the 5-minute results include the available (i.e. non-missing) observed values.

### TABLE 1. PMM Results for Random Missing Patterns

<table>
<thead>
<tr>
<th>missing</th>
<th>May 2012 5 minute</th>
<th>May 2012 20 seconds</th>
<th>January 2012 5 minute</th>
<th>January 2012 20 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MAPE</td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>Site A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>1.52</td>
<td>1.5%</td>
<td>1.31</td>
<td>22.7%</td>
</tr>
<tr>
<td>20%</td>
<td>2.13</td>
<td>2.1%</td>
<td>1.28</td>
<td>22.1%</td>
</tr>
<tr>
<td>30%</td>
<td>2.61</td>
<td>2.6%</td>
<td>1.30</td>
<td>22.8%</td>
</tr>
<tr>
<td>40%</td>
<td>2.92</td>
<td>2.9%</td>
<td>1.29</td>
<td>22.7%</td>
</tr>
<tr>
<td>50%</td>
<td>3.12</td>
<td>3.1%</td>
<td>1.30</td>
<td>22.8%</td>
</tr>
<tr>
<td>60%</td>
<td>3.37</td>
<td>3.4%</td>
<td>1.30</td>
<td>22.8%</td>
</tr>
</tbody>
</table>

| Site B  |       |        |       |        |       |        |       |        |
| 10%     | 2.15  | 1.9%   | 1.63  | 27.7%  | 2.14  | 3.8%   | 1.50  | 32.3%  |
| 20%     | 3.12  | 2.8%   | 1.64  | 27.8%  | 2.89  | 4.6%   | 1.49  | 32.2%  |
| 30%     | 3.63  | 3.3%   | 1.63  | 27.7%  | 3.49  | 5.4%   | 1.48  | 32.2%  |
| 40%     | 4.03  | 3.6%   | 1.63  | 27.5%  | 3.81  | 5.8%   | 1.48  | 31.9%  |
| 50%     | 4.43  | 4.0%   | 1.62  | 27.4%  | 4.19  | 6.4%   | 1.48  | 32.2%  |
| 60%     | 4.81  | 4.3%   | 1.63  | 27.8%  | 4.45  | 6.6%   | 1.48  | 32.0%  |
TABLE 2. Pairwise Regression Results for Random Missing Patterns

<table>
<thead>
<tr>
<th>Missing</th>
<th>May 2012 5 minute</th>
<th>May 2012 20 seconds</th>
<th>January 2012 5 minute</th>
<th>January 2012 20 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MAPE</td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>Site A</td>
<td>10%</td>
<td>1.94</td>
<td>1.9%</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>3.33</td>
<td>3.4%</td>
<td>1.65</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>4.57</td>
<td>4.7%</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>5.78</td>
<td>5.9%</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>7.11</td>
<td>7.3%</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>60%</td>
<td>8.13</td>
<td>8.4%</td>
<td>1.66</td>
</tr>
<tr>
<td>Site B</td>
<td>10%</td>
<td>3.30</td>
<td>3.1%</td>
<td>1.81</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>2.03</td>
<td>1.9%</td>
<td>1.82</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>4.34</td>
<td>4.2%</td>
<td>1.79</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>5.35</td>
<td>5.2%</td>
<td>1.79</td>
</tr>
<tr>
<td></td>
<td>50%</td>
<td>6.25</td>
<td>6.1%</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>60%</td>
<td>7.16</td>
<td>7.0%</td>
<td>1.80</td>
</tr>
</tbody>
</table>

Tables 3 and 4 show the results for PMM and pair-wise regression for missing days at the 20-second and 5-minute aggregation levels. The missing dates in May for Saturday, Wednesday, and Monday fall on the 5th, 16th, and 28th respectively. For January, the missing days fall on the 7th, 11th, and 23rd. Note that the difference in accuracy between PMM and pair-wise regression are more pronounced when entire days (Tables 3 and 4) are missing compared to missing at random (Tables 1 and 2), and that the PMM method consistently outperforms pairwise regression. It is also worth noting that the average error at the 5-minute aggregation level was between {1.29 to -2.05} for PMM and {3.78 to -7.29} for pairwise regression, indicating that PMM produces less bias. In this case, the entire days were removed from the dataset, and so both 20-second and 5-minute values contain only the imputation estimates.

TABLE 3. PMM Results for missing days

<table>
<thead>
<tr>
<th>Missing</th>
<th>May 2012 5 minute</th>
<th>May 2012 20 seconds</th>
<th>January 2012 5 minute</th>
<th>January 2012 20 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MAPE</td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>Site A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saturday</td>
<td>4.04</td>
<td>3.7%</td>
<td>1.27</td>
<td>21.3%</td>
</tr>
<tr>
<td>Wednesday</td>
<td>3.81</td>
<td>4.2%</td>
<td>1.12</td>
<td>20.6%</td>
</tr>
<tr>
<td>Monday</td>
<td>3.48</td>
<td>3.9%</td>
<td>1.28</td>
<td>26.4%</td>
</tr>
<tr>
<td>Site B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saturday</td>
<td>5.16</td>
<td>4.7%</td>
<td>1.52</td>
<td>25.6%</td>
</tr>
<tr>
<td>Wednesday</td>
<td>5.09</td>
<td>4.4%</td>
<td>1.58</td>
<td>24.8%</td>
</tr>
<tr>
<td>Monday</td>
<td>4.75</td>
<td>5.2%</td>
<td>1.60</td>
<td>32.1%</td>
</tr>
</tbody>
</table>
### TABLE 4. Pair-wise Regression Results for Missing Days

<table>
<thead>
<tr>
<th>Missing</th>
<th>May 2012 5 minute</th>
<th>May 2012 20 seconds</th>
<th>January 2012 5 minute</th>
<th>January 2012 20 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MAPE</td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>Site A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saturday</td>
<td>14.41</td>
<td>13.9%</td>
<td>1.80</td>
<td>32.4%</td>
</tr>
<tr>
<td>Wednesday</td>
<td>13.63</td>
<td>15.2%</td>
<td>1.60</td>
<td>34.0%</td>
</tr>
<tr>
<td>Monday</td>
<td>11.88</td>
<td>15.5%</td>
<td>1.74</td>
<td>41.2%</td>
</tr>
<tr>
<td>Site B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saturday</td>
<td>9.40</td>
<td>9.5%</td>
<td>1.76</td>
<td>32.5%</td>
</tr>
<tr>
<td>Wednesday</td>
<td>9.76</td>
<td>8.9%</td>
<td>1.75</td>
<td>29.2%</td>
</tr>
<tr>
<td>Monday</td>
<td>10.90</td>
<td>13.9%</td>
<td>1.75</td>
<td>41.5%</td>
</tr>
</tbody>
</table>

### TABLE 5. PMM Results for missing months

<table>
<thead>
<tr>
<th>Missing</th>
<th>5 minute</th>
<th>20 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>Site A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>May 2012</td>
<td>3.82</td>
<td>4.7%</td>
</tr>
<tr>
<td>January 2012</td>
<td>4.44</td>
<td>6.7%</td>
</tr>
<tr>
<td>Site B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>May 2012</td>
<td>5.41</td>
<td>4.8%</td>
</tr>
<tr>
<td>January 2012</td>
<td>6.53</td>
<td>11.4%</td>
</tr>
</tbody>
</table>

### TABLE 6. Pair-wise Regression Results for Missing Months

<table>
<thead>
<tr>
<th>Missing</th>
<th>5 minute</th>
<th>20 seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>Site A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>May 2012</td>
<td>14.45</td>
<td>14.7%</td>
</tr>
<tr>
<td>January 2012</td>
<td>14.75</td>
<td>16.9%</td>
</tr>
<tr>
<td>Site B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>May 2012</td>
<td>17.54</td>
<td>19.3%</td>
</tr>
<tr>
<td>January 2012</td>
<td>23.08</td>
<td>36.3%</td>
</tr>
</tbody>
</table>

### DISCUSSION AND CONCLUSIONS

It is clear from the results in the previous section the largest performance benefits from the PMM multiple imputation method are obtained at higher rates of missingness and at the 5-minute aggregation level. This demonstrates that the PMM method more accurately reflects the true statistical properties of the underlying data, such that the errors present at the 20-second level tend to cancel each other out in aggregation. This is further illustrated by the fact that the variance of the PMM imputed values is consistently between 60% - 80% of the true variance at the 20-second level, while the variance of the pairwise imputed values falls...
between 8% and 27% of the true variance. While the pairwise regression method performs reasonably well at the 5-minute level on random missing patterns, it is increasingly unreliable as the length of the missing interval is increased. The performance of the PMM method, on the other hand, is relatively unaffected by longer missing intervals.

Note that the pairwise linear model is used as the benchmark to illustrated the advantages of the proposed PMM method. Thus, it is clear that the PMM method can overcome many of the limitation of the pairwise linear model. One obvious limitation of the pairwise method is that, in using the median of all pairwise estimates, much of the variance present in the data is smoothed over. That is, the regression to the mean phenomenon is exacerbated by aggregating over multiple regression estimates. The PMM method, by using multiple regression model for each response, does not suffer from this limitation. Likewise, by incorporating the true variability of the measured data in a Monte Carlo procedure, the PMM method better represents the statistical properties of the data, resulting in even greater performance benefits when post imputation aggregation is performed.

The difference between computational complexity of the proposed algorithm and that of pairwise regression or other elementary methods bears some discussion. Using an appropriate statistical software and data management approach, pairwise linear regression can be performed nearly instantaneously for a week of data from a single detector cabinet in a single travel direction. Using the proposed methodology, we typically impute a year’s worth of data at a time for a single detector cabinet in both travel directions, which takes between 30 minutes and 3 hours depending on hardware configuration and rates of missingness. Thus, with reasonably efficient algorithm design, a year of data from 30 detector cabinets (or around 400 detectors) can be imputed in a day or two.

While the results reported here demonstrate that the proposed algorithm outperforms pairwise regression (e.g. (8) and (9)), there are a number of algorithms described in recent literature that would likely do the same in some scenarios. The contribution of this paper, then, can be summarized as follows: First, while direct comparison is not objective due to differences in testing methodology, data, and aggregation levels, the imputation results obtained through the proposed algorithm compare very favorably with other methods described in literature (e.g. (8) and (9)). Second, unlike any other imputation methods identified in literature, the proposed algorithm for volume imputation was shown to perform well on multiple challenging missing data scenarios including high rates of missing at random, missing entire days, and missing entire months. Third, while the multiple imputation scheme is more computationally complex than pairwise regression, the underlying imputation model (i.e. PMM) is readily interpretable, and can be easily understood by anyone with knowledge of linear regression.
ACKNOWLEDGEMENTS

This work was completed with the support of the Washington State Department of Transportation and the Pacific Northwest Transportation Consortium (PacTrans)

REFERENCES


