Markov Chain Monte Carlo Multiple Imputation for Incomplete ITS Data Using Bayesian Networks

Daiheng Ni and John D. Leonard II

Daiheng Ni
(Corresponding author)
School of Civil and Environmental Engineering
Georgia Institute of Technology
Mailing: 331687 Georgia Tech Station, Atlanta, GA 30332
Phone: 404-385-0567
E-mail: gte710q@prism.gatech.edu

John D. Leonard II
Georgia Institute of Technology
School of Civil & Environmental Engineering
Atlanta, GA 30332-0355
Phone: (404) 894-2360
Fax: (404) 894-2278
john.leonard@ce.gatech.edu

Words: 5,270
Tables: 0 × 250
Figures: 8 × 250
Total: 7,270

Submitted for presentation and publication at the
Transportation Research Record
July 12, 2004
Abstract.

The rich ITS data is a precious resource for transportation researchers and practitioners. However, the usability of such resource is greatly limited by the issue of data missing. A lot of imputation methods have been proposed in the past decade. However, some issues are still not or not sufficiently addressed. For example, the missing of entire records, temporal correlation in observations, natural characteristics in raw data, and unbiased estimates for missing values. With these in mind, this paper proposes an advanced imputation method which is based on the recent development in other disciplines, especially applied statistics. It uses a Bayesian network to learn from the raw data and a Markov chain Monte Carlo technique to sample from the probability distributions learned by the Bayesian network. On the other hand, it imputes the missing data multiple times and makes statistical inference about the result. In addition, it incorporates a time series model so that it allows data missing in entire rows – an unfavorable missing pattern that is frequently seen in ITS data. Empirical study shows that the proposed method is robust and very accurate. It is ideal for use as a high-quality imputation method for off-line application.

Keywords: Bayesian networks, Markov chain Monte Carlo, Gibbs sampler, Data augmentation, Multiple imputation, Time series, Incomplete data, Intelligent transportation systems (ITS).
INTRODUCTION

Intelligent transportation systems (ITS) have been deployed nationwide and, as a result, ITS data are accumulating at traffic management centers at a rate of megabytes to even gigabytes on a daily basis. This provides transportation researchers and practitioners a favorable position to examine the transportation system under analysis. However, a problem that bothers users of these ITS data is that such data typically suffer from missing data issue, while most analysis procedures assumes complete data. Missing data can be the result of missing values or erroneous values with the former being blank in observations and the latter being observations that are not physically meaningful (e.g., an observation with 10 vehicles, 0 speed, and 0 density during a 20-second interval). A common practice of treating incomplete data is the deletion of records that contain missing value(s), which may greatly reduce the size of useable data. For example, for a dataset with 1000 records and 4 variables, a missing rate of 5 percent for each variable may result in a loss of 20 percent of useable data in the worst case. Chandra and Al-Deek (1) reported a 15 percent missing rate on loop detectors' data on Interstate I-4. An empirical study on the ITS data collected by the Georgia NaviGAtor system at GA 400 showed an average missing rate between 4 to 14 percent.

Fortunately, recent studies showed that it is possible to solve the incomplete data problem by means of imputation – a process of filling in the missing values with estimates called imputes. A lot of imputation methods were proposed and these methods can generally be classified as ad hoc methods, traditional methods, and statistically-principled methods. Ad hoc methods involved some naïve techniques such as replacing the missing values with historical data (or data from neighboring locations) or filling in the missing values with historical averages (or the averages over neighboring locations). It was recognized later that replacement and average techniques might be too arbitrary and smoothened techniques such as linear temporal (or spatial) interpolation (or extrapolation) were developed. These techniques, called nearest neighbors, used data of one or more of the neighboring detectors to guess the missing value just as patching a hole on a piece of cloth. More recent research found that linear interpolation might also be subject to arbitrary error and that data of detectors beyond the nearest neighbors are able to provide useful information as well. This gave birth to traditional methods such as the Kalman filter method (2) and time series (ARIMA) method (3). Current development of imputation techniques is moving predominantly on a statistically-principled track. For example, Conklin and Smith (4) used local lane distribution patterns to estimate missing data. Chen et al. (5) proposed a linear regression-based method for imputing missing values using neighboring cell values in the time-space lattice. Smith et al. (6) reported a two-tiered approach where a less time-consuming technique, i.e., the historical averages approach, was used to impute in real-time during daytime, while a computationally intensive but more advanced technique - the expectation maximization (EM) approach - was employed to fine-tune the imputes and overwrite them during night. Smith et al. (7) also discussed a promising imputation method based on data augmentation (DA) which performed better than regular methods. Chandra and Al-Deek (8) compared a class of methods, including multiple regression methods, time series methods, and pair-wise regression methods, and tested their feasibility and accuracy. They found that pair-wise quadratic method with selective median performed better than the rest of the methods.

Most of the above methods are deterministic in nature because the outcome of each imputation is fixed and only one impute is estimated for each missing value. Therefore, these methods are called single imputation methods. However, such methods typically fail to reflect sampling variability about the true values, nor additional uncertainty about the right values to impute. On the other hand, statistically-principled methods such as DA typically assume that the data come from independent, identical (IID) draws from some multivariate probability distribution. This might not be true in ITS data because such data are typically time series data and temporal correlation is an important issue. Moreover, some advanced imputation methods assume that values are missing at random. However, ITS data typically have a worse case in that, if a value is missing, it is highly likely that the whole record is missing. For example, if traffic density is not recorded at some time point, very often this is due to system malfunction and other data at the same time, such as volume and speed, are not recorded as well.

To address the above issues, this paper proposes a stochastic method which shows some significant improvements over the existing methods. First, the method is based on Bayesian networks, an approach that is become increasingly popular in problem solving such as reasoning under uncertainty. A Bayesian network is particularly suited for imputing incomplete data because it offers a natural way to encode the correlations between and within variables. Second, a Markov chain Monte Carlo (MCMC) technique is employed to sample from the probability distributions learned by the Bayesian network. These samples, combined with observed data, are then
used to update the Bayesian network and new samples are drawn from the updated Bayesian network. The above process repeats until convergence is achieved. In short, the basic idea of this method is to solve a difficult incomplete data problem by solving manageable complete data problems iteratively and progressively. Third, the method contains a time series component to efficiently account for temporal correlation in ITS data. In this case, the assumption of missing values at random can be relaxed to missing records at random. Fourth, the concept of multiple imputation (9, 10) is introduced such that multiple draws are made for each missing value. By this, not only can one produce an unbiased estimate, but also have an understanding of how much confidence to put on the imputation result. Moreover, natural characteristics, such as natural variability and the relationships among variables, of the data are preserved.

This paper is arranged as follows. Theoretic development of the method is presented in the next section. This is followed by discussion on some implementation issues. The proposed method is applied to empirical data collected from GA 400 by Georgia NaviGAtor system and imputation results are presented. Finally conclusions are drawn and recommendations are made.

THEORY

This section provides the necessary theoretical background and highlights how these theories are applied to the problem of incomplete ITS data.

Bayesian Network

A Bayesian network (11) is a graphical model that encodes probabilistic relationships among variables of interest. The model can be represented as a directed acyclic graph (DAG) in which nodes represent variables and directed arcs represent dependencies, i.e., conditional probabilities. Missing arcs implies conditional independence. Thus the Bayesian network represents the joint distribution of all variables by means of independencies and conditional probabilities.

Figure 1 shows an example Bayesian network. A node Yᵢ is, at the same time, a variable (discrete or continuous). A directed arc represents the probability of a child node conditioning on its parent node, e.g., p(Y₂ | Y₁).

The joint probability distribution of all variables can be represented as:

\[ p(Y₁, Y₂, Y₃, Y₄) = p(Y₁)p(Y₂ | Y₁)p(Y₃ | Y₂)p(Y₄ | Y₃) \]

Generally, Let \( \{Y | Yᵢ \in Y, i = 1, 2, \ldots, n\} \) denote the set of variables, and \( \text{pa}(Yᵢ) \) denote the parent(s) of \( Yᵢ \). The joint probability distribution of \( Y \) can be represented in Markov condition (or factorization) as:

\[ p(Y) = \prod_{i=1}^{n} p(Yᵢ | \text{pa}(Yᵢ)) \]

The most salient feature of Bayesian networks is their capability of learning from data, either learning parameters (e.g., conditional probabilities) or learning structure (e.g., nodes and arcs of a Bayesian network) or both. Considering whether a dataset is complete and whether the network structure is known, applications of Bayesian networks generally fall into the following matrix:

<table>
<thead>
<tr>
<th></th>
<th>Complete data</th>
<th>Incomplete data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known structure</td>
<td>Network structure is specified</td>
<td>Network structure is specified</td>
</tr>
<tr>
<td></td>
<td>- Need to estimate parameters</td>
<td>- Need to estimate parameters</td>
</tr>
<tr>
<td></td>
<td>Data do not contain missing values</td>
<td>Data contain missing values</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Need to impute missing values</td>
</tr>
<tr>
<td>Unknown structure</td>
<td>Network structure is not specified</td>
<td>Network structure is not specified</td>
</tr>
<tr>
<td></td>
<td>- Need to select arcs &amp; estimate parameters</td>
<td>- Need to select arcs &amp; estimate parameters</td>
</tr>
<tr>
<td></td>
<td>Data do not contain missing values</td>
<td>Data contain missing values</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Need to impute missing values</td>
</tr>
</tbody>
</table>
Our interest is the cell categorized by “Known structure” and “Incomplete data” because the traffic data collected from ITS systems typically contain erroneous or missing values but the network structure (i.e., variables and their dependencies) is known beforehand. Therefore, the basic idea of imputation using Bayesian networks is to learn a model from the data, predict values for the missing ones, and update the model with the improved data and update imputes with the improved model. This process may need to iterate many times to achieve convergence.

For our problem, let \( Y_{\text{obs}} \) denote the observed part of \( Y \) and \( Y_{\text{mis}} \) denote the missing part of \( Y \), so that \( Y = (Y_{\text{obs}}, Y_{\text{mis}}) \). Let \( G \) denote the graphical model of a Bayesian network according to which the joint probability distribution can be factored. An inference with the Bayesian network is to compute the conditional distribution of \( Y_{\text{mis}} \) given \( Y_{\text{obs}} \) and the Bayesian network. A general expression for the inference is:

\[
p(Y_{\text{mis}} | Y_{\text{obs}}, G) = \frac{p(Y_{\text{mis}} | G) p(Y_{\text{obs}} | Y_{\text{mis}}, G)}{\int p(Y_{\text{obs}} | Y_{\text{mis}}, G) p(Y_{\text{mis}} | G) dY_{\text{mis}}}
\]

by which we can update our knowledge in the missing data \( Y_{\text{mis}} \) given the observed data \( Y_{\text{obs}} \) and the Bayesian network \( G \). The left-hand term, \( p(Y_{\text{mis}} | Y_{\text{obs}}, G) \) is known as the posterior probability, or the probability of \( Y_{\text{mis}} \) after considering the effect of \( Y_{\text{obs}} \) and \( G \). The term \( p(Y_{\text{mis}} | G) \) is called the prior probability of \( Y_{\text{mis}} \) given \( G \) alone. The term \( p(Y_{\text{obs}} | Y_{\text{mis}}, G) \) is called the likelihood which gives the probability of the evidence assuming the missing data \( Y_{\text{mis}} \) and that \( G \) is true. Finally, the denominator is independent of \( Y_{\text{mis}} \) and can be regarded as a normalizing or scaling factor. Obtaining the posterior distribution \( p(Y_{\text{mis}} | Y_{\text{obs}}, G) \) often involves the integration of high-dimensional functions which can be computationally very difficult. A simple and frequently employed method to this NP-hard problem (12) is the approximation method based on Markov chain Monte Carlo (MCMC) (13, 14).

**Markov chain Monte Carlo**

MCMC is a collection of methods to simulate direct draws from some complex distribution of interest. MCMC is so-named because one uses the previous sample value to randomly generate the next sample value, creating a Markov chain where the current value of the sample depends only on its previous value and is independent of its further earlier values.

**Gibbs Sampler**

Among the many choices, Gibbs sampler (15) is one of the simplest MCMC methods. Introduced in the context of image processing, the Gibbs sampler method decomposes a complex high-dimensional joint distribution to a series of univariate conditional distributions which are far easier to simulate than the complex joint distribution and usually have simple forms. Thus, one simulates \( n \) random variables sequentially from the \( n \) univariate conditionals rather than generating a single \( n \)-dimensional vector in a single pass using the full joint distribution.

Suppose a random vector \( Y \) can be partitioned into \( n \) subvectors, i.e., \( Y = (Y_1, Y_2, ..., Y_n) \). Let \( p(Y) \) denote the joint distribution to be simulated. This distribution can be simulated by iterative draws from conditional distribution of each subvector given all other subvectors which take their most recent drawn values. Suppose, at step \( t \), the state of \( Y \) is

\[
Y^{(t)} = (Y_1^{(t)}, Y_2^{(t)}, ..., Y_n^{(t)}),
\]

the state at step \( (t + 1) \)

\[
Y^{(t+1)} = (Y_1^{(t+1)}, Y_2^{(t+1)}, ..., Y_n^{(t+1)})
\]

can be obtained according to the following procedure:

\[
Y_1^{(t+1)} = p(Y_1 | Y_2^{(t)}, Y_3^{(t)}, ..., Y_n^{(t)})
\]

\[
Y_2^{(t+1)} = p(Y_2 | Y_1^{(t+1)}, Y_3^{(t)}, ..., Y_n^{(t)})
\]
\[ Y_{n}^{(t+1)} = p(Y_{n} \mid Y_{1}^{(t+1)}, Y_{2}^{(t+1)}, \ldots, Y_{n-1}^{(t+1)}) \]

Repeat the above process to obtain \( Y^{(t+2)}, Y^{(t+3)}, \) and so on. The sequence \( \{ Y^{(t)} : t = 1, 2, 3, \ldots \} \) forms a Markov chain whose stationary distribution is equal to \( p(Y) \), i.e., \( Y^{(0)} \rightarrow Y \) in distribution as \( t \rightarrow \infty \).

**Data Augmentation**

Data augmentation (16) is a special case of Gibbs sampler where a random vector \( Y \) is partitioned into two parts, \( Y = (Y_{1}, Y_{2}) \). Suppose the current state of \( Y \) is \( (Y_{1}^{(t)}, Y_{2}^{(t)}) \), then the two-component Gibbs sampler updates the state according to the following procedure:

- Draw \( Y_{1}^{(t+1)} \) from the conditional distribution \( p(Y_{1} \mid Y_{2}^{(t)}) \);
- Draw \( Y_{2}^{(t+1)} \) from the conditional distribution \( p(Y_{2} \mid Y_{1}^{(t+1)}) \);

Return to our problem of incomplete ITS data, one of the component typically corresponds to the parameter of interest, \( \Theta \), and the other corresponds to the missing data, \( Y_{\text{mis}} \). The data augmentation algorithm then iterates between an imputation step (I-step) which imputes the missing data given the probability distribution conditional on observed data, the current parameter value, and the Bayesian network and a Posterior step (P-step) which updates the parameter value by the probability distribution conditional on the observed data, the newly imputed missing data, and the Bayesian network:

\[
\text{I-Step: } \quad Y_{\text{mis}}^{(t+1)} = p(Y_{\text{mis}} \mid \Theta^{(t)}, Y_{\text{obs}}, G)
\]

\[
\text{P-Step: } \quad \Theta^{(t+1)} = p(\Theta \mid Y_{\text{mis}}^{(t+1)}, Y_{\text{obs}}, G)
\]

Given an initial value \( \Theta^{0} \) and the Bayesian network \( G \), and repeat the above I- and P-steps, one obtains a Markov chain \( \{ (\Theta^{(t)}, Y_{\text{mis}}^{(t)}) : t = 1, 2, 3, \ldots \} \) whose stationary distribution is \( \pi(\Theta, Y_{\text{mis}} \mid Y_{\text{obs}}, G) \). The sequence \( \{ \Theta^{(t)} : t = 1, 2, 3, \ldots \} \) then has \( \pi(\Theta \mid Y_{\text{obs}}, G) \) as its stationary distribution and \( \{ Y_{\text{mis}}^{(t)} : t = 1, 2, 3, \ldots \} \) then has \( \pi(Y_{\text{mis}} \mid Y_{\text{obs}}, G) \) as its stationary distribution.

**Time Series**

Theoretical development up to this point seems to have solved our problem. However, the above method is based on several implicit assumptions, among which the independent, identical distribution (IID) assumption calls for special attention. More specially, suppose one has a rectangular dataset \( Y = \{ y_{\tau} : \tau = 1, 2, \ldots, T \} \) whose rows are samples (or records) and columns are variables, the IID assumption states that all the samples come from a common multivariate probability distribution and every sample is independent of every one else. However, traffic data collected by ITS systems are typically temporally correlated and, thus, violate the independent assumption. Therefore, a time series component is called in the method.

In addition to the IID assumption, the above method assumes that data is missing at random (MAR). This assumption states that, loosely speaking, the probability that a value is missing has nothing to do with the value itself. In the rectangular data model, this assumption typically results in missing data randomly interspaced in the rectangle. If, however, values in a record are totally missing, this record is typically removed from consideration because it contributes no more information to the imputation but slowing down convergence (17). However, such a missing pattern happens to be the case mostly seen in ITS data. For example, if traffic count at some time point is missing, it is highly probable that the ITS system is not functioning well and, as a result, speed, density, and other data at the same time may not be recorded as well. This, again, calls for a time series component in the model so that the missing rows can be predicted by time series and refined by imputation.
To deal with the issues discussed above, an autoregressive integrated moving average time series model (19), ARIMA(p, d, q), is considered here:

$$\Phi(B)\Delta^d y_\tau = \Omega(B)\epsilon_\tau$$

where p is the order of the autoregressive part, d is the order of the differencing, q is the order of the moving-average process, $\tau$ is time indices, B is the backshift operator, i.e.,

$$By_\tau = y_{\tau-1}$$

$\Phi(B)$ is the autoregressive operator, represented as a polynomial in the backshift operator, i.e.,

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 ... - \phi_p B^p$$

$\Omega(B)$ is the moving-average operator, represented as a polynomial in the backshift operator, i.e.,

$$\Omega(B) = 1 - \omega_1 B - \omega_2 B^2 ... - \omega_q B^q$$

$\Delta$ is the differencing operator, i.e.,

$$\Delta^d = (1 - B)^d$$

and $E = \{\epsilon_\tau\}$ is a white noise, i.e., normal distribution with 0 mean and variance $\Sigma$:

$$\epsilon_\tau \sim N(0, \Sigma).$$

Multiple imputation

Most existing imputation methods estimate a single value for a missing datum. As a result, there is no direct measure for the variability and uncertainty about this estimate, so there is usually a lack of confidence on the part of the user in using the imputation result. Take a different perspective. Let's think of the missing value as the unknown parameter whose true value is of our interest. The standard statistical way to estimate the true value is to draw some random samples from the population and make inference from these samples. Single imputation represents only one
draw which is highly likely to give a biased estimate. A better and statistically sound approach would be to take multiple draws and make an inference based on them. By this, not only can one produce an unbiased estimate, but also have an understanding of how much confidence to put on the result. This is the premise based on which a multiple imputation (9, 10) technique works.

Return to our problem of incomplete ITS data. Multiple imputation for the Gibbs sampler with ARIMA model can be formulated as follows. Run the Gibbs sampler for a sufficiently long time, chop off the first portion, say tens or hundreds of iterations, which is the "warming-up" period to achieve convergence, and take samples after every k iterations to obtain m sets of imputes. Alternatively, one can make m shorter runs of the Gibbs sampler, obtain a set of imputes from each run after convergence. These m sets of imputes are then used to generate m sets of complete data which are then combined to create an overall set of complete data as well as to estimate the uncertainty about the imputation.

IMPLEMENTATION

Here we consider the problem of incomplete multivariate ITS data where a record contains values of several variables and some records are missing.

The Data Model

Let $Y$ denote the $T \times N$ matrix of a complete dataset as schematically represented in Figure 2. The dataset consists of a N-dimensional time series:

$$Y = \{y_t : \tau = 1,2,...,T\} \text{ and } y_{\tau,i} = \{y_{\tau,i} : i = 1,2,...,N\}$$

of which rows are time series samples and columns are variables (e.g., volume, speed, density, etc.). Some values of this matrix are not observed and a question mark denotes a missing value. If data are missing at reasonable rates and the missing exhibits a random pattern, a regular multiple imputation method for multivariate data based on data augmentation (17, 18) suffices. Here, we consider a more unfavorable case where values in some records are totally missing, but it is assumed that the missing records occur in a random pattern, so the missing mechanism still conforms to the MAR assumption in a loose sense. Note that imputation for missing records in a systematical manner (e.g., extended periods of time) represents the most unfavorable case which deserves further research efforts.

The Bayesian Network

The Bayesian network, in a Markovian sense, for the Gibbs sampler with time series component is represented in Figure 3. In this network, we are not interested in the joint distribution of all variables, but the stationary conditional distributions:

$$\pi(Y_{\text{mis}} \mid Y_{\text{obs}}, G) = p(Y_{\text{mis}} \mid \Theta^{(i)}, Y_{\text{obs}}, G) \text{ as } t \to \infty, \text{ and}$$

$$\pi(\Theta \mid Y_{\text{obs}}, G) = p(\Theta \mid Y_{\text{mis}}^{(i)}, Y_{\text{obs}}, G) \text{ as } t \to \infty$$

The Algorithm

The general idea of the algorithm can be stated as an iterative procedure that involves:

- Learning a Bayesian network from the data.
- Predicting values for the missing ones.

More specifically, the algorithm of Markov chain Monte Carlo multiple imputation based on Bayesian network is specified as follows.
1. Determine ARIMA parameters \( p, d, \) and \( q \) based on prior knowledge and \( Y_{\text{obs}} \).

2. Initialize Bayesian network parameters \( \Phi^{(0)}, \Sigma^{(0)}, \Omega^{(0)}, \) and \( E^{(0)} \).

3. Loop:
   1) Impute the missing values \( Y_{\text{mis}} \) given \( Y_{\text{obs}} \), and the Bayesian network \( G \) with parameters \( \Phi, \Sigma, \Omega, \) and \( E \).
   2) Learn Bayesian network parameters from \( Y_{\text{com}} = (Y_{\text{mis}}, Y_{\text{obs}}) \)
      (1) Draw \( \Phi \) given \( Y_{\text{obs}} \), the latest \( Y_{\text{mis}} \), \( \Sigma, \Omega, \) and \( E \).
      (2) Draw \( \Sigma \) given \( Y_{\text{obs}} \), the latest \( Y_{\text{mis}} \), the latest \( \Phi, \Omega, \) and \( E \).
      (3) Draw \( \Omega \) given \( Y_{\text{obs}} \), the latest \( Y_{\text{mis}} \), the latest \( \Phi, \) the latest \( \Sigma, \) and \( E \).
      (4) Draw \( E \) given \( Y_{\text{obs}} \), the latest \( Y_{\text{mis}} \), the latest \( \Phi, \) the latest \( \Sigma, \) and the latest \( \Omega \).
   4. Run 3 sufficiently long to obtain multiple imputations of \( Y_{\text{mis}} \) after convergence, or run 3 multiple times to obtain an imputation of \( Y_{\text{mis}} \) from each run after convergence.

5. Combine the multiple imputations of \( Y_{\text{mis}} \) to construct an overall imputation of \( Y_{\text{mis}} \) and make inference about the imputation.

**Convergence**

As stochastic methods, Markov chain Monte Carlo algorithms converge to probability distributions. Convergence by \( k \) iterations means that \( (\Theta^{(t)}, Y_{\text{mis}}^{(t)}) \) is independent of \( (\Theta^{(t+k)}, Y_{\text{mis}}^{(t+k)}) \) for any \( t > 0 \). Also, it is sufficient for the distribution of \( \Theta^{(t)} \) to have converged to \( \pi(\Theta | Y_{\text{obs}}) \) and the distribution of \( Y_{\text{mis}}^{(t)} \) to \( \pi(Y_{\text{mis}} | Y_{\text{obs}}) \). In applications of multiple imputation for missing data, the goal is to simulate independent draws from the stationary distribution of \( Y_{\text{mis}}^{(t)}, \pi(Y_{\text{mis}} | Y_{\text{obs}}) \), and assessment of convergence can be performed based on monitoring successive outputs of \( Y_{\text{mis}}^{(t)} \).

Typically, the rate of convergence depends on missing rates and starting values or distribution. It is straightforward that higher missing rates typically result in longer time to converge. Even within a single application, starting values at the tails of observed-data posterior distribution normally lead to more iterations than starting values close to the center of the distribution. On the other hand, it is of interest to roughly estimate the number of iterations, \( k \), so that, after a warming-up period that phases out the influence of starting values or distribution, one allows enough cycles between multiple imputations to ensure that they are statistically independent.

Among various methods, output analysis is a handy tool to identify the warming-up period. For a multidimensional problem, output analysis can be performed based on target variables such as components of \( Y_{\text{mis}} \) or some scalar function of \( Y_{\text{mis}} \). A time series plot of the target variables is particularly helpful to identify the warming-up period. To further investigate the relationships among outputs of successive iterations, the autocorrelation function (ACF) is typically employed.

**RESULTS**

**The Dataset**

The dataset used in this study was collected from GA 400 by Georgia NaviGAtor - the ITS system of Georgia. Traffic conditions on the study site were monitored by video cameras which were deployed approximately every one third mile of the road in each direction. Each camera constitutes an observation station and watches all the lanes at this location. An image processing software program was running in the background to extract traffic data from the
videos. Traffic conditions were sampled every 20 seconds and each record contains values of a variety of variables, among which traffic count, speed, and density were of major interest in this study. Empirical studies showed that values of count and speed were generally accurate, but density was less reliable due primarily to the facts that was is calculated from some equation that lacked theoretical ground and it exhibited the highest variability in the raw data.

Empirical tests were performed on several days an over several stations and the test results are consistent over time and station. Without losing generality, the following discussion focuses on the test performed on the observations collected at station 4000036 on Tuesday, Sept. 10th, 2002. This station was located at Pitts Rd. on the northbound GA 400 between exits 5B and 6. There were 4 lanes at this location. The traffic samples started at 5:31:00 and ended at 23:51:00, containing a total of 3301 records. A procedure was devised to generate an array of non-repeating random numbers based on a pre-specified missing rate. These random numbers were then used to delete the corresponding records in the complete data. The proposed imputation method was then applied on the resulted incomplete data and evaluation was performed by comparing the imputed data and the true data. Note that an attractive feature of the imputation method is that it is able to learn from the correlations within and among variables and estimate imputes for all the variables simultaneously.

Imputation and Aggregation

The raw data collected from the field were in 20-second intervals and exhibited high variability. Most traffic analysis procedures works on merged data such as those in 5-minute or 15-minute intervals. Therefore, there are basically two schools of thoughts regarding the order of aggregation and imputation. Smith et al (7) aggregated 1-minute data into 10-minute intervals and then performed imputation, while Chandra and Al-Deek (8) aggregated 30-second data into 5-minute intervals before imputation. Aggregation-before-imputation seems to help reduce variance, improve computation efficiency, and average out noise. However, this approach has its limitations. In practice, one rarely has control on where the missing values occur, so aggregation-before-imputation might accidentally incorporate missing values and/or pre-imputed values into the aggregated data based on which the intended imputation is performed. This means that one might be working on a modified data rather than the raw data and the aggregation may alter the natural characteristics encoded in the raw data. Moreover, this approach may result in loss of useable information and/or introducing extra error in the aggregated data. With these issues in mind, this paper is going to follow an imputation-before-aggregation approach, i.e., imputation is made directly on the 20-second raw data and the imputed data are then aggregated into 5-minute intervals for practical consumption. By this way, one does not need to worry about where the missing data occur. Though 20-second data exhibit higher variability than aggregated data, but they contain rich information regarding the correlations within and among variables and such information is the basis on which the proposed imputation works. On the other hand, aggregation after imputation is able to give a more reliable and clean trend because all the missing values have been filled with educated estimates.

Imputation Results

Imputation results are presented in this subsection. A 30 percent missing rate was simulated in the imputation process and each missing value was imputed 5 times, resulting in 5 sets of imputed complete data based on which statistical inference was made. An overall imputation was generated by taking the mean of the 5 sets of data and the variability of the multiply imputed data as well as the uncertainty introduced by the imputation process were also assessed.

Analysis of Imputed Values

A natural starting point of assessing imputation quality is the analysis of imputed values. Figure 4 gives diagonal plots where imputed values are plotted against actual values and series plots where imputed values are plotted on top of actual values. The figure shows that speed has the best fit because, in the diagonal plot, the data points are closely distributed along the 45 degree line which indicates an ideal fit and, in the series plot, imputed values match the actual values very well. Traffic count also indicates a good fit even though the diagonal plot shows more deviation. Density seems to give the least satisfactory result. However, this is due primarily to the facts that raw data of this variable are not vary reliable, as stated above, and it exhibits the highest variability. Given these, a trend of fit can still be seen in its plots.

To evaluate how the proposed imputation method preserves the natural characteristics of the raw data, the natural variability within variables and the relationships between variables were examined. The right half of Figure 4
shows that, for each variable, the imputed values match the actual values and both exhibit approximate the same variability. Figure 5 plots the speed-flow and flow-density relationships of the actual and the imputed values. It shows that the imputed values preserve very well the original relationships among variables in the raw data.

Convergence

An important issue concerning the applicability of the proposed method is the rate of convergence. By assessing convergence, one obtains a basic understanding on how long the warming-up period takes and how many iterations is needed for two imputations apart at least that far to be independent. As discussed above, the analysis is performed by means of autocorrelation function (ACF) and time series plot of imputation result of successive iterations. Figure 6 shows these plots based on the mean of imputed traffic count where the time series plot shows result of the first 100 iterations. It can be seen from the ACF that sample correlation disappears for imputations more than 3 iterations apart and the time series plot indicates a warming-up period of 3-5 iterations.

Imputed Complete Data

To make inference on multiple imputation results, Rubin (9) proposed a procedure, based on imputed complete data, which yields the following results in this study:

<table>
<thead>
<tr>
<th></th>
<th>Traffic count</th>
<th>Speed (mi/hr)</th>
<th>Density (veh/mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True mean</td>
<td>30.22</td>
<td>60.52</td>
<td>127.83</td>
</tr>
<tr>
<td>Grand mean of all imputations</td>
<td>30.24</td>
<td>60.53</td>
<td>128.43</td>
</tr>
<tr>
<td>Within-imputation variance</td>
<td>156.22</td>
<td>59.22</td>
<td>6618.60</td>
</tr>
<tr>
<td>Between-imputation variance</td>
<td>0.0002</td>
<td>0.00004</td>
<td>0.01167</td>
</tr>
<tr>
<td>Total variance</td>
<td>156.22</td>
<td>59.22</td>
<td>6618.70</td>
</tr>
<tr>
<td>95% confidence intervals</td>
<td>(5.74, 54.74)</td>
<td>(45.45, 75.62)</td>
<td>(0, 287.89)</td>
</tr>
</tbody>
</table>

It can be seen that there is strong evidence that the imputed means are not statistically different than the true means. This implies that the imputation result is unbiased. On the other hand, within-imputation variance is very high comparing with between-imputation variance. This implies that the high total variance is due primarily to the natural variability of the dataset, as is reflected by within-imputation variance, and the uncertainty caused by missing data is very low, as is reflected by between-imputation variance.

To give an overview of the imputation quality, time series plots of imputed and actual complete data are displayed in Figure 7 which generally confirms our findings with the imputed values.

Merged Complete Data

To smooth out the high variability in the 20-second data, both the imputed and the actual data were merged into 5-minute intervals, as comparable to similar studies. Assessment of imputation quality was also performed at this aggregation level to highlight the comparability with those studies. Figure 8 shows diagonal plots and time series plots of the imputed and the actual data after aggregation. It can be seen that both sets of data fit perfectly well at this level.

To quantitatively evaluate the imputation quality, Imputation error was analyzed and results are listed below. Measures computed included mean of error (Mean), error standard deviation (Std), root mean squared error (RMSE), mean absolute percentage error (MAPE), and 95% confidence interval. The results show that the imputation error is not statistically different than 0, which confirms again that the proposed method is able to impute unbiased estimates for the missing data. In absolute terms, the imputation method results in a root mean squared error of 1.5 vehicles for traffic count, 0.59 mi/hr for speed and 15.3 veh/mi for density. Note that these measures are based on 4 lanes and in 5-minute intervals. Put it in relative terms, the mean absolute percentage error is below 4.3% for traffic count, 0.8% for speed, and 9.7% for density.
### Computational Efficiency

The computational efficiency of the imputation method largely depends on the size of the problem. In practice, one may need to decompose a large dataset into a series of small batches and apply the imputation method to each of them. There are two competing aspects in this process. One is to keep the batch size small to achieve computation efficiency. The other is to use large batch size so that the method can learn from more information and possibly yields better estimation. Experiences in this study show that a batch size of 100 might be appropriate. For example, the data in this study are in 20-second intervals, so a batch covers 2000 seconds. It is estimated that 19 seconds are needed to impute a batch and this means the method is running slightly faster than real time and, during the 2000 seconds, approximately 100 stations can be processed, which happens to be about the same size of the study site.

Tests on other days and stations are also performed and test results are consistent with what are reported above. This shows that the imputation method is robust over time and location.

### CONCLUSION

This paper proposes an advanced imputation method to deal with incomplete ITS data. It uses a Bayesian network to learn the correlations encoded between and within variables. It also uses a Markov chain Monte Carlo technique to simulate random draws from the probability distributions learned by the Bayesian network. The incomplete data problem is then solved by iteratively and progressively solving complete problems until the algorithm converges. On the other hand, each missing value is imputed multiple times so that multiple sets of imputations are produced. An overall imputation is generated from the multiple sets and statistical inference can be made. In addition, the method also incorporates a time series model so that the missing values are predicted based on the trend in the raw data and then further fine-tuned based on the correlations embedded in the data. By this way, the issue of data missing in entire records is effectively accounted for. Therefore, the proposed method is based on a sound theoretical ground.

In practical terms, the proposed method follows an imputation-before-aggregation approach in that it works directly on the raw data (20-second intervals in this study) which eliminates the possibility of introducing extra error during imputation. The imputed data are then aggregated into longer intervals (5-minute in this study) for practical consumption. This is the natural way of performing imputation because an accurate imputation performed on inaccurate data doesn’t make any sense. On the other hand, empirical study shows that the proposed method is very accurate. Graphical comparison shows a close fit between the actual data and the imputed data and quantitative assessment reveals a very small imputation error. In addition to its high accuracy, this method is quite robust. For example, the above results are achieved at a missing rate of 30 percent, which is pretty high in real situation. Moreover, this method is capable of producing unbiased estimates for the missing data and preserving the natural characteristics (e.g., variability within variables and relationships among variables) of the raw data. This is very attractive because, for example, a seemingly accurate imputation but with the wrong flow-speed-density relationship is useless to traffic analysts.

There are three types of missing data problems: missing values in random, missing records (a record consists of all values in a row) in random, and missing records systematically (e.g., in extended periods of time) and they are increasingly difficult in that order. This research deals with the second type, i.e., missing records in random, and it is recognized that the third type is a more difficult and critical scenario to deal with. It is also recognized that, due to its very nature, the imputation method might be computationally intensive. Hopefully, empirical studies show that it is still possible to apply the method online. As a last note, though this method is proposed using incomplete ITS data as an illustrative example, the method applies to other general problems with similar nature.

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Speed (mi/hr)</th>
<th>Density (veh/mi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.02</td>
<td>0.01</td>
<td>0.60</td>
</tr>
<tr>
<td>Std</td>
<td>1.53</td>
<td>0.59</td>
<td>15.35</td>
</tr>
<tr>
<td>95% C.I.</td>
<td>(-0.18, 0.22)</td>
<td>(-0.07  0.09)</td>
<td>(-1.44  2.64)</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.52</td>
<td>0.59</td>
<td>15.32</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.043</td>
<td>0.008</td>
<td>0.097</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENT

The authors thank the anonymous referees whose insightful comments helped improve the quality of this paper.

REFERENCES

LIST OF TABLES AND FIGURES

Tables

None

Figures

FIGURE 1. An example Bayesian network
FIGURE 2. Multivariate dataset with values fully missing in some records
FIGURE 3. The Bayesian network for Gibbs sampler with time series component
FIGURE 4. Comparison of imputed values and actual values
FIGURE 5. Relationships among variables of imputed values and actual values
FIGURE 6. Assessment of convergence
FIGURE 7. Comparison of imputed complete data and actual complete data (20-second intervals)
FIGURE 8. Comparison of imputed complete data and actual complete data (5-minute intervals)
FIGURE 1. An example Bayesian network
FIGURE 2. Multivariate dataset with values fully missing in some records
FIGURE 3. The Bayesian network for Gibbs sampler with time series component
FIGURE 4. Comparison of imputed values and actual values
FIGURE 5. Relationships among variables of imputed values and actual values
FIGURE 6. Assessment of convergence
FIGURE 7. Comparison of imputed complete data and actual complete data (20-second intervals)
FIGURE 8. Comparison of imputed complete data and actual complete data (5-minute intervals)