University of Massachusetts Amherst

From the Selected Works of Daiheng Ni

March 1, 2005

Multiple Imputation Scheme for Overcoming the Missing Values and Variability Issues in ITS Data

Daiheng Ni, *University of Massachusetts - Amherst*

John D. Leonard II

Angshuman Guin

Chunxia Feng

Available at: https://works.bepress.com/daiheng_ni/7/

Multiple Imputation Scheme for Overcoming the Missing Values and Variability Issues in ITS Data

Daiheng Ni¹; John D. Leonard II²; Angshuman Guin³; and Chunxia Feng⁴

Abstract: Traffic engineering studies such as validating *Highway Capacity Manual* (HCM) models require complete and reliable field data. However, the wealth of intelligent transportation systems (ITS) data is sometimes rendered useless for these purposes because of missing values in the data. Many imputation techniques have been developed in the past with virtually all of them imputing a single value for a missing datum. While this provides somewhat simple and fast estimates, it does not eliminate the possibility of producing biased results and it also fails to account for the uncertainty brought about by missing data. To overcome these limitations, a multiple imputation scheme is developed which provides multiple estimates for a missing value, simulating multiple draws from a population to estimate the unknown parameter. This paper also develops a framework of imputation which gives a broad perspective so that one can relate imputation methods to each other.

DOI: 10.1061/(ASCE)0733-947X(2005)131:12(931)

CE Database subject headings: Intelligent transportation systems; Traffic capacity; Data processing.

Introduction

Validating *Highway Capacity Manual* (HCM) models relies heavily on complete and reliable field data. Intelligent transportation systems (ITS) accumulate a tremendous amount of traffic data on a daily basis and these data could be an ideal resource for HCM model validation. However, a major hurdle in applying these data has been the missing data issue because it sometimes renders an entire dataset useless. Researchers at the Texas Transportation Institute (TTI) reported a missing rate between 16 and 93%. Chandra and Al-Deek (2004) reported a 15% missing rate on loop detectors' data on Interstate I-4. Researchers at the Georgia Institute of Technology reported a missing rate between 4 and 14% on Georgia 400 data.

The American Association of State Highway and Transportation Officials (AASHTO) *Guidelines for Traffic Data Programs* (AASHTO 1992) does not recommend substituting estimated

values for missing or edit-rejected data (i.e., imputation) because this introduces "errors which cannot be quantified." Nevertheless, the limitations of reduced data (after eliminating samples with missing values from the original dataset) has been widely recognized primarily because of its propensity to bias one's view on the target system and lead to erroneous results. In response, various imputation techniques have been developed in the past decade. The majority of the existing imputation techniques propose substituting a missing value with a single value. However, this approach is limited because a single draw may be biased and the uncertainty caused by missing data is not accounted for. To address these issues, a multiple imputation scheme is developed where a missing value is imputed multiple times, simulating multiple draws from a population to obtain an estimate of the unknown parameter.

Contributions of this paper include the following: (1) This paper develops an operable scheme to impute incomplete ITS data based on the original multiple imputation approach developed by Rubin (1987 and 1996). It is interesting to note that this paper seems to be the first, if appropriate, which introduces the multiple imputation approach to address incomplete ITS data problem. (2) This paper presents a framework of imputation for ITS data, and this framework provides readers a favorable perspective to examine the whole body of research in this field and to relate past and current research endeavors to each other. More importantly, the framework can also help identify new imputation techniques by entering proper cells in the framework. (3) This paper discusses the relative advantages of the imputation-first approach and the aggregation-first approach. (4) This paper implements and validates the multiple imputation approach and the study results provide researchers and practitioners a basic understanding of the usefulness of this approach.

¹School of Civil and Environmental Engineering, Georgia Institute of Technology, 790 Atlantic Dr., NW, Atlanta, GA 30332-0355. E-mail: daiheng.ni@ce.gatech.edu

²School of Civil and Environmental Engineering, Georgia Institute of Technology, 790 Atlantic Dr., NW, Atlanta, GA 30332-0355. E-mail: john.leonard@ce.gatech.edu

³URS Corporation, Atlanta Office, Atlanta, GA. E-mail: angshu31@hotmail.com

⁴School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332. E-mail: gt5962a@prism.gatech.edu

Note. Discussion open until May 1, 2006. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on March 25, 2004; approved on January 31, 2005. This paper is part of the *Journal of Transportation Engineering*, Vol. 131, No. 12, December 1, 2005. ©ASCE, ISSN 0733-947X/2005/12-931-938/\$25.00.

Fig. 1. Framework of imputation for ITS data

Review of Existing Techniques for ITS Data Imputation

Early techniques of imputing missing data involved some ad-hoc methods such as replacement or average methods. For example, temporal replacement uses historical data of the same location to replace the missing data and spatial average replaces the missing data using the average of neighboring locations. Later, it was recognized that replacement or average 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 estimate 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 more advanced techniques such as the Kalman filter method (Dailey 1993), time series (ARIMA) method (Nihan 1997), and lane distribution method (Conklin and Smith 2002). Current development of imputation techniques is moving predominantly on a statistically principled track. For example, Chen et al. (2003) proposed a linear regression-based methodology for imputing missing values using neighboring cell values in the time-space lattice. Smith and Babiceana (2004) reported a two-tiered approach where a less time-consuming technique (i.e., the historical averages approach) is used to impute in real time during daytime, while a computationally intensive but more advanced technique (i.e., the expectation maximization (EM) approach) is employed to fine tune the imputes (i.e., estimated values) and overwrite them during the night. Zhong et al. (2004) developed a class of advanced models based on genetic algorithms (GAs), time delay

neural network (TDNN), and locally weighted regression (LWR) and showed higher accuracy than traditional imputation methods. Chandra and Al-Deek (2004) 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 the pair-wise quadratic method with selective median performed better than the rest of the methods.

Framework of Imputation

The application of imputation in traffic data has revealed three dimensions along which imputation techniques evolve. Fig. 1 shows a framework with these dimensions. The first dimension, methodology, is the main theme we discussed above. Examples of methodology are illustrated in the vertical column as “replacement,” “interpolation,” etc. The second dimension, domain, refers to the attributes of the data used to perform imputation and example attributes are listed horizontally as time (e.g., using historical information), space (e.g., using neighboring information), or both (e.g., using both historical and neighboring information). The third dimension, parameter, means which variables are involved in imputation (e.g., flow, speed, density, etc.). With this framework, it is easy to position the above-mentioned imputation methods in the framework and find the relationship among them. For example, Chandra and Al-Deek (2004) is located in the cell (methodology=regression, domain=composite, parameter=flow), while Smith and Babiceana (2004) is located in the cell (methodology=EM/DA, domain=composite, parameter=composite). In addition, the framework can also help identify new imputation techniques by entering proper cells in the framework.

The above imputation methods impute only one estimate for each missing value, and hence, these techniques can be called single imputation methods, as illustrated in Fig. 1. Unlike single imputation, the multiple imputation (MI) method (Rubin 1987, 1996) replaces each missing value with a set of plausible values to represent the uncertainty about the right value to impute. The multiple imputation technique can work on top of various imputation methods, as listed in Fig. 1. Examples of the underlying statistically principled methods include: the regression method (Rubin 1987), the propensity score method (Rosenbaum and Rubin 1983; Rubin 1987; Lavori et al. 1995), the expectation maximization method (Dempster et al. 1977; Schafer 1997), the data augmentation (DA) method (Tanner and Wong 1987), the Markov chain Monte Carlo (MCMC) method (Gilks et al. 1996; Schafer 1997), etc. To show the application of multiple imputation in ITS data, this research employs EM/DA as the underlying imputation method and the results in this paper are based on imputing traffic counts, so the position of this research is located in the cell (methodology=EM/DA, MI, domain=composite, and parameter=flow).

Multiple Imputation Scheme

This section develops a multiple imputation procedure and discusses the role of data aggregation in the imputation.

Multiple Imputation Procedure

The procedure of multiple imputation is outlined in the following three steps.

Filling the Missing Data n Times

The first step of multiple imputation is to estimate multiple values for each missing datum. This simulates multiple random draws from a population in order to estimate the unknown parameter. There is no general guideline regarding which statistically principled technique to choose, but empirical studies show that for monotone missing data patterns a regression or a propensity scores method is more appropriate, while an EM/DA or MCMC method works better for an arbitrary missing data pattern. Take EM/DA for example, EM is used first to generate maximum-likelihood estimates of the missing values which is then used as input for DA. Next, DA is run for k iterations, where k is set large enough to guarantee convergence. This produces a random draw of parameters from their posterior distribution. Imputing the missing data under these random parameter values results in one imputation. Repeating the whole process n times produces n sets of imputed data.

Analyzing the n Imputed Data Sets

For the n sets of imputed data, our main interest is to quantify the variability of the multiply imputed data as well as the uncertainty introduced by missing data. Let

$$\bar{v} = \sum_{i=1}^n \hat{v}_i \quad (1)$$

where $\hat{v}_i$ =mean of the imputes of the i th imputation, $i=1, 2, \dots, n$; and $\bar{v}$ =grand mean of all imputes. Total variance of all imputes can be decomposed into two terms:

$$\text{Var}^T = \text{Var}^W + \left(1 + \frac{1}{n}\right) \text{Var}^B \quad (2)$$

where Var^T =total variance. Var^W is the within-imputation variance which preserves the nature variability. This component is equivalent to the variance that would exhibit if there were no missing data and is computed by simply averaging the variances of each imputation

$$\text{Var}^W = \frac{1}{n} \sum_{i=1}^n \text{Var}_i^W$$

Var^B is the between-imputation variance which explains the uncertainty introduced by missing data. This variance measures how the estimated values vary from imputation to imputation and is computed as

$$\text{Var}^B = \left(\frac{1}{n-1}\right) \sum_{i=1}^n (\hat{v}_i - \bar{v})^2$$

If the estimated values vary greatly from imputation to imputation, this means the uncertainty introduced by missing data is high and Var^B should be large. Otherwise, Var^B should be small.

Combining the n Results for Inference

Combining the n sets of imputed data is quite simple and the most common practice is simply to take the average of the n sets.

Imputation before Aggregation versus Aggregation before Imputation

Raw ITS data are often collected with short intervals such as 20 s and 1 min. However, traffic engineering studies typically necessitate longer intervals such as 5 and 15 min. A basic question is whether we should impute before aggregate or aggregate before impute. Smith et al. (2003) aggregate 1–10 min data and then perform imputation, while Chandra and Al-Deek (2004) aggregate 30–5 min data 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 over where the missing values should appear and is, therefore, unable to clearly delineate “good” and “bad” data. Aggregation before imputation might accidentally incorporate missing values and/or preimputes into the aggregated data based on which the intended imputation is going to be performed. This means that one is working on modified data rather than raw data and this aggregation may alter the natural relation embedded in the raw data. On the other hand, this approach may result in loss of usable information and/or introducing extra error in the aggregated data. With these issues in mind, this paper follows the imputation-before-aggregation approach, i.e., imputation is first performed directly on the raw data (in 20 s bins) and the imputed data are aggregated next. By this way, one does not need to worry about delineation of good and bad data. Though the raw data exhibit higher variability, much of that variability is attributable to white noise. The raw data contain more information regarding how the system works as well as the relations among variables of interest. The basic idea of imputation is to learn from available information and estimate what is missing. Therefore, raw data will be more helpful in restoring the original information. Also, aggregation after imputation is able to give a more reliable and clean trend

Fig. 2. Study site: GA 400

because all the missing values have been filled with educated estimates and this represents the best estimation one has on the real system.

Study Site and Data

The procedure outlined in the previous section was validated with real ITS data from GA 400. This section presents an overview of the study site and test data.

Study Site

Test data used in this study came from GA 400, which is a toll road to the north of the Atlanta metropolitan area. Traffic on this road, as shaded in green in Fig. 2, is monitored by Georgia NaviGator—the ITS system of Georgia. The surveillance system covers the section between I-285 (to the south) and Old Milton Parkway (to the north), a stretch of about 20.2 km (12.56 mi), and this is our study site.

Data Set

Traffic conditions on the study site are monitored by video cameras, which are deployed approximately every 0.54 km (one third mile) of the road for each direction. Each camera constitutes an observation station (or simply a station) and watches all the lanes at this location. An image processing software runs in the background to extract traffic data from the videos. Simulated loops are placed over the lanes to detect vehicles and these loops are called detectors with each detector corresponding to a single

Table 1. Summary of the Incomplete Data at 30% Missing Rate

(a) Number Observations=798; number of variables=4			
	Number Missing	% Missing	
Detector_1	227	28.45	
Detector_2	254	31.83	
Detector_3	244	30.58	
Detector_4	233	29.20	
(b) Matrix of missingness patterns ^a			
Count	Pattern	Count	Pattern
194	1111	79	1110
80	0111	34	0110
90	1011	35	1010
25	0011	17	0010
82	1101	30	1100
31	0101	14	0100
47	1001	14	1000
16	0001	10	0000
(c) Means and standard deviations of observed data ^b			
	Mean	Standard deviation	
Detector_1	9.87215	3.84198	
Detector_2	10.0717	2.84667	
Detector_3	9.93141	2.84799	
Detector_4	8.77699	3.08620	

^a1=observed; 0=missing; and count=number of observations with the specified pattern.

^bUnit: vehicle count.

Table 2. Summary of Statistical Analysis of the Imputation Errors (958 Samples in Each Imputation)

	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5
Mean	0.09	0.10	0.17	0.08	0.27
Variance	16.28	14.09	14.37	14.51	14.10
Standard deviation	4.03	3.75	3.79	3.81	3.75
T-test statistics	0.73	0.80	1.38	0.65	2.25

lane. Traffic conditions are sampled every 20 s and columns provided in a sample include detector ID, sample start time, classified volumes, time occupancy, time mean speed, level of service, density, etc.

Multiple Imputation and Results

This section details how the multiple imputation scheme is applied and presents validation results at various perspectives.

Validation Procedure of Multiple Imputation

The validation of the multiple imputation scheme is conducted as follows. First, we obtain field data and randomly choose a subset for validation. The chosen datasets must be good (i.e., without missing values) through a sufficiently long period (e.g., 20 h). For each data set, i.e., the “complete data,” we randomly eliminate some values to simulate data missing and we called the resulted data set “incomplete data.” We impute the incomplete data multiple times and obtain multiple versions of “imputed data.” Next, we combine the multiple versions of imputed data and obtain the “combined data.” Then, we take out all “combined imputes” for comparison with their “actual values” by means of statistical tests [regular statistical tests if no autocorrelation is involved or Ni et al. (2004) otherwise].

Results of Multiple Imputation

Multiple imputation is validated using Ga 400 data set. To give in-depth analysis of the validation results, the following discussion focuses on data of the day October 1, 2003 at Station 4000044 (see Fig. 2).

Summary of the Incomplete Data

The missing mechanism is simulated by generating an array of nonrepeating random numbers using a random number generator based on a prespecified missing rate. These random numbers are then used as keys to enter the complete data to determine which values to eliminate. The complete data contains 798 observations (cases or records) and each observation consists of four variables (lanes or detectors). Table 1 summarizes the resulting incomplete data at a 30% missing rate. The number of missing values under this rate is $794 \times 4 \times 30\% \approx 958$.

Analyzing Imputed Data

To perform multiple imputation, this study identifies existing software programs and chooses one to serve our need. A few software programs are identified such as SOLAS 3.0 (Statistical Solutions 2004) and NORM (Schafer 1997, 1999). NORM (currently version 2.03) is selected in this study because it is sound in principle, simple to use, and readily accessible. Once an input data file has

Fig. 3. Diagonal plot of imputed versus actual values (with 30% missing)

Fig. 4. Histogram of imputation error (with 30% missing)

been loaded, NORM displays a summary of the observed data. This summary includes the number of and per cent missing of each variable, as well as the means and standard deviations of the observed data. After examining the data summary, an expectation-maximization procedure is run. This algorithm is a preliminary step that estimates imputes for the missing values. Following the EM procedure is a data augmentation procedure which is an iterative process that fine tunes the imputes generated in the EM step. The final step in the analysis is to combine the imputes and report the result. The report provides the overall estimate and the associated standard errors, degrees of freedom, p values, and confidence interval.

Table 3. Summary of Imputation Quality Under Different Missing Rates (Station 4000044)

Missing rate	Mean of errors	SD of errors	MAPE of imputes	Overall MAPE
0.00	0.00	0.00	0.00	0.00
0.10	0.01	2.93	0.30	0.03
0.20	-0.01	2.65	0.25	0.06
0.30	-0.03	2.98	0.30	0.09
0.40	0.01	3.03	0.32	0.13
0.50	-0.01	3.08	0.33	0.16

Note: SD=standard deviation and MAPE=mean absolute percentage error.

In this study, we perform imputation five times, resulting in five versions/columns of imputes. The five columns of imputes are contrasted by their corresponding actual values which are placed in the sixth column. Since the missing data are simulated by random elimination, the six columns can be viewed as six random processes (as opposed to a time series). On the other hand, each version of imputes are paired up with their corresponding actual values so that the imputation errors can be computed. Statistical analysis is then performed on the imputation errors and the results are summarized in Table 2.

Table 2 shows that the imputations are quite stable because the variation of means and variances is small from imputation to imputation. A two-tailed t -test is performed after checking the necessary conditions and the Null hypothesis here is: H_0 : The paired imputation errors are not significantly different than zero. With level of significance $\alpha=0.05$ and critical value of 1.96, it can be seen that four of the imputations strongly support the null hypothesis while the last one fails to.

Fig. 5. Comparison of complete data and imputed data by detectors (with 30% missing)

Fig. 6. Plot of imputation robustness

Combining Imputed Data

For usability in practice, the multiple versions of imputed data need to be combined into an overall one. A simple way to do this is to average the multiple imputations. The statistical analysis mentioned above is, again, applied to the residuals obtained by pairing the combined imputes and their corresponding actual values. The result of the t-test in this case also supports the null hypothesis, i.e., imputation errors are not significantly different than 0.

In addition to the basic statistical analysis, variance of imputation is analyzed further. In this example, we have $\text{Var}^W = 10.45604$ and $\text{Var}^B = 0.00097$. This implies that although the within-imputation variance (nature variability) is high, the between-imputation variance is very low, i.e., the uncertainty caused by missing data is low since there is much information to restore the actual values. The total variance is $\text{Var}^T = 10.4572$ and this translates to 3.23376 for standard error of the mean.

Comparison of Combined Imputes and Their Actual Values

To see the imputation quality, the following figures present the results of comparing the combined imputes and their actual values. A 30% missing rate is implied here unless explicitly mentioned otherwise.

Fig. 3 contracts combined imputes against their actual values. An ideal imputation would be a 45° line, as shown in Fig. 3. Though the plot shows some deviation around the line, data points are generally evenly scattered at both sides of the line. The trend of the data points in Fig. 3 also suggests that lower values are likely to be overestimated while higher values tend to be underestimated. However, in practical use, such a bias tends to be canceled out when aggregating the data to longer time intervals, as can be seen in Fig. 5. In case that the aggregation fails to cancel the bias, an adjustment procedure might be necessary.

Fig. 4 presents the frequency of imputation error. The histogram roughly exhibits a bell shape, indicating that the imputation error is approximately normally distributed. This is a necessary condition for performing the t-tests in previous sections.

The above discussions focus on comparing imputes and their actual values. Now let us examine the entire data set. Fig. 5 gives a detector-by-detector comparison of the complete data (solid lines) and the imputed data (dotted lines) in time series. These plots show that the dotted lines chase and fit the red lines quite well.

To show the robustness of the multiple imputation scheme, tests were repeated at every 10% increment in missing rate. For each level of missing rate, multiple imputation is replicated five times and statistics are collected for each replication. These statistics include mean of errors, standard deviation of errors, mean absolute percentage error (MAPE), and overall MAPE. The first three statistics are based on combined imputes while the last one is based on the entire data set. The five replicates are then averaged to give a set of overall statistics which are presented in Table 3. It can be seen that the MAPE of imputes is generally around 30%, but the MAPE of the entire data set is very small as can be seen from the column overall MAPE. This means that the multiple imputation scheme is quite robust under different missing rates. In Table 3, results for missing rates higher than 50% are not listed because these scenarios are generally regarded

Fig. 7. Time series plot of aggregated data (5 min per step, sum over all four lanes)

as impractical and the usability of these data sets are greatly questionable. However, it still makes sense to examine the trend as missing rate varies from 0 to 1. Fig. 6 shows how overall MAPE varies as missing rate increases. It can be seen that the overall MAPE increases steadily almost linearly up to somewhere around 90% missing rate and then increases exponentially when approaching 100% missing rate. It is interesting to notice that a 30% missing rate corresponds to an overall MAPE of about 10% which is generally acceptable in practice.

To verify the effect of imputation before aggregation as well as to provide a basis to compare with the aggregation-before-imputation approach, the 20 s data are merged to be 5 min data for both the combined data and the complete data. Without loss of generality, this process is still based on 30% missing rate. Fig. 7 shows a time series plot of the aggregate data sets with the solid line as complete data and the dash-dotted line as combined data. The two lines fit each other very well. Quantitatively, the overall MAPE of these two curves is 0.0216. This result is comparable with that of Smith et al. (2003), but the former eliminates the possibility of introducing extra error into imputation, provides extra information about the uncertainty of making the imputation, and the preservation of the natural variability of the observed data.

Conclusion

One of the goals of HCM is to fully replicate field conditions. To achieve this goal, efforts of HCM model development, validation, and refinement have to work closely with field data. However, a major problem with field data is the issue of data missing which sometimes can render the field data useless or lead to erroneous results. Imputation for missing data is a feasible and low-cost solution to this issue.

This paper summarizes the current practice of imputing missing values in ITS data and develops a framework of imputation where existing imputation methods can be related to each other and new imputation methods can be identified by entering proper cells in the framework. A multiple imputation scheme is outlined where a missing value is imputed multiple times and this simulates a random sampling process to estimate the unknown parameter. In addition to the high imputation quality, the multiple imputation scheme merits many advantages such as yielding unbiased estimates for the missing values, preserving the natural variability of the observed data, and providing a measure of the uncertainty introduced by missing data.

The results obtained from this study are based on the premise that data points are missing at random under different missing rates. However, real world traffic surveillance systems sometimes fail to record data for an extended period of time. To deal with such data, formal investigation is strongly recommended before applying this imputation scheme.

It is suggested that the AASHTO's guideline be reconsidered for the following reasons: one, imputation has proved to be able to achieve reasonable accuracy as demonstrated in this and previous studies; two, imputation is able to preserve the original relationship among variables as well as their natural variability; three, the uncertainty introduced by missing data can be quantified and this enables users to make educated decisions on either incorporating imputes in their analysis or not.

References

- AASHTO. (1992).
- Chandra, C., and Al-Deek, H. (2004). "New algorithms for filtering and imputation of real time and archived dual-loop detector data in the I-4 data warehouse." *Proc., 83rd Transportation Research Board (TRB) Annual Meeting*, TRB, National Research Council, Washington D.C., Preprint CD-ROM.
- Chen, C., Kwon, J., Rice, J., Skabardonis, A., and Varaiya, P. O. (2003). "Detecting errors and imputing missing data for single-loop surveillance systems." *Proc., 82nd Transportation Research Board (TRB) Annual Meeting*, TRB, National Research Council, Washington D.C., Preprint CD-ROM.
- Conklin, J. H. and Smith, B. L. (2002). "The use of local lane distribution patterns for the estimation of missing data in transportation management systems." *Transportation Research Record 1811*, Transportation Research Board, Washington, D.C., 50–56.
- Dailey, D. J. (1993). "Improved error detection for inductive loop sensors." *Rep. No. WA-RD 3001*, Washington State Department of Transportation, Olympia, Wash.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum-likelihood estimation from incomplete data via the EM algorithm (with discussion)." *J. R. Stat. Soc. Ser. B. Methodol.*, 39, 1–38.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., eds. (1996). *Markov chain Monte Carlo in Practice*, Chapman & Hall, London.
- Lavori, P. W., Dawson, R., and Shera, D. (1995). "A multiple imputation strategy for clinical trials with truncation of patient data." *Stat. Med.*, 14, 1913–1925.
- Ni, D., Leonard, J. D., Guin, A., and Williams, B. M. (2004). "A systematic approach for validating traffic simulation models." *Proc., 83rd Transportation Research Board (TRB) Annual Meeting*, TRB, National Research Council, Washington D.C., Preprint CD-ROM.
- Nihan, N. (1997). "Aid to determining freeway metering rates and detecting loop errors." *J. Transp. Eng.*, 123(6), 454–458.
- Rosenbaum, P. R., and Rubin, D. B. (1983). "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70, 41–55.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, Wiley, New York.
- Rubin, D. B. (1996). "Multiple imputation after 18+ years." *J. Am. Stat. Assoc.*, 91, 473–489.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*, Chapman & Hall, New York.
- Schafer, J. L. (1999). "NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT." (<http://www.stat.psu.edu/~jls/misofwa.html>), accessed February 4, 2004.
- Smith, B., Scherer, W., and Conklin, J. (2003). "Exploring imputation techniques for missing data in transportation management systems." *Proc., 82nd Transportation Research Board (TRB) Annual Meeting*, TRB, National Research Council, Washington D.C., Preprint CD-ROM.
- Smith, B., and Babiceanu, S. (2004). "An investigation of extraction transformation and loading (ETL) techniques for traffic data warehouses." *Proc., 83rd Transportation Research Board (TRB) Annual Meeting*, TRB, National Research Council, Washington D.C., Preprint CD-ROM.
- Statistical Solutions. (2004). "SOLAS for Missing Data Analysis and Multiple Imputation." (<http://www.statsol.ie/solas/solas.htm>), accessed October 16, 2004.
- Tanner, M. A. and Wong, W. H. (1987). "The calculation of posterior distributions by data augmentation (with discussion)." *J. Am. Stat. Assoc.*, 82, 528–550.
- Zhong, M., Sharma, S., and Lingras, P. (2004). "Genetically designed models for accurate imputations of missing traffic counts." *Proc., 83rd Transportation Research Board (TRB) Annual Meeting*, TRB, National Research Council, Washington D.C., Preprint CD-ROM.