MANAGING MISSING PAVEMENT PERFORMANCE DATA IN PAVEMENT MANAGEMENT SYSTEM

by

J. Farhan[1] and T. F. Fwa[2]

[1]Research Fellow and [2]Professor

Department of Civil and Environmental Engineering
National University of Singapore
10 Kent Ridge Crescent
SINGAPORE 119260

Total Number of Words

Number of words in text:			=	4086	words
Number of tables:	4	(4 x 250)	=	1000	words equivalent
Number of figures:	2	(2 x 250)	=	500	words equivalent
Total number of words			=	5586	words equivalent

Corresponding author: Professor Fwa Tien Fang
Department of Civil and Environmental Engineering
National University of Singapore
10 Kent Ridge Crescent
SINGAPORE 119260

ABSTRACT
Missing data in pavement condition and performance records of pavement management systems (PMS) are ubiquitous in practice. Imputation of missing data is often required in the analysis of pavement performance and decision making for pavement management. The traditional methods of handling missing data by pavement engineering professionals include deletion of affected records, and imputation of missing data either by means of interpolation substitution, mean substitution, or regression substitution. Today, the advancement of computer technology has permitted the use of computationally complex stochastic Multiple Imputation algorithms to improve the accuracy of missing data estimates. This paper examines the effects of different available missing-data imputation techniques in handling missing pavement performance data in pavement management systems. The methods of Multiple Imputation are also examined to take into account the stochastic nature of the data imputation problem. Demonstrative examples using actual records from LTPP database are presented to illustrate the relative merits of different missing data imputation techniques.

INTRODUCTION
The basic requirement of a pavement management system is to have an efficient pavement condition and performance data collection program to its support decision making process. In order to ensure the data collected meet the needs of the pavement management decision making process, highway agencies have developed data quality management programs (*1*, *2*, *3*) entailing procedures and guidelines for managing the quality of pavement data collection activities in terms of quality control and assurance. Data quality assurance is the process of profiling the data to discover inconsistencies, and other anomalies in the data and performing data cleansing activities such as removal of outliers and imputation of missing data. As transportation agencies heavily rely on data driven applications, the need for complete and accurate data is increasing.

Missing data in databases has been one of the most prevalent problems in pavement management systems (*4*). According to NCHRP (*5*), 61 percent of the highway agencies reported employing software routines to check for missing data elements, and some agencies reported mitigating missing data issues through recollection (*6*). For example, the Missouri Department of Transportation (MoDOT) five-year condition data reported that only the 1999 PSR data were fully complete, and the 1998 PSR data, being the second most complete dataset, had only records of 54 percent of the data (*4*). A quality assurance computer program developed by the Colorado Department of Transportation (CDOT) found it necessary to check for duplicate records, missing segments, incorrect highway limits, missing or incorrect highways, incorrect pavement type, and incorrect raw distress values (*2*). Zhang and Smadi (*7*) listed various data quality checks in Iowa DoT, including missing data.

While the principles of statistical quality assurance in terms of imputation of missing data are well developed, their application and performance to the imputation of pavement management data is unclear. Most data archival facilities in pavement management systems do not have a reliable system for dealing with missing data. This paper proposes a multiple imputation approach to address the missing pavement condition data issue, and analyses feasibility and applicability of the approach in comparison to four other existing imputation techniques. The first technique examined in this study is listwise deletion, and the second is mean substitution, while the last two techniques are linear interpolation and regression substitution. The following section presents a brief overview of the existing imputation procedures which have been implemented in various other fields.

EXISTING DATA IMPUTATION PRACTICES

Traditionally, several approaches have been employed for the purpose of estimating missing data. A brief overview of the four basic approaches is presented in this section.

Listwise Deletion

This is by far the most common approach involving neglecting cases with missing data and to run analyses on remaining data. This leads to a loss of reliability as the available sample size for potential analyses is reduced, although it produces unbiased parameter estimates in the case where the data is missing at random. Several works (*8*, *9*, *10*) have demonstrated the implications of simply removing cases using the listwise deletion method (LD) on the original data set.

Ad-hoc Imputation using Basic Statistics

Basic statistics can be applied to fill in the gaps of missing data to maintain sample size for analyses. However, many of such ad hoc imputation methods come with implausible assumptions such as ignoring the stochastic nature of the unobserved values. These methods impute the missing data only once using a single calculated imputation value. The following are some basic statistics that have been employed for this purpose.

Mean Substitution

In this approach the missing physical values are imputed using the mean value of a data set of a particular pavement distress over time. However, it adds no new information since the overall mean, with or without replacing missing data, will remain constant, and the variance will be artificially decreased proportionally to the number of missing data. In addition, since certain distresses evolve over time or are significantly correlated with other distresses, substituting those by mean values will result in considerable loss in correlation.

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad (1)$$

Interpolation using Adjacent Data Points

The missing data are computed by interpolation from the adjacent available data points, which graphically amounts to substituting missing data by connecting with a straight line the point just prior to the missing data with the point just following the missing data. This method assumes a linear correlation in the data, that is, that each observation is to some extent related to and therefore most similar to the previous observation. Yang et al. (*11*) applied this approach in forecasting pavement condition rating in Texas. Bennett (*10*) suggested this as one of the possible approaches to imputing pavement condition data, and is represented by the following equation in case of three data points (x_1,y_1), (x_2,y_2) and (x_3,y_3),

$$y_2 = \frac{(x_2 - x_1)(y_3 - y_1)}{(x_3 - x_1)} + y_1 \qquad (2)$$

Regression Substitution

This approach involves fitting a least-squares regression line to the data on the basis of available information such as pavement age, traffic volume and load. The missing data are then be replaced by the values predicted by this regression line. This model assumes a linear relationship between the dependent variable y_i and the p-vector of regressors x_i, and is modeled with the so called noise term ε_i that adds noise to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = x_i'\beta + \varepsilon_i \qquad i = \{1, 2, \ldots, n\} \qquad (3)$$

where ′ denotes the transpose, so that $x_i'\beta$ is the inner product between vectors x_i and β. Often these n equations are stacked together and written in vector form as,

$$y = X\beta + \varepsilon \quad (4)$$

Expectation Maximization Algorithm

The Expectation Maximization Algorithm (EM) is an iterative regression technique in which the missing variables are regressed on the available data and any additional variables provided as inputs to the algorithm. First, a vector of means and a covariance matrix are calculated using all available data. The means are then imputed for missing values in each variable which serve as a starting value for the imputation. Next, variables with missing values are regressed on all the other available variables. The imputed mean values are then replaced with estimates calculated from the regression equations, and the means and covariances are recalculated. Regression equations and imputations are iteratively calculated, and the process continues until the mean and covariance matrix values converge (*8*, *9*).

A review of the literature indicates that the effectiveness of the data imputation methods relies strongly on the problem domain such as the number of cases, number of variables, and patterns of missing data (*12*, *13*). There is no clear indication of dominance of one imputation method over another, in general, hence this paper contributes to the literature by comparing existing data imputation approaches in the context of pavement management, and proposes a multiple imputation approach in resolving missing data issue.

CONCEPT OF MULTIPLE IMPUTATION

This paper proposes a multiple imputation approach specific to resolving the data imputation issues in the context of pavement management system and compares the effectiveness of the proposed approach against the existing approaches. Multiple Imputation is a technique in which the missing values are replaced by $m > 1$ plausible values drawn from their predictive distribution. The variation among the m imputations reflects the uncertainty with which the missing values can be predicted from the observed ones. As a result, there are m complete data sets. Rubin (*14*) identified that an important limitation of single imputation methods is that "standard variance formulas applied to the filled-in data systematically underestimated the variance of estimates"

In Rubin's method for multiple imputed inference (*14*), each of the simulated complete data sets is analyzed by standard statistical methods, and the results (estimates and standard errors) are combined to produce estimates and confidence intervals incorporating missing data uncertainty. The technique is performed using Data Augmentation (DA) Algorithm (*15*), however Expected Maximization Algorithm is considered a preferred approach in establishing initial estimates such as mean and covariance for DA to begin with (*12*).

Expected Maximization Algorithm (EM)

Dempster et al. (*16*) published a paper titled Maximum Likelihood from Incomplete Data via the "EM" Algorithm. In this paper they presented an iterative regression technique for calculating descriptive statistics on a data set with missing values in such a way that statistical inferences could still be made from the data. The EM algorithm is a general method for obtaining maximum likelihood estimates of parameters in problems with incomplete data. Consider an incomplete data matrix with the observed data defined as Y_{obs}, missing data as Y_{mis} , and a vector of parameters as θ. Hence, complete data, Y_{com}, can be defined as $Y_{com} = (Y_{obs}, Y_{mis})$. With the complete data log-likelihood function, $L(\theta) = f(Y_{com}|\theta)$ and the observed

data log-likelihood function, $L(\theta) = f(Y_{obs}|\theta)$, the expected complete data log-likelihood function can be defined as,

$$Q(\theta|\theta') = E\{ln[f(Y_{com}|\theta)]Y_{obs}, \theta'\} \quad (5)$$

The EM algorithm begins with some value of θ and alternates between two steps (*17*) as follows:

(i) Expectation step (E-step), i.e. Computing $Q(\theta|\theta^{(t)})$ as a function of θ, and

(ii) Maximization step (M-step), i.e. Find $\theta^{(t+1)}$ that maximizes $Q(\theta|\theta^{(t)})$

The increase in the log-likelihood function $L(\theta)$ is observed with each iteration of the EM algorithm until convergence (*16*), and the rate of convergence is proportional to the amount of unobserved or missing information in a data matrix (*18*).

Data Augmentation Algorithm (DA)

The Data Augmentation Algorithm requires starting values for the mean and covariance matrix, and an appropriate approach is to calculate these values using the EM algorithm. Data Augmentation makes use of Multiple Imputation, and the premise behind generating multiple imputations is that instead of using a point estimate as the imputed value, several estimates can be combined to calculate the imputed value. By using multiple points, the analyst is using a distribution of data to find the imputation, and this not only can result in better estimates, but it provides insight in to how much variance there is in the estimate.

Data augmentation (DA) process is similar in nature to that of EM algorithm i.e. an iterative process which alternately fills in the missing data while crafting inferences about the unknown parameters, however in contrast to the EM algorithm; this is performed in a stochastic manner (*19*). A random imputation of missing data under assumed values of the parameters is performed by DA, followed by estimating of new parameters from a Bayesian posterior distribution based on the observed and imputed data (*20*). Beginning at some value of θ, each iteration of the DA algorithm alternates between two steps (*20*) as follows:

(i) Imputation step (I-step): Draws $Y_{mis}^{t+1} \sim P\left(Y_{mis} \mid Y_{obs}, \theta^{(t)}\right)$, and

(ii) (ii) Posterior step (P-step): Draws $\theta^{(t+1)} \sim P\left(\theta \mid Y_{obs}, \theta^{(t+1)}\right)$

This process of alternately imputing and establishing missing data and parameters respectively creates a Markov chain that finally converges in distribution (*20*).

PROPOSED PROCEDURE OF MULTIPLE IMPUTATION

The procedure of the Multiple Imputation method adopted in the study involves the following steps:

Step I: Data Transformation

Data for all variables are transformed to approximately normal before imputation using a logit, log or square root transformation function and then transformed back to their original scale after imputation. The logit or logistic transformation (*21*) is defined as:

$$\log it(p) = \log\left(\frac{p}{1-p}\right) \quad (6)$$

where p stands for probability or proportion.

Step II: Imputation using EM

Estimates of missing values are generated for the data matrix using the EM algorithm with the convergence criterion, which is the maximum relative parameter change in the value of any parameter during iterative process, is set as 0.0001.

Step III: Imputation using DA

The initial estimates from the EM algorithm serve as the basis for the DA algorithm to generate multiple imputed data matrices as explained in the preceding section.

Step IV: Synthesis of Estimates

Since the data augmentation algorithm is stochastic in nature, there will be a slight variation between the imputed data matrices. Consequently, when any standard data analysis procedure is applied to each set of data, the results will differ slightly from one analysis to another. The final set of estimates is derived by averaging over these estimates following a set of rules provided by Rubin (*14*). The standard data analysis can be considered a linear regression analysis with regression equation consisting of one independent variable *Z*, and can be represented as follows,

$$\hat{X}_j = b_0 + b_1 Z_j \quad (7)$$

where $\hat{X}_j$ = estimate from data set *j*, and b_0, b_1 represent regression coefficients.

Rubin (*14*) presented this step for combining results from a data analysis performed *m* times, once for each of *m* imputed data sets, to obtain a single set of results as follows. From each analysis, one must first calculate and save the estimates and standard errors. Suppose that $\hat{X}_j$ = estimate from data set *j*, and Y_j = squared standard error for $\hat{X}_j$

The overall estimate is the average of the individual estimates,

$$\overline{X} = \frac{1}{m}\sum_{j=1}^{m} \hat{X}_j \quad (8)$$

For the overall standard error, obtain the within-imputation variance,

$$\overline{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j \quad (9)$$

and the between-imputation variance,

$$B = \frac{1}{m-1}\sum_{j=1}^{m} \left(\hat{X}_j - \overline{X}\right)^2 \quad (10)$$

The total variance is

$$T = \overline{Y} + \left(1 + \frac{1}{m}\right)B \quad (11)$$

The overall standard error is the square root of *T*. Confidence intervals are obtained by taking the overall estimate plus or minus a number of standard errors, where that number is a quantile of Student's t-distribution with degrees of freedom,

$$df = (m-1)\left(1 + \frac{m\overline{Y}}{(m+1)B}\right)^2 \quad (12)$$

A significance test of the null hypothesis $X = 0$ is performed by comparing the ratio to the same t-distribution as follows,

$$t = \overline{X} \Big/ \sqrt{T} \quad (13)$$

The entire approach is explained on two typical pavement performance databases in the following section.

ILLUSTRATIVE EXAMPLE

Pavement Performance Database

A typical highway pavement condition survey database was used in this study to assess the performance of each imputation approach in estimating missing distress data as shown in Table 1. The databases include seven time-series, transverse crack density in crack per mile (lane 1 and 2), average IRI in inch per mile and rut in inch (left and right wheelpath), and average longitudinal cracking in feet.

TABLE 1 Typical Pavement Performance Database of a Section

Year	Transverse Crack Density (Crack/mi)		Average IRI (in/mi)	Rutting (in)		Avg. Longitudinal Cracking (ft)
	Lane 1 (outside)	Lane 2 (inside)		R-RUT	L-RUT	
1980	2.75	2.13	48.18	0.17	0.15	1.26
1981	4.01	2.15	48.74	0.17	0.16	1.62
1982	4.75	2.22	49.18	0.18	0.20	2.91
1983	4.84	2.27	49.37	0.19	0.20	2.95
1984	6.2	2.35	49.45	0.20	0.20	3.20
1985	6.21	2.37	49.65	0.20	0.20	3.41
1986	6.58	2.39	49.78	0.21	0.21	3.54
1987	7.09	2.42	50.23	0.21	0.21	3.65
1988	7.94	2.49	50.44	0.21	0.21	4.85
1989	7.98	2.53	51.01	0.21	0.21	5.07
1990	8.58	2.55	51.05	0.22	0.22	5.14
1991	8.81	2.55	51.1	0.22	0.22	5.48
1992	8.86	2.56	51.13	0.22	0.22	6.00
1993	9.51	2.57	51.93	0.23	0.22	6.01
1994	9.62	2.58	52.75	0.23	0.23	6.03
1995	9.96	2.58	52.92	0.23	0.23	6.30
1996	9.97	2.59	53.08	0.24	0.24	6.60
1997	11.45	2.6	53.23	0.24	0.24	6.63
1998	11.68	2.62	53.44	0.24	0.24	6.79
1999	11.82	2.64	53.45	0.24	0.24	6.97
2000	12.06	2.64	53.75	0.24	0.24	7.02
2001	12.08	2.68	54.31	0.25	0.24	7.36
2002	12.53	2.7	54.52	0.25	0.25	7.66
2003	12.75	2.71	54.63	0.26	0.26	8.67
2004	14.32	2.77	54.71	0.26	0.26	9.08
2005	14.71	2.81	54.79	0.29	0.26	9.3
2006	14.96	2.88	55.53	0.30	0.26	9.52
2007	16.03	2.88	55.61	0.31	0.26	9.59
2008	16.69	2.92	56.16	0.31	0.26	9.66
2009	17.76	3.05	57.46	0.38	0.28	10.16

Evaluation Concept

In order to test all of the techniques, delineated in the preceding sections, data points are removed randomly from the pavement performance data set. The premise of removing data

records as is to construct a missing data pattern in order to evaluate each of the imputation approach explained earlier. This may not be a typical pattern for missing data in pavement performance database; however it was used in order to facilitate the comparison of various imputation approaches discussed. It must be noted that removing data in this manner does not violate the missing completely at random (MCAR) assumption, as long as the value of the data is independent of the missing data mechanism. Once the missing data is imputed, real values are compared against imputed values to assess performance of each missing data estimation method. The assessment of the imputation capability of the proposed approach is measured using the mean absolute error (MAE), which is a quantity used to measure how close forecasts or predictions are to the actual outcomes. The mean absolute error is an average of the absolute errors, and is given by the following equation,

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (14)$$

where, f_i is the imputed value, y_i is the actual value, and n represents the number of observations.
This paper employs this technique to validate the proposed approach, and the imputed results from the proposed approach are compared against the results from the traditional approaches described earlier.

Multiple Imputation using Maximum Likelihood

In order to assess the performance of the proposed approach in estimating missing pavement performance data, density histograms, and probability distribution plots were evaluated to determine if the data were normally distributed. An appropriate transformation function was used, such as logit, log or square root transformation, to approximate the data to normal distribution. This was found to give the best approximation of variables to normality. Anderson-Darling (AD) test (*22*) was employed to evaluate conformation of transformed data to normal distribution with a 95 percent confidence interval, and the outcome is represented in terms of AD statistics and P-value as follows.

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\ln\left(F\left(Y_i\right)\right) + \ln\left(1 - F\left(Y_{n+1-i}\right)\right)\right) \quad (14)$$

$$AD^* = AD\left(1 + \frac{4}{n} - \frac{25}{n^2}\right) \quad (13)$$

The hypotheses of the AD* test are as follows,
Null hypothesis (H_0): Data is sampled from a population that is normally distributed, and Alternative hypothesis (H_1): Data is sampled from a population that is not normally distributed. Hence, if AD* statistic exceeds a given critical value, which in this case is 0.787, at $\alpha = 0.05$ or P-value is less than α, we reject null hypothesis.

Given the normally distributed data, the Expected Maximization algorithm was employed with the convergence criteria set to 0.0001. The EM algorithm converged after 15 iterations for all the given pavement performance parameters. The EM estimates serve as starting values for the Data Augmentation (DA) process. Since the convergence behavior of DA is the same as EM algorithm, the number of iterations needed for the convergence of the DA algorithm is more or less the same as EM's for the pavement performance data matrix. The data imputation algorithm was configured for 5 imputations after every 50 iterations for a total of 250 iterations while ensuring enough cycles between each imputed data matrix to ensure convergence.

The performance of the proposed approach in estimating missing pavement condition data is then compared against existing methods used in practice. The following four methods are considered, the descriptions for all of which have been presented earlier.

- Method I: Ad hoc imputation method by mean substitution

- Method II: Ad hoc imputation method by interpolation using adjacent points
- Method III: Ad hoc imputation method by linear regression substitution
- Method IV: Multiple Imputation proposed in this study

Assessment of Results

The relative quality of the results from the four methods are assessed using mean absolute error in percentage, as tabulated in Table 2 and plotted in Fig. 1. As can be seen from the figure, the mean substitution method resulted in the highest amount of deviations of the imputed values from the actual values, followed by the regression substitution method and the interpolation method.

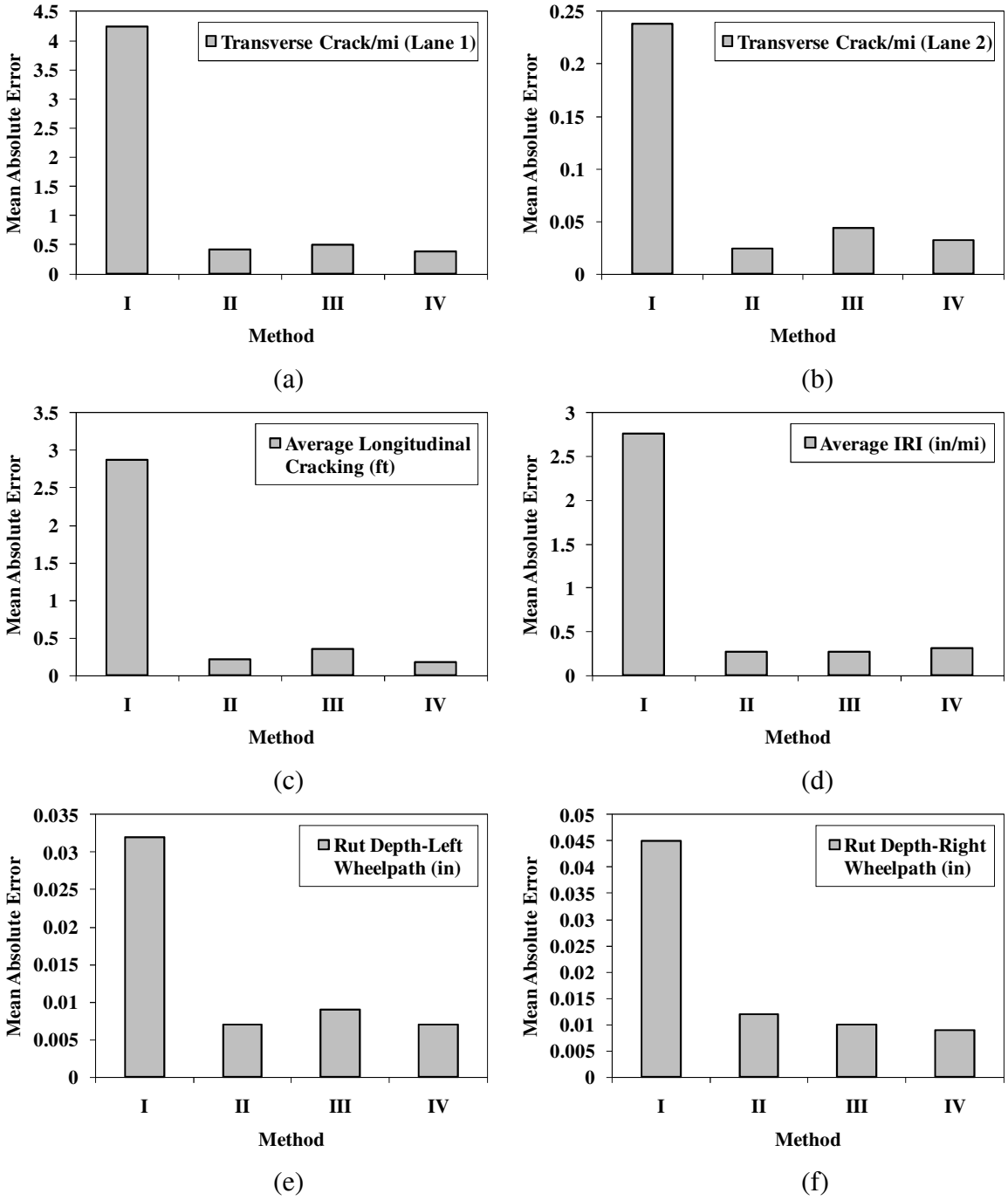


FIGURE 1 Mean absolute error of actual and imputed values from four methods.

Note: Method I refers to Mean Substitution, Method II refers to Interpolation, Method III refers to Regression Substitution, and Method IV refers to Multiple Imputation.

However, the Multiple Imputation method proposed in this study yielded the smallest errors for all distress types. Since the value of the mean absolute error varies across distresses, the aggregated errors by each method across distresses will shed profound insights into the overall robustness of each method in imputing missing data. The overall mean absolute error of the imputed values using Method I is 34.70%, while Method III results in 13.59% followed by 9.62% and 2.77% using Method II and Method IV respectively. The result reinforces the conclusion that the Method I generates the worst estimates of the missing data values, followed by Method III and Method II.

TABLE 2 Mean Absolute Error between Observed and Estimated Values

Imputation Method	Transverse Crack Density (Crack/mi)		Average IRI (in/mi)	Rut Depth (in)		Average Longitudinal Cracking (ft)
	Lane 1 (outside)	Lane 2 (inside)		Right RUT	Left RUT	
Mean Substitution	4.246	0.238	2.762	0.045	0.032	2.87
Interpolation	0.424	0.024	0.27	0.012	0.007	0.219
Regression Substitution	0.496	0.044	0.274	0.01	0.009	0.346
Multiple Imputation	0.386	0.032	0.307	0.009	0.007	0.176

APPLICATION ILLUSTRATION USING (LTTP) DATABASE

Based on the discussion in the preceding section, it is found that the Multiple Imputation method could closely approximate missing values from the actual values. This section presents an illustration of the application of the Multiple Imputation method to the LTTP (Long Term Pavement Performance) data obtained from the LTPP database (*23*).

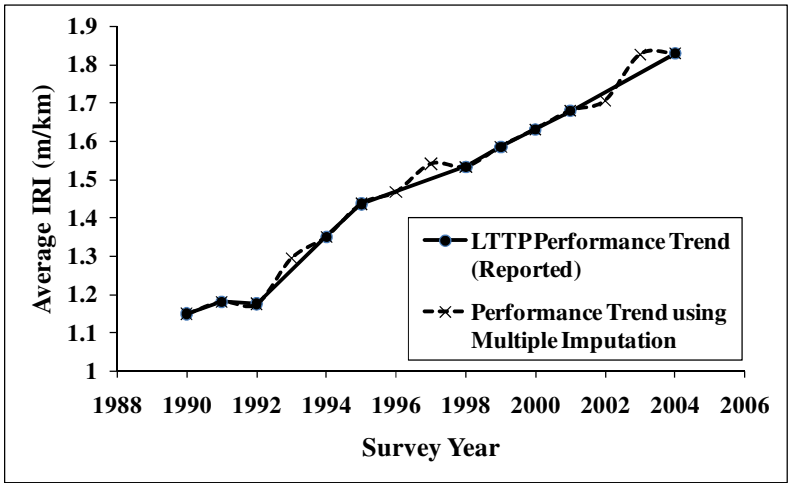
TABLE 3 Long Term Pavement Performance (LTTP) Data

Year	Average-IRI	Transverse Crack
1989	-	0
1990	1.1495	2
1991	1.1810	1
1992	1.1754	5
1993	1.3124	5
1994	1.3514	5
1995	1.4368	5
1996	1.4321	5
1997	1.5721	6
1998	1.5332	7
1999	1.586	7
2000	1.6316	7
2001	1.6794	9
2002	1.6719	10
2003	1.7042	24
2004	1.8294	62

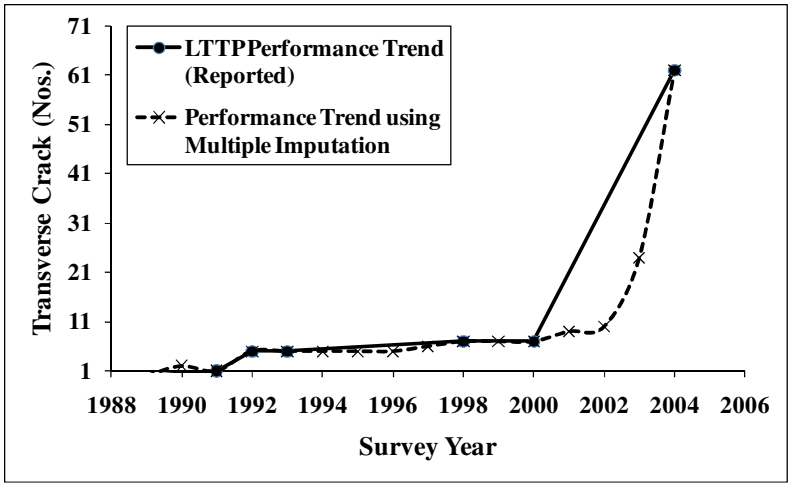
Note: Shaded cells represent imputed data for the missing values

The data set collected from LTPP database is shown in Table 3. It can be seen from the table that the data for some years are missing and the trend had been approximated as linear over time, even though in reality it might not be the case. Since there can be deviations from linearity which cannot be captured if the data is missing for certain number of years, the proposed Multiple Imputation method is applied to impute missing data. The data are imputed 5 times. The final imputed values are presented in Table 3.

The performance trends obtained from the Multiple Imputation method for roughness, and cracking are plotted in Fig. 2 along with the LTPP piece-wise linear performance curves. Significant differences are noticed between the two predicted performance curves. It can be expected that the pavement management decisions made from the two sets of performance curves could also be quite different.



(a) Multiple Imputation of LTPP data and roughness performance trend



(c) Multiple Imputation of LTPP data and cracking performance trend

FIGURE 2 Multiple Imputation of LTPP data and performance trends.

CONCLUSIONS

This paper has proposed a Multiple Imputation approach to address the missing data in pavement condition and performance database of a pavement management system. The rationale and applicability of the proposed approach has been explained. The quality of the

imputed data values by the proposed approach was assessed against values obtained using common conventional methods, including the listwise deletion method, the mean substitution method, and the interpolation method. It was found that the proposed approach is superior to other imputation approaches. It produced smaller deviation from actual values from the analysis using cross-validation technique. The proposed approach was also applied to data from the LTTP database and it was demonstrated that imputation of missing data using the proposed approach produces significantly different trends form the conventional linear interpolation method.

REFERENCES

1. Larson, C. D. and Forma, E. H. Application of Analytic Hierarchy Process to Select Project Scope for Video Logging and Pavement Condition Data Collection, In *Transportation Research Record: Journal of the Transportation Research Board, No. 1990*, Transportation Research Board, Washington, D.C., 2007, pp. 40-47.
2. Keleman, M., Henry, S. and Farrokhyar, A. *Pavement Management Manual*. Colorado Department of Transportation, Denver, CO., 2003.
3. National Cooperative Highway Research Program (NCHRP). *Automated Pavement Distress Collection Techniques*. National Cooperative Highway Research Program Synthesis Report No. 334, Transportation Research Board, Washington, D.C., 2004.
4. Amado, V. and Bernhardt, K. L. S. Knowledge Discovery in Pavement Condition Data. In *the 81st Annual Meeting of the Transportation Research Board (TRB)*, Washington D.C., 2002.
5. National Cooperative Highway Research Program (NCHRP). *Quality Management of Pavement Condition Data Collection*. National Cooperative Highway Research Program Synthesis Report No. 401, Transportation Research Board, Washington, D.C., 2009.
6. Lindly, J. K., Bell, F. and Sharif U. Specifying Automated Pavement Condition Surveys. Journal of the Transportation Research Forum, Vol. 44, No. 3, 2005, pp. 19-32.
7. Zhang and Smadi. "What is Missing in Quality Control of Contracted Pavement Distress Data Collection?" In the *90th Annual Meeting of Transportation Research Board*, Washington, D.C., 2009.
8. Allison P. D. *Missing Data*. Sage Publications, Inc., Thousand Oaks, CA., 2001.
9. Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*. 2nd edition, John Wiley, New York, 2002.
10. Bennett, C. R. Sectioning of Road Data for Pavement. *In the 6th International Conference on Managing Pavements*, Queensland, Australia, 2004.
11. Yang, J., Lu, J. J. and Gunaratne, M. Application of Neural Network Models for Forecasting of Pavement Crack Index and Pavement Condition Rating. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1699*, Transportation Research Board, Washington, D.C., 2003, pp. 3-12.
12. Schafer, J. L. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London, 1997.
13. Rubin, D. B. Inference and Missing Data, Biometrika, Vol. 63, No. 3, 1976, pp. 581-592.
14. Rubin, D. B. *Multiple Imputation for Survey Nonresponse*. Wiley, New York, 1987.
15. Tanner, M. A and Wong, W. H. The Calculation of Posterior Distributions by Data Augmentation. Journal of American Statistical Association, Vol. 82, 1987, pp.528–550.
16. Dempster, A. P., Laird, N. M. and Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Vol. 39, No. 1, 1977, pp.1-38.
17. Ripley, B. D. *Pattern Recognition and Neural Networks*, Cambridge Univ. Press, Cambridge, 1996.

18. Fraley, C. On Computing the Largest Fraction of Missing Information for the EM Algorithm and the Worst Linear Function for Data Augmentation. Computational Statistics & Data Analysis, Vol. 31, 1999. pp. 13–26.
19. Schafer, J. L. and Olsen, M. K. Multivariate Behavioral Research, Vol. 33, 1998, pp. 545–571.
20. Schafer, J. L. and Rubin, D. B. Multiple Imputation for Missing Data Problems, Short course presented at Joint Statistical Meetings, Dallas, TX, August, 1998.
21. Hill, T. and Lewicki, P. *Statistics: Methods and Applications*, Statsoft, Inc. 2006, pp. 652
22. Anderson, T. W., Darling, D. A. Asymptotic Theory of Certain "Goodness-of-fit" Criteria Based on Stochastic Processes. Annals of Mathematical Statistics, Vol. 23, 1952, pp.193–212.
23. Long Term Pavement Performance (LTTP) Database (2014). LTTP InfoPave. www.infopave.com. Accessed June 17, 2014.