Airport Pavement Missing Data Management and Imputation with Stochastic Multiple Imputation Model

J. Farhan and T. F. Fwa

In practice, missing data in pavement condition databases have been one of the most prevalent problems in airport pavement management systems. Missing data present problems in pavement performance analysis and uncertainties in pavement management decision making. A number of data imputation approaches are available for handling missing data. This paper examines the limitations of the conventional data imputation methods and proposes a stochastic multiple imputation (MI) approach to overcome major limitations associated with conventional data imputation methods. A case study is presented to appraise the effectiveness of the proposed approach against three conventional data imputation methods, namely, substitution by mean, substitution by interpolation, and substitution by regression methods. The roughness and friction data of a 4-km-long runway pavement and the roughness data of a 4-km-long taxiway pavement were considered in the study. The effectiveness of auxiliary variables in data imputation models was also demonstrated. Results from the performance appraisal indicated that the proposed stochastic MI method yielded the smallest errors for the roughness as well as friction data. Furthermore, the substitution by mean method resulted in imputed values with the highest amount of deviations from the observed values, followed by the substitution by regression method, and the substitution by interpolation method. Therefore, it is concluded that the proposed stochastic MI method outperformed conventional methods in handling missing runway and taxiway pavement roughness and friction data and provides an effective approach to impute missing data required in an airport pavement management system.

FAA encouraged airports to adopt a systematic method to collect data for the maintenance management of airport pavements (*1*). FAA also recommends that airports maintain minimal requirements on pavement inspections and record keeping, thereby ensuring that the data collected met the needs of the pavement management decision-making process in regard to quality control and assurance. Data quality assurance can be defined as the process of profiling the data to discover inconsistencies and anomalies such as missing data. As most modern airports rely heavily on data-driven applications, the need for complete and accurate pavement management data is growing. However, missing data in databases have been one of the most prevalent problems in pavement management systems (*2*). According to an NCHRP Synthesis Report (*3*), 61% of the pavement agencies reported using software routines to check for missing data elements, and some agencies reported mitigating missing data issues through re-collection (*4*).

Most data archival facilities in pavement management systems do not have a reliable system for dealing with missing runway performance and safety-related data. Although the principles of statistical quality assurance, in regard to the imputation of missing data, are well developed, their performance and application to the imputation of pavement management data are unclear. Therefore, the aim of this paper is threefold. First, present an overview and limitations of conventional methods to handle missing data. Second, propose a stochastic multiple imputation (MI) technique to mitigate the missing data issue. Last, examine the implications of missing data on the performance and safety characteristics of airport pavements and appraise the relative performance of conventional and proposed methods in handling missing data.

EXISTING DATA IMPUTATION METHODS

Traditionally, several approaches have been used for the purpose of estimating missing data. These techniques are categorized as case deletion, prediction rules, and least square approximation approaches, as briefly discussed below.

Deletion Methods

The case deletion method is by far the most common and simplest approach. It involves neglecting cases with missing data, using only the remaining data in the database for analyses (*5*). Two common techniques are used for deleting data with missing entries: listwise deletion and pairwise deletion. (Editor's note: listwise refers to the process whereby data are searched list by list or record by record for missing values; pairwise refers to the process whereby covariance between two variables is computed from paired-complete data, excluding cases with missing values on one or both of the variables.) In the former, the entire list of observations containing missing data is removed from the database (*6*). This method results in a reduction in the sample size. In the latter method, the incomplete record is retained only if the missing variable is not required in the analysis under consideration. Allison refers to the fact that unless the data are missing completely at random (MCAR), pairwise deletion produces estimates that may be seriously biased (*7*).

Department of Civil and Environmental Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Corresponding author: T. F. Fwa, ceefwatf@nus.edu.sg.

Transportation Research Record: Journal of the Transportation Research Board, No. 2336, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 43–54.
DOI: 10.3141/2336-06

Nonstochastic Imputation Methods

Nonstochastic imputation involves substituting a plausible value for data that are missing. These nonstochastic imputation methods fill in missing values to maintain a full sample so as to analyze completed data. The most common techniques included in this category are substitution methods (mean, interpolation, or regression) and pattern-matching imputation methods. In the mean substitution approach, the missing values are substituted by a single constant value for a particular variable. The constant value is equivalent to the mean of available values for that variable. Although the method is simple, it reduces the variance of the variable, which also attenuates covariances that the variable has with other variables in the data set (*6*). In the interpolation substitution approach, missing values are replaced by using linear interpolation. This method typically assumes a linear trend in the vicinity of the missing data, which is a simplistic (and may not be valid) assumption (*8, 9*). The regression substitution approach involves regression on the nonmissing data to predict expected values for the missing data (*10, 11*). This technique tends to underestimate the variance and the covariance of the data (*6*). Pattern-matching imputation methods impute values on the basis of matching the record with missing data with similar records without missing data (*12*). Bennett (*13*) observed that this method results in less bias than does listwise deletion or mean imputation; however strong evidence of the accuracy of this method is unavailable (*12*).

Stochastic Imputation Methods

Stochastic imputation methods reflect the uncertainty of predicted missing values by adding a stochastic, or random, component (*14*). Auxiliary variables deemed valuable in imputation analysis may be included to improve the prediction of missing values (*15*). The expectation maximization (EM) algorithm is an iterative regression technique in which the missing variables are regressed on the available data (*6, 12, 13*). This method involves two steps, namely, expectation (E) and maximization (M) (*6*). The first step computes the expected value for the missing data, and the second is aimed at maximizing the likelihood function of the expected variables obtained in the previous step. The EM strategy is based on the notion that the missing data have information that is useful in estimating various parameters, such as mean vector and covariance matrix, and the estimated parameter has information that is useful in finding the most likely value of the missing data (*13*).

Limitation of Existing Methods

Each of the data imputation approaches described in the preceding sections contains assumptions that lead to limitations in its applications. Deletion methods result in reduced sample size and can be a serious impediment for analyses. Similarly, mean substitution and regression methods underestimate the variance and the covariance of the data, and the pattern-matching imputation method is found to produce inaccurate results at times. The defect of the EM algorithm is that the standard errors and confidence intervals are not provided. In light of the limitations of the existing data imputation approaches, this paper proposes a stochastic MI model using auxiliary variables to impute missing airport pavement performance data. The proposed approach is explained in the following sections, and a case study is presented to illustrate the effectiveness of the approach as compared with existing methods.

CONCEPT OF MI APPROACH FOR MISSING DATA

The MI approach is an improvement over the EM approach. It involves the degree of similarity or difference between several imputed data sets as additional information for the standard errors of parameter estimates. This improvement makes the solution less biased than the EM algorithm (*16*). Rubin pointed out 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" (*17*).

The MI replaces the missing values by $m > 1$ plausible values drawn from their predictive distributions. The variation among the m imputations reflects the uncertainty with which the missing values can be predicted from the observed values. The final results are obtained by averaging the parameter estimates across the multiple analyses, which results in an unbiased parameter estimate. These combined standard errors from the multiple imputed data sets are used for significance testing or the construction of confidence intervals around these parameter estimates or both. The technique can be performed by using the data augmentation (DA) algorithm (*18*). However, the EM algorithm is considered a preferred approach in establishing initial estimates such as mean and covariance for DA to begin with and is used in the present study (*19*).

EM Algorithm

The EM algorithm is a general method for obtaining maximum likelihood estimates of parameters in problems with incomplete data (*20*). 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 [such as means (μ) and covariance matrix (V)]. Therefore, complete data, Y_{com}, can be defined as $Y_{\text{com}} = (Y_{\text{obs}}, Y_{\text{mis}})$. With the complete data log likelihood function, $\ln L_c(\theta) = \ln f(Y_{\text{com}}|\theta)$, the expected complete data log likelihood function can be defined as

$$Q(\theta|\theta') = \mathrm{E}\{\ln[f(Y_{\text{com}}|\theta)]Y_{\text{obs}}, \theta^t\} \qquad (1)$$

where

θ^t = current parameter estimate,
θ = new parameter to update, and
t = iteration number.

The EM algorithm begins with some value of θ and alternates between two steps as follows (*21*):

1. Expectation step (E step), that is, compute $Q(\theta|\theta^{(t)})$ as a function of θ and
2. Maximization step (M step), that is, 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 (*20*), and the rate of convergence is proportional to the amount of unobserved or missing information in a data matrix (*22*).

DA Algorithm

The DA algorithm requires starting values for the mean and covariance matrix, and an appropriate approach is to calculate these values by using the EM algorithm. DA makes use of MI, and the premise behind generating m/s is that instead of a point estimate being used 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 also provides insight into how much variance there is in the estimate.

The DA process is similar in nature to that of the EM algorithm, that is, it is an iterative process that alternately fills in the missing data while crafting inferences about the unknown parameters. However, in contrast to the EM algorithm, this process is performed in a stochastic manner (*23*). A random imputation of missing data under assumed values of the parameters is performed by DA, followed by the estimation of new parameters from a Bayesian posterior distribution based on the observed and imputed data (*24*). Beginning at some value of parameter θ, each iteration of the DA algorithm alternates between two steps as follows (*24*):

1. Imputation step (I step): draws $Y_{\text{mis}}^{t+1} \sim P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^{(t)})$ and
2. Posterior step (P step): draws $\theta^{(t+1)} \sim P(\theta|Y_{\text{obs}}, \theta^{(t+1)})$.

This process of alternately imputing and establishing missing data and parameters, respectively, creates a Markov chain that finally converges in distribution (*24*).

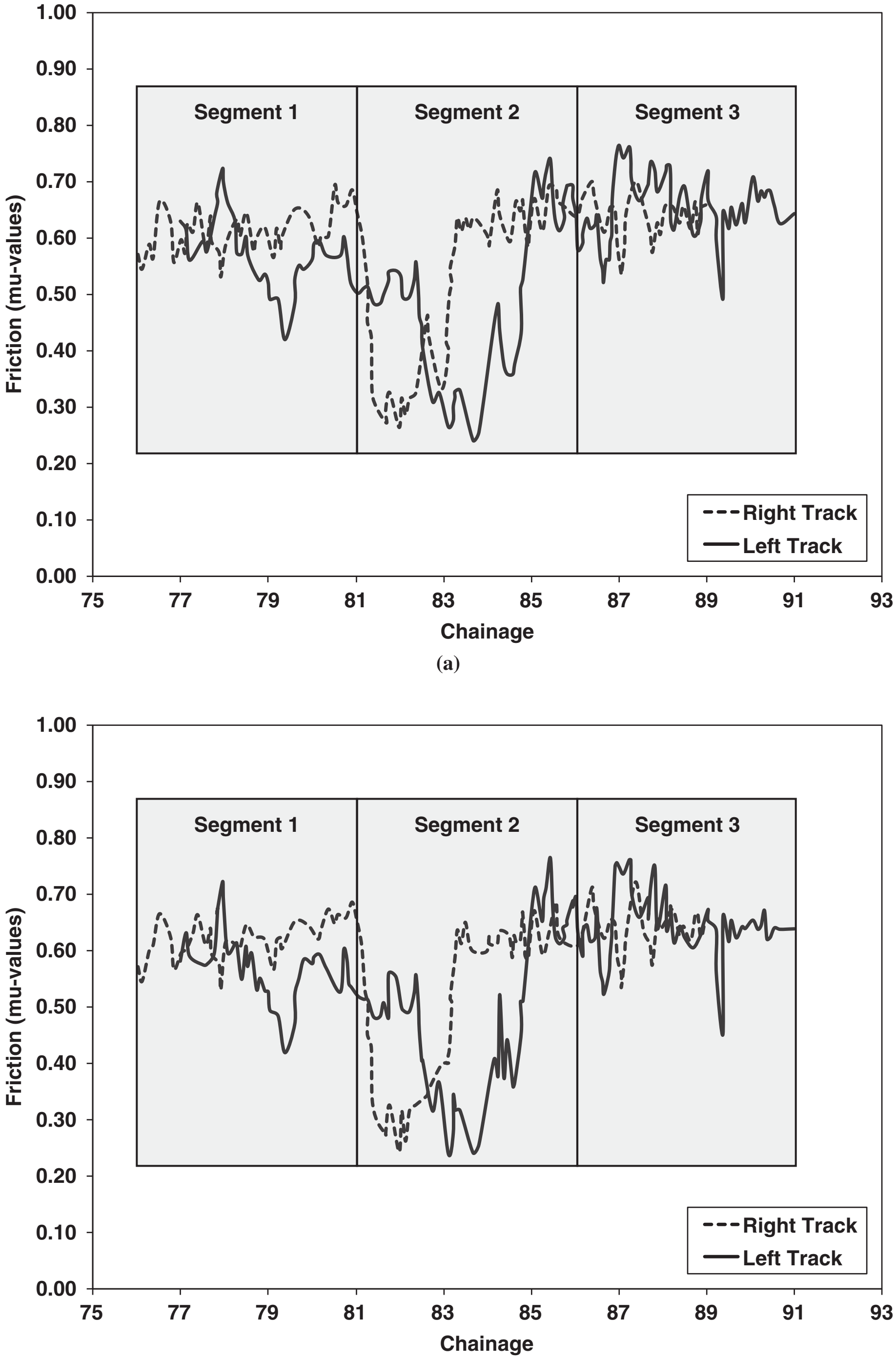
PROPOSED MI PROCEDURE FOR AIRPORT PAVEMENT MISSING DATA

For the purposes of illustration, the roughness and friction data of a 4-km-long runway pavement and the roughness data of a 4-km-long taxiway pavement were considered. The runway and the taxiway were divided uniformly into four segments, each 1 km in length, for the analysis. These four segments are further divided into five subsections each, for roughness characterization, in regard to straightedge index (SE) (*25*), international roughness index (IRI) (ASTM E1926), and Boeing bump index (BBI) (*26*). Table 1 shows the roughness indexes computed from the measured roughness data. As for the friction data analysis, the runway was divided into three segments of equal length. The friction data are given in Figure 1*a*.

TABLE 1 Airport Pavement Roughness Characteristics from Measured Data

Roughness Index	Runway			Taxiway			Runway			Taxiway		
	LR	CR	RR	LT	CT	RT	LR	CR	RR	LT	CT	RT
Segment 1							Segment 2					
SE												
1	0.25	0.29	0.30	0.67	0.26	0.22	0.11	0.06	0.21	0.30	0.13	0.19
2	0.22	0.49	0.30	0.24	0.22	0.28	0.10	0.31	0.18	0.17	0.10	0.16
3	0.08	0.27	0.31	0.19	0.32	0.24	0.05	0.05	0.20	0.06	0.15	0.17
4	0.32	0.17	0.19	0.20	0.30	0.25	0.35	0.10	0.08	0.09	0.30	0.27
5	0.10	0.21	0.17	0.22	0.26	0.13	0.16	0.23	0.24	0.11	0.19	0.11
BBI												
1	0.27	0.38	0.22	0.47	0.25	0.18	0.14	0.10	0.22	0.54	0.27	0.30
2	0.29	0.46	0.34	0.23	0.29	0.27	0.16	0.35	0.18	0.29	0.13	0.20
3	0.18	0.48	0.27	0.33	0.25	0.25	0.10	0.13	0.17	0.13	0.15	0.14
4	0.36	0.30	0.32	0.31	0.24	0.23	0.26	0.19	0.10	0.08	0.23	0.17
5	0.19	0.26	0.26	0.23	0.25	0.12	0.16	0.24	0.23	0.16	0.25	0.18
IRI												
1	0.36	0.35	0.34	1.03	0.24	0.29	0.10	0.06	0.25	0.20	0.08	0.24
2	0.24	0.54	0.42	0.29	0.30	0.30	0.12	0.25	0.14	0.15	0.10	0.14
3	0.07	0.29	0.14	0.13	0.34	0.37	0.04	0.06	0.21	0.05	0.17	0.26
4	0.39	0.22	0.17	0.20	0.42	0.35	0.44	0.10	0.08	0.10	0.41	0.30
5	0.11	0.19	0.12	0.17	0.26	0.13	0.24	0.22	0.33	0.11	0.24	0.17
Segment 3							Segment 4					
SE												
1	0.11	0.11	0.29	0.11	0.05	0.02	0.26	0.25	0.10	0.15	0.36	0.17
2	0.18	0.09	0.15	0.04	0.30	0.06	0.15	0.08	0.10	0.15	0.21	0.15
3	0.11	0.13	0.25	0.14	0.35	0.18	0.49	0.46	0.85	0.30	0.17	0.32
4	0.29	0.15	0.24	0.11	0.11	0.04	0.41	0.36	0.27	0.41	0.51	0.47
5	0.30	0.49	0.37	0.15	0.17	0.15	0.46	0.60	0.52	1.49	1.37	1.76
BBI												
1	0.17	0.15	0.23	0.10	0.09	0.03	0.27	0.31	0.16	0.22	0.39	0.33
2	0.22	0.13	0.15	0.07	0.23	0.12	0.20	0.14	0.16	0.22	0.24	0.21
3	0.14	0.14	0.18	0.11	0.20	0.25	0.98	1.00	1.19	0.23	0.16	0.25
4	0.22	0.19	0.22	0.11	0.10	0.06	0.65	0.65	0.55	0.64	0.84	0.70
5	0.38	0.55	0.35	0.15	0.16	0.14	0.51	0.75	0.71	1.90	2.08	2.18
IRI												
1	0.12	0.11	0.41	0.14	0.06	0.02	0.32	0.32	0.19	0.16	0.35	0.21
2	0.17	0.09	0.14	0.05	0.39	0.06	0.12	0.11	0.69	0.18	0.20	0.14
3	0.15	0.18	0.22	0.14	0.53	0.19	0.31	0.31	0.80	0.33	0.16	0.31
4	0.36	0.12	0.42	0.15	0.18	0.04	0.27	0.33	0.20	0.37	0.42	0.55
5	0.41	0.57	0.42	0.16	0.16	0.20	0.48	0.57	0.53	1.16	1.41	2.16

NOTE: L, C, and R represent left, center, and right tracks respectively. R and T represent runway and taxiway respectively. SE, BBI, and IRI reflect Straightedge Index in mm, Boeing Bump Index, and International Roughness Index in m/km respectively.



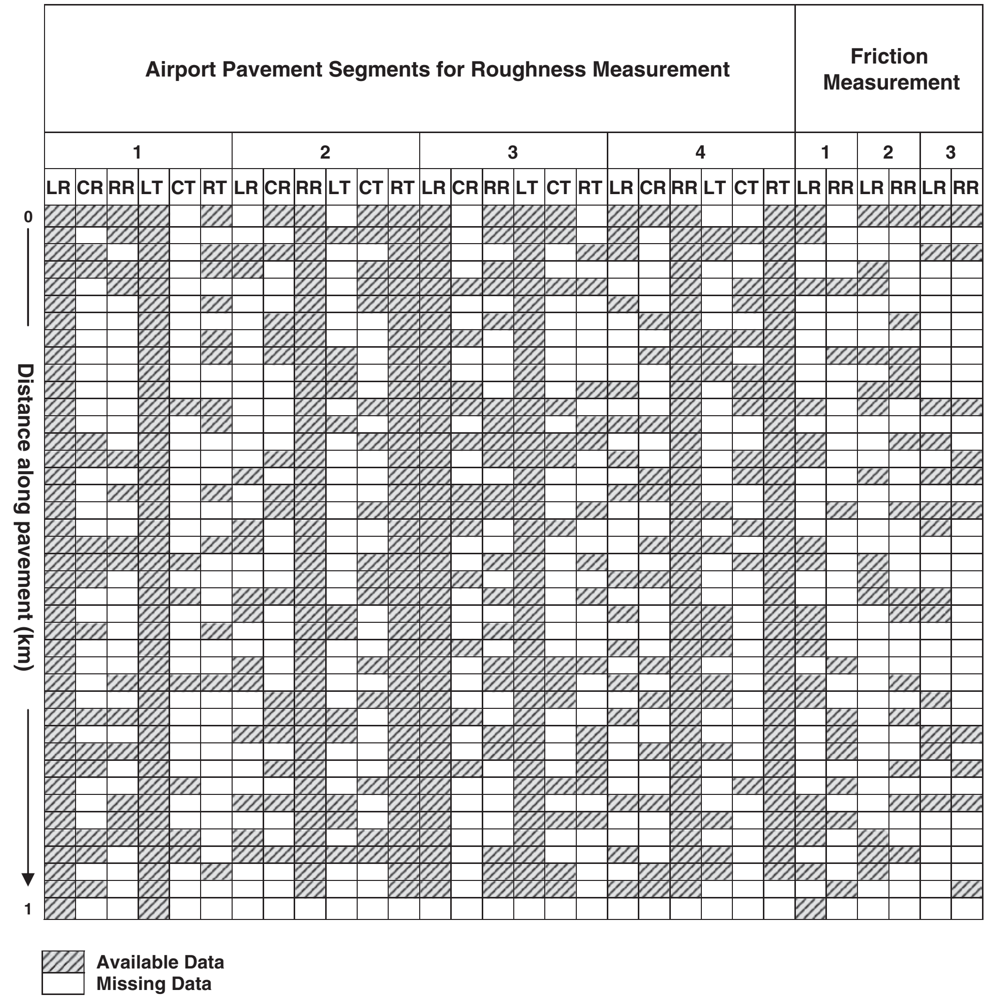
(a)

(b)

FIGURE 1 Runway friction measurements in Mu-Meter readings: (*a*) before data imputation and (*b*) after data imputation with the proposed approach.

For the analysis of the effectiveness of the different data imputation methods, data in the pavement roughness and friction records were randomly deleted to create various patterns of data missing completely at random (MCAR) at the rate of 35%. MCAR is representative of a scenario in which the missing observations simply represent a random sample from within all observations in the data set. Because no structural association exists between missing and observed data, missing values do not alter the original distributional relationships between variables. A graphical representation of the missing data pattern for the entire stretch of runway and taxiway is depicted in Figure 2.

The basic step in the MI method is to create values to be substituted for the missing data; therefore a need arises to identify some model that will allow the creation of imputes based on auxiliary or other variables in the data set. Under the multivariate normal imputation model, the imputation of an observation is based on



L, C, and R represent left, centre, and right tracks respectively
R and T represent runway and taxiway respectively
Roughness and friction data were collected at equal intervals along the runway and taxiway represented by number of rows

FIGURE 2 State of missing data and pattern in airport pavement database.

regressing a variable, with missing data, on the other variables in the data set. Because the regression method is used to impute the values for the missing data, the imputation model is selected to be rich enough to preserve the associations or relationships between variables. For instance, under Segment 1, the roughness data in the left runway track are included as an auxiliary variable to impute missing roughness data in the right track. The model would take the following form:

$$r_i = \alpha l_i + \beta x_i + \gamma + s_{rlx}\varepsilon_i \qquad (2)$$

where

r and l = right and left runway track roughness variables, respectively;
x = location of measurement from any reference point;
ε = random draw that follows the normal (0, 1) distribution;
s_{rlx} = square root of mean square error;
α, β, and γ = calibration constants; and
i = number of data points in the data set concerned.

Following the imputation procedure just described, the imputed value will contain a random error component. Each time imputation is performed a slightly different result will be obtained, followed by the estimation of new parameters from a Bayesian posterior distribution on the basis of the observed and imputed data (*24*). The MI method procedure adopted in the study involves the following steps:

Step I. Data Transformation. Transform the data for all variables to approximately normal before imputation using a logit, log, or square root transformation function. Next, transform back to their original scale after imputation. The logit or logistic transformation is defined as (*27*)

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \qquad (3)$$

where p stands for probability or proportion. In the case of elevation profile, a constant value to the data before the log transformation is applied can be added to handle negative values.

TABLE 2 Airport Pavement Roughness from Imputed Data Using MI

	Runway			Taxiway			Runway			Taxiway		
Roughness Index	LR	CR	RR	LT	CT	RT	LR	CR	RR	LT	CT	RT
Segment 1							Segment 2					
SE												
1	0.25	0.31	0.25	0.67	0.42	0.33	0.10	0.19	0.21	0.29	0.13	0.19
2	0.22	0.43	0.25	0.24	0.49	0.36	0.14	0.33	0.18	0.24	0.10	0.16
3	0.08	0.29	0.26	0.19	0.52	0.16	0.08	0.21	0.20	0.28	0.15	0.17
4	0.32	0.36	0.15	0.20	0.50	0.24	0.43	0.17	0.08	0.22	0.30	0.27
5	0.10	0.26	0.30	0.22	0.34	0.24	0.17	0.21	0.24	0.25	0.19	0.11
BBI												
1	0.27	0.25	0.21	0.47	0.36	0.21	0.13	0.22	0.22	0.56	0.27	0.30
2	0.29	0.40	0.27	0.23	0.35	0.26	0.17	0.34	0.18	0.27	0.13	0.20
3	0.18	0.32	0.32	0.33	0.30	0.22	0.10	0.23	0.17	0.24	0.15	0.14
4	0.36	0.42	0.42	0.31	0.49	0.25	0.28	0.17	0.10	0.23	0.23	0.17
5	0.19	0.31	0.31	0.23	0.37	0.19	0.14	0.28	0.23	0.18	0.25	0.18
IRI												
1	0.36	0.47	0.33	1.03	0.24	0.45	0.12	0.05	0.25	0.15	0.08	0.24
2	0.24	0.37	0.27	0.29	0.30	0.57	0.14	0.19	0.14	0.22	0.10	0.14
3	0.07	0.30	0.44	0.13	0.34	0.15	0.07	0.04	0.21	0.14	0.17	0.26
4	0.39	0.49	0.47	0.20	0.48	0.19	0.53	0.06	0.08	0.31	0.41	0.30
5	0.11	0.27	0.49	0.17	0.27	0.31	0.28	0.21	0.33	0.32	0.24	0.17
Segment 3							Segment 4					
SE												
1	0.11	0.18	0.34	0.11	0.05	0.15	0.27	0.25	0.20	0.32	0.36	0.39
2	0.18	0.37	0.19	0.04	0.30	0.12	0.16	0.08	0.20	0.27	0.21	0.10
3	0.11	0.23	0.29	0.14	0.35	0.24	0.57	0.46	0.89	0.25	0.17	0.17
4	0.29	0.33	0.08	0.11	0.11	0.24	0.53	0.36	0.32	0.41	0.51	0.43
5	0.30	0.40	0.37	0.15	0.17	0.24	0.59	0.60	0.53	1.28	1.37	1.35
BBI												
1	0.17	0.16	0.33	0.10	0.09	0.10	0.29	0.31	0.48	0.31	0.39	0.41
2	0.22	0.25	0.15	0.07	0.23	0.11	0.18	0.14	0.32	0.42	0.24	0.12
3	0.14	0.24	0.18	0.11	0.20	0.26	0.92	1.00	1.38	0.31	0.16	0.14
4	0.22	0.19	0.15	0.11	0.10	0.14	0.65	0.65	0.53	0.77	0.84	0.74
5	0.38	0.34	0.33	0.15	0.16	0.20	0.50	0.75	0.72	1.69	2.08	2.12
IRI												
1	0.12	0.19	0.37	0.14	0.06	0.17	0.29	0.32	0.14	0.39	0.35	0.47
2	0.17	0.09	0.20	0.05	0.39	0.11	0.17	0.11	0.65	0.18	0.20	0.14
3	0.15	0.27	0.47	0.14	0.53	0.20	1.06	0.31	1.14	0.16	0.16	0.20
4	0.36	0.42	0.48	0.15	0.18	0.04	0.20	0.33	0.24	1.05	0.42	0.33
5	0.41	0.59	0.47	0.16	0.16	0.31	0.61	0.57	0.38	0.90	1.41	1.07

Step II. Imputation Using EM. Generate estimates of missing values for the data matrix by using the EM algorithm with the convergence criterion that the maximum relative parameter change in the value of any parameter during iterative process is less than 0.0001.

Step III. Imputation Using DA. With the initial parameter estimates from the EM algorithm serving as the basis for the DA algorithm, generate imputed data and new parameter estimates, as explained in the preceding section. The commonly adopted practice of 10 imputations is applied in this study (*6, 19*).

Step IV. Synthesis of Estimates. Average over the multiple estimates to obtain the final set of estimates (*17*).

With the imputed data obtained as described above, the roughness indexes and friction values are computed for the runway and taxiway segments. Auxiliary variables were used in the imputation analysis, and the imputation process infuses the imputed values with the information from the auxiliary variables. For instance, for Segment 1, the roughness data in the left runway track are included as an auxiliary variable to impute missing roughness data in the right track. Similarly, track runway and taxiway roughness data are incorporated into the imputation analysis as auxiliary variables for Segments 2, 3, and 4, respectively. The final results are summarized in Table 2 for roughness and Figure 1*b* for friction.

COMPARISON OF EXISTING IMPUTATION STRATEGIES AGAINST THE PROPOSED MI APPROACH

In this section the performance of the proposed MI approach for estimating missing airport pavement performance and friction data is compared against the following three existing imputation methods: (*a*) the substitution by mean method, (*b*) substitution by linear interpolation and extrapolation using adjacent points, and (*c*) substitution by regression method.

Method of Comparison

A number of measures can be used to evaluate imputation accuracy, such as root mean square error (RMSE), based on squared errors, as shown in Equation 4 (*28*). In addition, measures based on absolute error such as the mean absolute percentage error (MAPE) shown in Equation *5* are predominantly used. Because MAPE is to some extent scale dependent, such as when very low values or integers are evaluated (i.e., a value of one or two), the size of the measure can be easily inflated. Therefore, it is best to use a combination of measures to evaluate the accuracy of imputation. Both RMSE and MAPE, as in Equations 4 and 5, respectively, are used to compare the accuracy of missing data imputation by using different approaches.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(O_t - I_t)^2}{n}} \tag{4}$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{O_t - I_t}{O_t}\right| \tag{5}$$

where *O, I,* and *n* stand for observed values, imputed values, and total number of values imputed, respectively.

Comparison of Imputation Results

The imputation results by the proposed MI approach are summarized in Table 2. The corresponding imputation results for the three existing methods, namely, the substitution by mean method, the substitution by linear interpolation and extrapolation method, and the substitution by regression method, are presented in Tables 3, 4, and 5, respectively.

The relative quality of data imputation from the four methods is assessed by using MAPE and RMSE, as presented in Figures 3 and 4. As can be seen from the figures, the mean substitution method resulted largely in imputed values with the highest amount of deviations from the observed values, followed by the regression substitution method and the interpolation method. The stochastic MI method, proposed in this study, yielded the smallest errors for roughness as well as for friction data. Because the roughness of the runway and taxiway increases progressively from Segments 1 to 4, the second worst performing method switches from substitution by regression to substitution by linear interpolation in Segment 4. Nevertheless, the proposed MI approach performs consistently better than any of the other approaches analyzed for missing data imputation.

TABLE 3 Airport Pavement Roughness Imputed Data Using Substitution by Mean

Roughness Index	Runway			Taxiway			Runway			Taxiway		
	LR	CR	RR	LT	CT	RT	LR	CR	RR	LT	CT	RT
Segment 1							Segment 2					
SE												
1	0.25	0.15	0.15	0.67	0.00	0.67	0.26	0.21	0.21	0.18	0.43	0.19
2	0.22	0.26	0.23	0.24	0.09	0.29	0.36	0.42	0.18	0.14	0.36	0.16
3	0.08	0.47	0.37	0.19	0.84	0.15	0.27	0.25	0.20	0.05	0.27	0.17
4	0.32	0.81	1.18	0.20	0.25	0.50	0.31	0.24	0.08	0.06	0.40	0.27
5	0.10	0.60	0.57	0.22	0.75	0.58	0.58	0.52	0.24	0.28	0.46	0.11
BBI												
1	0.27	0.08	0.10	0.47	0.01	0.42	0.24	0.16	0.22	0.14	0.38	0.30
2	0.29	0.15	0.12	0.23	0.07	0.27	0.26	0.27	0.18	0.12	0.23	0.20
3	0.18	0.34	0.28	0.33	0.43	0.11	0.28	0.21	0.17	0.04	0.22	0.14
4	0.36	0.67	0.62	0.31	0.30	0.36	0.24	0.16	0.10	0.06	0.27	0.17
5	0.19	0.57	0.50	0.23	0.54	0.38	0.39	0.33	0.23	0.16	0.32	0.18
IRI												
1	0.36	0.24	0.22	1.03	0.00	0.88	0.30	0.27	0.25	0.28	0.62	0.24
2	0.24	0.34	0.27	0.29	0.13	0.36	0.40	0.53	0.14	0.15	0.54	0.14
3	0.07	0.39	0.59	0.13	1.14	0.21	0.35	0.34	0.21	0.04	0.35	0.26
4	0.39	1.30	1.76	0.20	0.43	0.73	0.29	0.27	0.08	0.08	0.59	0.30
5	0.11	0.77	0.97	0.17	0.90	0.72	0.78	0.79	0.33	0.43	0.65	0.17
Segment 3							Segment 4					
SE												
1	0.11	0.45	0.80	0.11	0.12	0.71	1.99	0.25	1.46	1.27	0.36	0.20
2	0.18	0.43	0.22	0.04	0.22	0.06	2.19	0.08	1.34	1.33	0.21	0.24
3	0.11	0.11	0.41	0.14	0.02	0.39	1.95	0.46	2.06	0.71	0.17	0.21
4	0.29	0.80	0.55	0.11	0.11	0.14	0.24	0.36	1.49	1.05	0.51	0.51
5	0.30	0.41	0.40	0.15	0.24	0.11	2.28	0.60	3.05	2.23	1.37	0.04
BBI												
1	0.17	0.30	0.43	0.10	0.10	0.37	1.57	0.31	1.17	0.70	0.39	0.14
2	0.22	0.29	0.17	0.07	0.15	0.07	1.64	0.14	1.46	0.78	0.24	0.16
3	0.14	0.10	0.25	0.11	0.02	0.28	1.49	1.00	1.65	0.73	0.16	0.14
4	0.22	0.55	0.44	0.11	0.11	0.13	0.34	0.65	1.14	0.68	0.84	0.27
5	0.38	0.31	0.42	0.15	0.15	0.10	1.81	0.75	2.22	1.99	2.08	0.04
IRI												
1	0.12	0.53	1.06	0.14	0.19	1.13	2.37	0.32	1.52	1.96	0.35	0.33
2	0.17	0.56	0.31	0.05	0.32	0.05	2.48	0.11	1.92	1.85	0.20	0.30
3	0.15	0.12	0.61	0.14	0.04	0.55	2.17	0.31	2.80	0.61	0.16	0.31
4	0.36	0.97	0.57	0.15	0.16	0.23	0.22	0.33	1.75	1.54	0.42	0.76
5	0.41	0.64	0.45	0.16	0.35	0.12	2.63	0.57	3.93	2.61	1.41	0.09

TABLE 4 Airport Pavement Roughness Imputed Data Using Substitution by Interpolation

	Runway			Taxiway			Runway			Taxiway		
Roughness Index	LR	CR	RR	LT	CT	RT	LR	CR	RR	LT	CT	RT
Segment 1							Segment 2					
SE												
1	0.25	0.09	0.06	0.67	0.01	0.20	0.04	0.05	0.21	0.01	0.13	0.19
2	0.22	0.16	0.04	0.24	0.00	0.12	0.03	0.06	0.18	0.12	0.03	0.16
3	0.08	0.09	0.11	0.19	0.09	0.02	0.03	0.03	0.20	0.01	0.03	0.17
4	0.32	0.13	0.16	0.20	0.02	0.08	0.02	0.04	0.08	0.04	0.03	0.27
5	0.10	0.13	0.06	0.22	0.11	0.00	0.03	0.11	0.24	0.04	0.07	0.11
BBI												
1	0.27	0.08	0.06	0.47	0.00	0.16	0.07	0.10	0.22	0.03	0.20	0.30
2	0.29	0.15	0.07	0.23	0.00	0.12	0.04	0.11	0.18	0.11	0.06	0.20
3	0.18	0.24	0.21	0.33	0.13	0.04	0.05	0.04	0.17	0.01	0.02	0.14
4	0.36	0.20	0.23	0.31	0.06	0.12	0.04	0.08	0.10	0.04	0.05	0.17
5	0.19	0.22	0.13	0.23	0.16	0.00	0.05	0.15	0.23	0.08	0.10	0.18
IRI												
1	0.36	0.14	0.10	1.03	0.02	0.24	0.07	0.05	0.25	0.02	0.16	0.24
2	0.24	0.21	0.03	0.29	0.00	0.18	0.02	0.06	0.14	0.17	0.04	0.14
3	0.07	0.08	0.14	0.13	0.08	0.02	0.05	0.04	0.21	0.01	0.04	0.26
4	0.39	0.17	0.14	0.20	0.03	0.11	0.03	0.04	0.08	0.06	0.03	0.30
5	0.11	0.17	0.10	0.17	0.11	0.00	0.03	0.11	0.33	0.05	0.09	0.17
Segment 3							Segment 4					
SE												
1	0.11	0.01	0.18	0.11	0.05	0.01	0.07	0.25	0.02	0.12	0.36	0.14
2	0.18	0.05	0.12	0.04	0.06	0.04	0.02	0.08	0.07	0.08	0.21	0.01
3	0.11	0.03	0.09	0.14	0.00	0.15	0.58	0.46	0.75	0.05	0.17	0.02
4	0.29	0.02	0.11	0.11	0.03	0.02	0.16	0.36	0.30	0.11	0.51	0.23
5	0.30	0.00	0.17	0.15	0.11	0.06	0.22	0.60	0.07	0.94	1.37	0.01
BBI												
1	0.17	0.02	0.21	0.10	0.05	0.02	0.07	0.31	0.01	0.18	0.39	0.14
2	0.22	0.07	0.10	0.07	0.11	0.08	0.04	0.14	0.14	0.13	0.24	0.01
3	0.14	0.04	0.12	0.11	0.01	0.21	0.77	1.00	0.99	0.05	0.16	0.04
4	0.22	0.03	0.13	0.11	0.04	0.04	0.31	0.65	0.48	0.21	0.84	0.24
5	0.38	0.00	0.24	0.15	0.11	0.06	0.19	0.75	0.11	0.89	2.08	0.03
IRI												
1	0.12	0.02	0.24	0.14	0.09	0.01	0.12	0.32	0.05	0.16	0.35	0.25
2	0.17	0.06	0.17	0.05	0.05	0.04	0.03	0.11	0.07	0.08	0.20	0.01
3	0.15	0.03	0.11	0.14	0.01	0.16	0.58	0.31	0.97	0.07	0.16	0.02
4	0.36	0.02	0.11	0.15	0.03	0.02	0.26	0.33	0.35	0.12	0.42	0.30
5	0.41	0.00	0.26	0.16	0.15	0.09	0.29	0.57	0.10	0.99	1.41	0.02

The benefits of using auxiliary variables in the analysis of the proposed MI approach can be seen from Figure 3. For instance, LR served as an auxiliary variable for imputing roughness data of CR and RR in Segment 1 resulting in an increased MAPE or RMSE value for CR in comparison with RR as shown in Figure 3, *a* and *b*. The same phenomenon is observed in Segments 2 and 3, in which RR and LR are used as auxiliary variables, respectively, as shown in Figure 3, *c* through *f*. This result is indicative of the situation in which a better correlation between right track and left track roughness measurements exists than that of between centerline and left and right track.

Implications of Imputation Results on Airport Pavement Maintenance Management

For the present case study of airport pavement missing data and subsequent imputation using the methods discussed in the preceding section, Figure 5 plots the imputed results of roughness indexes against the actual values of roughness indexes for each of the four imputation methods analyzed. It can be observed from the figure that the proposed MI approach produced the least deviations from the line of equality, and the substitution by mean method produced the widest deviation from the equality line. The substitution by interpolation method resulted in underestimation of roughness values in practically all of the cases.

The quantitative assessment of the imputation performance of the proposed approach against existing methods in imputing airport pavement missing data can be measured by using the Pearson correlation coefficient *r,* which reflects the degree of a linear relationship between any two sets of results evaluated as follows (*29*):

$$r = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} \times \sqrt{n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2}} \qquad (6)$$

where

x_i = value from observation i on variable X,
y_i = value from observation i on variable Y, and
n = number of values in each data set, $i = 1, \ldots, n$.

TABLE 5 Airport Pavement Roughness Imputed Data Using Substitution by Regression

Roughness	Runway			Taxiway			Runway			Taxiway		
Index	LR	CR	RR	LT	CT	RT	LR	CR	RR	LT	CT	RT
Segment 1							Segment 2					
SE												
1	0.25	0.21	0.25	0.67	0.68	0.31	0.20	0.17	0.21	0.19	0.13	0.19
2	0.22	0.43	0.46	0.24	0.17	0.24	0.23	0.39	0.18	0.19	0.28	0.16
3	0.08	0.21	0.18	0.19	0.22	0.21	0.21	0.20	0.20	0.09	0.26	0.17
4	0.32	0.42	0.36	0.20	0.44	0.15	0.14	0.26	0.08	0.18	0.19	0.27
5	0.10	0.29	0.24	0.22	0.40	0.15	0.09	0.45	0.24	0.14	0.33	0.11
BBI												
1	0.27	0.22	0.25	0.47	0.48	0.19	0.15	0.13	0.22	0.22	0.23	0.30
2	0.29	0.40	0.38	0.23	0.24	0.25	0.20	0.25	0.18	0.18	0.17	0.20
3	0.18	0.27	0.29	0.33	0.32	0.19	0.20	0.19	0.17	0.08	0.16	0.14
4	0.36	0.45	0.35	0.31	0.45	0.17	0.09	0.17	0.10	0.12	0.15	0.17
5	0.19	0.33	0.27	0.23	0.28	0.15	0.09	0.29	0.23	0.16	0.27	0.18
IRI												
1	0.36	0.32	0.36	1.03	1.04	0.45	0.25	0.26	0.25	0.28	0.15	0.24
2	0.24	0.41	0.59	0.29	0.13	0.25	0.22	0.51	0.14	0.20	0.39	0.14
3	0.07	0.17	0.20	0.13	0.20	0.31	0.27	0.25	0.21	0.13	0.42	0.26
4	0.39	0.49	0.44	0.20	0.47	0.16	0.16	0.29	0.08	0.17	0.22	0.30
5	0.11	0.35	0.33	0.17	0.43	0.14	0.11	0.67	0.33	0.13	0.48	0.17
Segment 3							Segment 4					
SE												
1	0.11	0.14	0.48	0.11	0.10	0.25	0.22	0.25	0.43	0.40	0.36	0.42
2	0.18	0.25	0.39	0.04	0.20	0.13	0.18	0.08	0.19	0.36	0.21	0.11
3	0.11	0.10	0.29	0.14	0.04	0.45	0.66	0.46	0.85	0.29	0.17	0.08
4	0.29	0.17	0.15	0.11	0.11	0.15	0.39	0.36	0.59	1.11	0.51	0.37
5	0.30	0.07	0.23	0.15	0.21	0.24	0.67	0.60	0.82	1.60	1.37	1.21
BBI												
1	0.17	0.09	0.36	0.10	0.06	0.12	0.27	0.31	0.47	0.30	0.39	0.43
2	0.22	0.17	0.31	0.07	0.15	0.09	0.13	0.14	0.22	0.25	0.24	0.16
3	0.14	0.10	0.20	0.11	0.04	0.34	1.14	1.00	1.46	0.31	0.16	0.12
4	0.22	0.11	0.15	0.11	0.11	0.15	0.69	0.65	0.88	0.71	0.84	0.70
5	0.38	0.07	0.44	0.15	0.14	0.20	0.80	0.75	1.05	1.56	2.08	1.85
IRI												
1	0.12	0.14	0.62	0.14	0.18	0.37	0.26	0.32	0.61	0.66	0.35	0.49
2	0.17	0.33	0.53	0.05	0.29	0.14	0.20	0.11	0.21	0.42	0.20	0.12
3	0.15	0.13	0.49	0.14	0.06	0.59	0.56	0.31	0.71	0.24	0.16	0.08
4	0.36	0.22	0.19	0.15	0.14	0.16	0.42	0.33	0.70	1.40	0.42	0.32
5	0.41	0.09	0.20	0.16	0.28	0.32	0.59	0.57	0.68	1.74	1.41	1.29

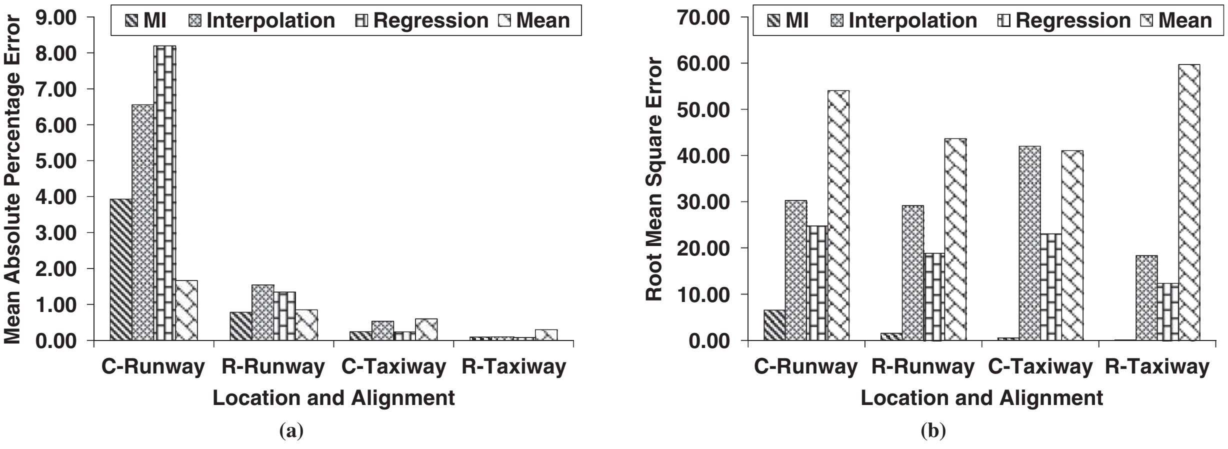


FIGURE 3 Appraisal of accuracy of various methods for imputing airport pavement roughness missing data: (*a*) MAPE of imputed Segment 1 roughness data and (*b*) RMSE of imputed Segment 1 roughness data.

(continued on next page)

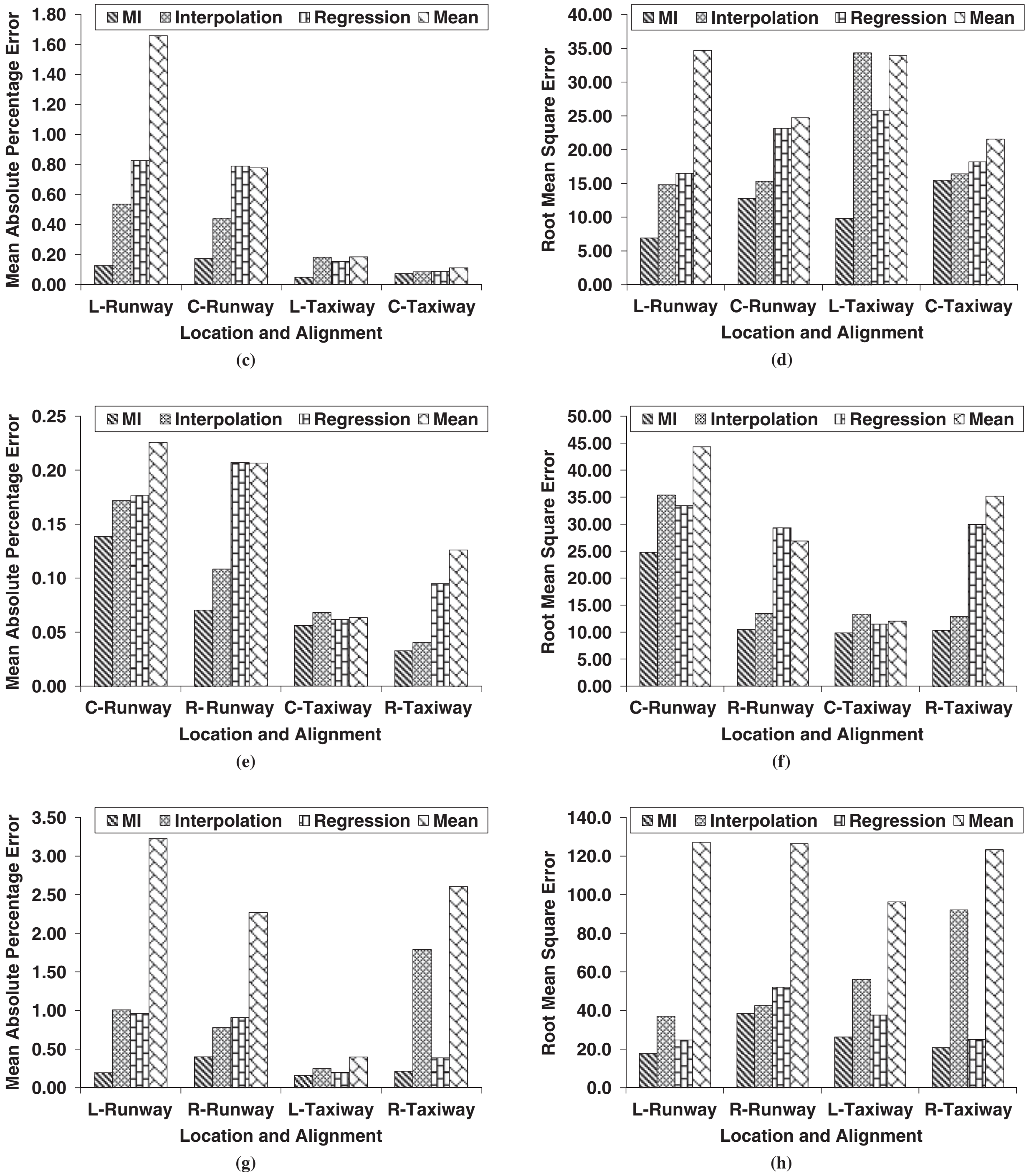


FIGURE 3 *(continued)* Appraisal of accuracy of various methods for imputing airport pavement roughness missing data: (*c*) MAPE of imputed Segment 2 roughness data, (*d*) RMSE of imputed Segment 2 roughness data, (*e*) MAPE of imputed Segment 3 roughness data, (*f*) RMSE of imputed Segment 3 roughness data, (*g*) MAPE of imputed Segment 4 roughness data, and (*h*) RMSE of imputed Segment 4 roughness data.

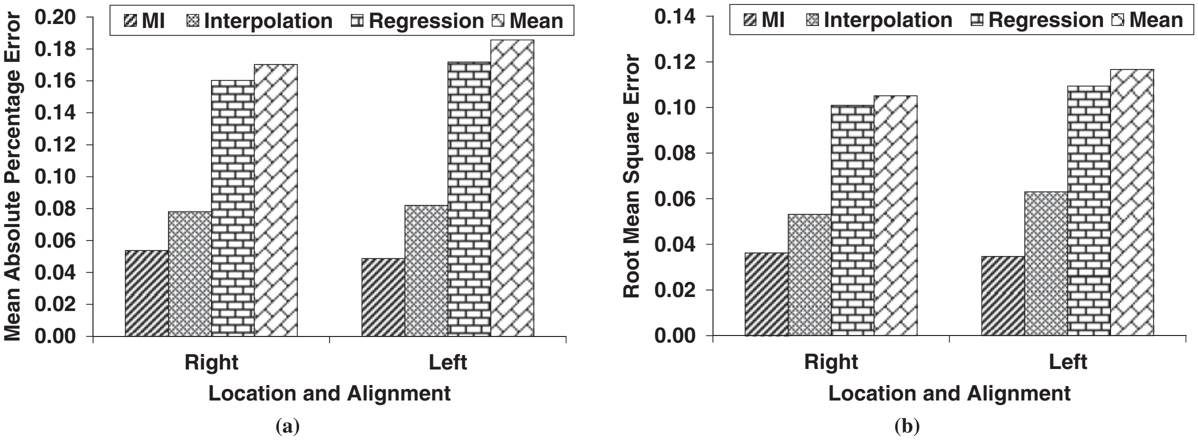


FIGURE 4 Appraisal of accuracy of various methods for imputing airport pavement friction missing data: (*a*) MAPE of imputed friction data and (*b*) RMSE of imputed friction data.

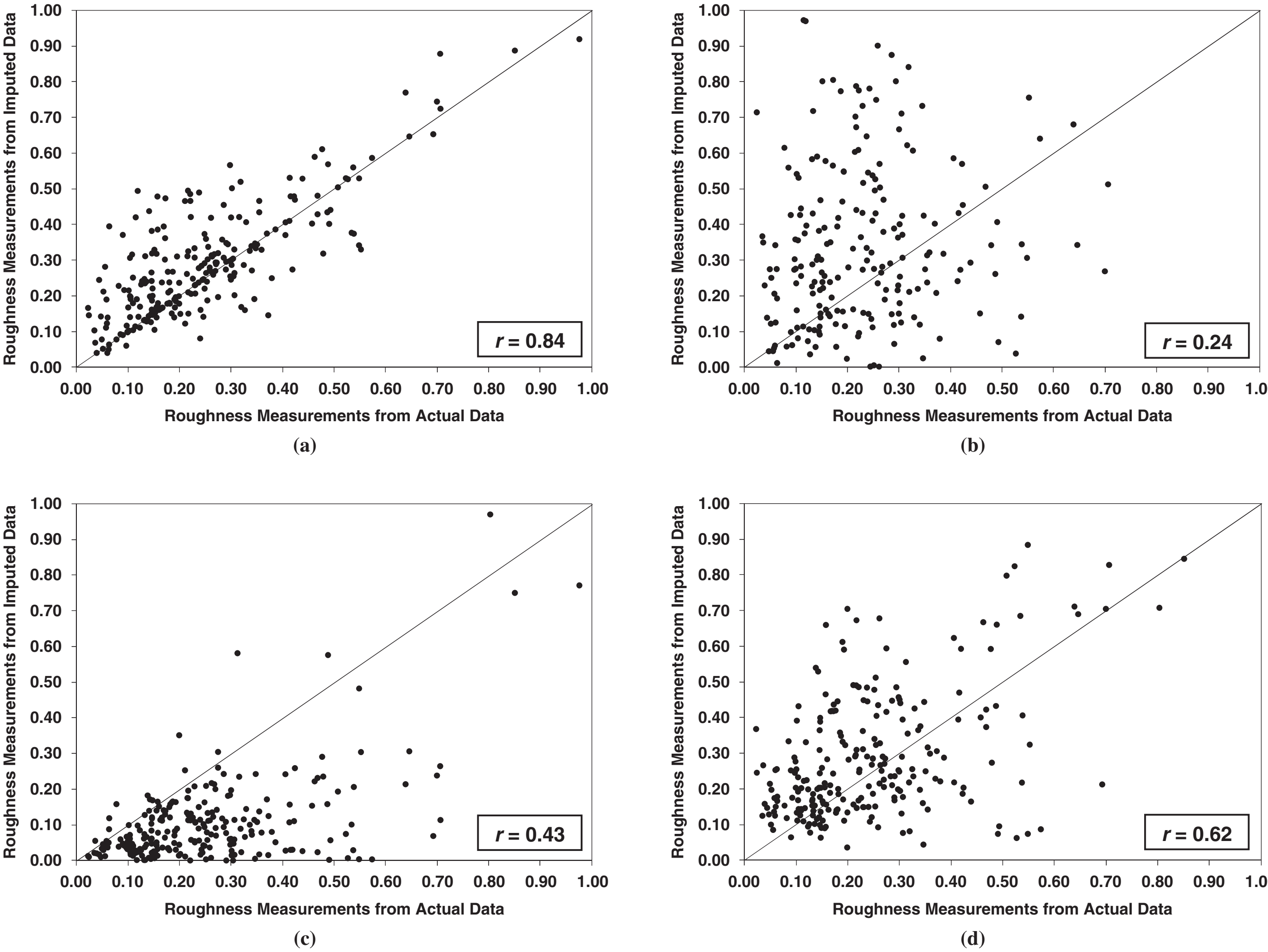


FIGURE 5 Scatter plots between SE, BBI, and IRI obtained via imputed and actual data: (*a*) substitution by multiple imputation, (*b*) substitution by mean, (*c*) substitution by interpolation, and (*d*) substitution by regression.

The degree of correlation between indexes obtained by imputing missing data with the proposed MI approach and by actual data is 0.84. The correlation between indexes obtained through actual and imputed data by using substitution by mean, interpolation, and regression is 0.24, 0.43, and 0.62, respectively. The correlation results are consistent with the MAPE and RMSE results in the preceding section, signifying the superior and robust performance of the proposed MI approach for missing data imputation of airport pavement performance data.

CONCLUSIONS

This paper has appraised four different data imputation methods for handling missing data in airport pavement condition databases. The four methods include a proposed stochastic MI method and three existing methods, namely, the substitution by mean, substitution by interpolation, and substitution by regression. For illustration, roughness and friction data of a 4-km-long runway pavement and roughness data of a 4-km-long taxiway were considered. Data were randomly deleted to create different patterns of data sets with missing data. The applicability and relative quality of the proposed approach in handling missing data were analyzed in comparison with the three existing imputation techniques. The effectiveness of using auxiliary variables in the runway and taxiway performance data imputation models is also demonstrated. The proposed stochastic MI method yielded the smallest errors for the roughness as well as for friction data. The mean substitution method resulted in imputed values with the highest amount of deviation from the observed values, followed by the regression substitution method and the interpolation method. Therefore, it is concluded that the proposed stochastic MI method outperformed the conventional methods in handling missing runway and taxiway pavement roughness and friction data and provides an effective approach to impute missing data required in an airport pavement management system.

REFERENCES

1. *Guidelines and Procedures for Maintenance of Airport Pavements.* FAA, Dec. 3, 1982.
2. Amado, V., and K. L. S. Bernhardt. Knowledge Discovery in Pavement Condition Data. Presented at 81st Annual Meeting of the Transportation Research Board, Washington D.C., 2002.
3. *NCHRP Synthesis of Highway Practices 401: Quality Management of Pavement Condition Data Collection.* TRB, National Research Council, Washington, D.C., 2009.
4. Lindly, J. K., F. Bell, and U. Sharif. Specifying Automated Pavement Condition Surveys. *Journal of the Transportation Research Forum,* Vol. 44, No. 3, 2005, pp. 19–32.
5. Schafer, J. L., and J. W. Graham. Missing Data: Our View of the State of the Art. *Psychological Methods,* Vol. 7, No. 2, 2002, pp. 147–177.
6. Little, R. J. A., and D. B. Rubin. *Statistical Analysis with Missing Data.* Wiley, New York, 1987.
7. Allison, P. D. *Missing Data: Quantitative Applications in the Social Sciences.* Sage, Thousand Oaks, Calif., 2002.
8. Yang, J., J. J. Lu, and M. Gunaratne. 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,* TRB, National Research Council, Washington, D.C., 2003, pp. 03–12.
9. Bennett, C. R. Sectioning of Road Data for Pavement. Presented at 6th International Conference on Managing Pavements, Queensland, Australia, 2004.
10. Buck, S. F. A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. *Journal of the Royal Statistical Society,* Series B 22, 1960, pp. 302–307.
11. Wasito, I. *Least Squares Algorithms with Nearest Neighbour Techniques for Imputing Missing Data Values.* PhD dissertation. University of London, 2003.
12. Roth, P. L. Missing Data: A Conceptual Review for Applied Psychologists. *Personnel Psychology,* Vol. 47, 1994, pp. 537–570.
13. Bennett, D. A. How Can I Deal with Missing Data in My Study? *Australian and New Zealand Journal of Public Health,* Vol. 25, 2001, pp. 464–469.
14. Marwala, T. *Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques,* 1st ed. IGI Global, New York, 2009, pp. 09.
15. Collins, L. M., J. L. Schafer, and C. Kam. A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods,* Vol. 6, 2001, pp. 330–351.
16. Acock, A. C. Working with Missing Values. *Journal of Marriage and Family,* Vol. 67, 2005, pp. 1012–1028.
17. Rubin, D. B. *Multiple Imputation for Survey Nonresponse.* John Wiley, New York, 1987.
18. Tanner, M. A., and W. H. Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of American Statistical Association,* Vol. 82, 1987, pp. 528–550.
19. Schafer, J. L. *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London, 1997.
20. Dempster, A. P., N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society,* Vol. 39, No. 1, 1977, pp. 01–38.
21. Ripley, B. D. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, United Kingdom, 1996.
22. 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.
23. Schafer, J. L., and M. K. Olsen. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research,* Vol. 33, 1998, pp. 545–571.
24. Schafer, J. L., and D. B. Rubin. Multiple Imputation for Missing Data Problems. Presented at Joint Statistical Meetings, Dallas, Tex., Aug. 1998.
25. Rapol, J. Evaluation of Grade and Straightedge Tolerances in Federal Aviation Administration Pavement Construction Specifications. Presented at Federal Aviation Administration Airport Technology Transfer Conference, FAA Airport Technology R&D Branch, Atlantic City, N.J., 2002.
26. *Runway Roughness Measurement, Quantification, and Application—The Boeing Method.* Document # D6-81746. Boeing Commercial Airplane Group, USA, Nov. 1995.
27. Hill, T., and P. Lewicki. *Statistics: Methods and Applications.* Statsoft, Inc., Tulsa, Okla., 2006, pp. 652.
28. Armstrong, J. S., and F. Collopy. Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons. *International Journal of Forecasting,* Vol. 8, 1992, pp. 69–80.
29. Neter, J., W. Wasserman, and M. H. Kutner. *Applied Linear Models: Regression, Analysis of Variance, and Experimental Designs.* Richard D. Irwin, Inc., Homewood, Ill., 1990, pp. 38–44, 62–104.

The Aircraft–Airport Compatibility Committee peer-reviewed this paper.