PART 2

Pavement Management Systems

Effect of Proportion of Missing Data on Application of Data Imputation in Pavement Management Systems

Javed Farhan, Bagus H. Setiadji, and Tien Fang Fwa

Instances of missing data are common in pavement condition–performance databases. A common practice today is to apply statistical imputation methods to replace the missing data with imputed values. Pavement management decision makers must know the uncertainty and errors involved in the use of data sets with imputed values in their analysis. Equally important information of practical significance is the maximum allowable proportion of missing data (i.e., the level of missing data) that can still produce results with an acceptable magnitude of error or risk when the imputed data are used. This paper proposes a procedure for determining such useful information. A numerical example analyzing pavement roughness data is presented to demonstrate the procedure through evaluating the error and reliability characteristics of imputed data. The roughness data of three road sections were obtained from the Long-Term Pavement Performance database. From these data records, data sets with different proportions of missing data were randomly generated to study the effect of level of missing data. The analysis shows that the errors of imputed data tend to increase with the level of missing data and that their magnitudes are significantly influenced by the effect of pavement rehabilitation. On the application of data imputation in pavement management systems, the study suggests that, at a 95% confidence level, 25% of missing data appears to be a reasonable allowable maximum limit for analyzing time series data on pavement roughness that include no rehabilitation within the analysis period. When pavement rehabilitation occurs within the analysis period, the maximum proportion of imputed data should be limited to 15%.

Engineering analysis and decision making in a pavement management system (PMS) are data-driven processes heavily dependent on the quality and accuracy of data records. Unfortunately, in practice, the records in most PMSs are missing data (*1–3*). Therefore, management of missing data is an important element in the engineering analysis and decision making of a PMS. An NCHRP Synthesis Report reports that 61% of the pavement agencies in United States included in a survey used a software routine to check for missing data (*4*).

Because pavement condition and performance data are time-specific information, re-collection of missing past records through a field survey is not possible. In this situation, the PMS engineer has the option of discarding the records having missing data and proceeding with the remaining records. This option is not always desirable, as it means that the engineering analysis proceeds with a reduced data space and ignores some recorded data that could have important implications for pavement maintenance or traffic operations. Procedures that are increasingly being adopted today include applying suitable data imputation techniques to supply the incomplete records with imputed values and performing engineering analysis without discarding those records (*4–6*).

In applying data imputation methods to manage missing data records in PMS, one must be aware that the techniques are statistical and that uncertainties are involved in the imputed data values. Knowing the likely magnitudes of the errors involved and the reliability of the data set containing imputed data would allow engineers to make informed decisions on whether to discard the incomplete data records or to proceed with the full set of records made complete with imputed data. Therefore, a relevant issue is to determine the upper limit of the proportion of missing data at which filling the incomplete data records with imputed data would still provide an accurate representation of the pavement condition. This issue is the focus of the present research. By using pavement roughness data from the Long-Term Pavement Performance (LTPP) database, this study examines the ways in which different proportions of missing data would affect the accuracy and reliability of imputed data sets.

SIGNIFICANCE

The theory and principle of statistical quality assurance in regard to the imputation of missing data are well developed and have been applied by researchers and practitioners in various fields of study, notably in the disciplines of medicine and social sciences (*7–10*). The issue of the upper limit of missing data for the application of data imputation has also been addressed by researchers in those disciplines. For instance, Schafer suggested using statistical data imputation approaches in medical research only when less than 5% of the data is missing (*11*). In contrast, in dealing with missing data in public health studies, Peng et al. recommended 20% missing data as the maximum limit for educational research (*12*). However, in their studies of palliative and end-of-life care, Preston et al. recommended that high rates of attrition or missing data are tolerable and that it is more important to design a clear statistical analysis plan to account for missing data and attrition (*13*).

Little and Rubin introduced the concept of missing data to highlight the importance of the influence of the pattern of missing data

J. Farhan and T. F. Fwa, Department of Civil and Environmental Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. B. H. Setiadji, Department of Civil Engineering, Diponegoro University, Kota Semarang, Jawa Tengab 50277, Indonesia. Corresponding author: T. F. Fwa, ceefwatf@nus.edu.sg.

Transportation Research Record: Journal of the Transportation Research Board, No. 2523, Transportation Research Board, Washington, D.C., 2015, pp. 21–31.
DOI: 10.3141/2523-03

on (*a*) the overall bias introduced and (*b*) the proportion of missing data that is too high for creating a reasonably "complete" data set (*14*). For example, in the case for which a very high proportion of data (much higher than 20%) was "missing completely at random," one could still re-create the data set with imputed data and capture the essential characteristics of the original data records. Schlomer et al. concurred that the pattern of missing data is a major factor of consideration but stressed that, regardless of the level of missing data, one must determine whether the resultant imputed data set has adequate statistical power to detect the effects of interest (*15*).

Research in various disciplines has made clear that no simple guidelines can be set for the maximum allowable proportion of missing data universally covering all fields of study. The effect of the proportion of missing data on the quality of analysis when imputed data sets are used depends on the nature and characteristics of the data as well as on the pattern of missing data, and statistical analyses must be performed by the decision maker to manage the missing data and use them properly.

To the authors' knowledge, the use of imputed data sets in pavement management analysis, their impact on data quality and reliability, and the possible bias they introduce to the analysis have not been studied in the field of pavement management. No guidelines are available about the data management procedure necessary to deal with data sets that contain different extents of missing data. As missing data are commonly encountered in records of pavement management data, the availability of the aforementioned information related to the use of imputed data would have high practical significance. This paper attempts to provide some information to bridge this knowledge gap, in part, by analyzing the effects of missing data in pavement roughness records.

APPROACH AND METHODOLOGY

Scope

The common types of pavement condition and performance data that are regularly collected in a typical pavement management system include pavement distress data (such as cracks, ruts, potholes, and depressions), roughness, friction, and structural condition. Because the nature and characteristics of each of these types of data are quite different from one another, they will likely be affected by missing data differently. A major research effort would be required to examine the effects of missing data on all types of pavement condition–performance data.

The scope of the present study is limited to the analysis of the effect of missing data in pavement roughness records. The framework and concept of the proposed analysis will be described in this section and followed by a demonstration that uses an example involving actual records of pavement roughness data. Analysis of the numerical example demonstrates that useful informative insight can be gained into the quality of imputed data obtained, the magnitude of errors involved as the proportion of missing data increases, and the statistical reliability implications of the imputed data set as a function of the proportion of missing data.

Framework of Analysis

For the purpose of studying the error and reliability characteristics of imputed data, complete records of pavement roughness data without any missing data were first obtained. These full records of actual measured roughness data will serve as the base reference for assessing the quality and reliability characteristics of data sets containing imputed data. The data sets containing missing data are artificially generated randomly from the original complete data records.

The proposed analysis consists of the following steps:

1. Selection of complete data records. The FHWA's LTPP database offers a convenient source for the selection of pavement roughness data records for the present study (*16*). The roughness data are reported as units of the international roughness index (IRI).
2. Creation of data sets having different levels of missing data and different patterns of missing data. To study the effect of the level of missing data (i.e., proportion of missing data), at least six equally spaced levels of missing data were first identified. Next, for each specified level of data missingness, a random process following the missing-completely-at-random rules was employed to generate a data set containing the correct (say, *n*) number of missing items of data by randomly deleting *n* data points from the original complete data records. This random deletion process is repeated nine times to produce 10 randomly generated data sets, each with a different pattern of missing data, for each of the six or more levels of missing data studied.
3. Computation of imputed values for each data set containing missing data. For each of the data sets containing missing data generated in Step 2, a suitable data imputation method is applied to compute a data value for each of the missing items of data. At the end of this step, all the data sets with missing data generated in Step 2 are transformed into data sets containing imputed data values (i.e., 10 data sets containing imputed data for each level of missing data). The technique of multiple imputation (MI) was adopted for computing imputed data in this study. The imputed value for each missing data item in each of the 10 generated data sets is obtained as the mean value of 10 imputation runs. The concept and procedure of computation of the MI technique are explained in the next section.
4. Performance of error and reliability analysis. By using the original complete data records as the base reference, the errors of the imputed data can be computed and analyzed. The variation of the errors with the level of missing data can be examined. The statistical reliability of the imputed data sets at different levels of missing data is established by means of hypothesis testing.

MI Technique for Data Imputation

The most widely used method today in performing data imputation for missing data is the MI technique introduced by Rubin (*17*). This method is known to produce unbiased imputed data and parameter estimates (*14, 17, 18*). The current authors demonstrated in earlier work that the MI method outperformed conventional methods (such as the deletion method and the substitution methods using mean, interpolation, or regression) in handling missing pavement condition–performance data (*19*). The MI process performs better than those other methods because it iteratively estimates statistical parameters from existing and estimated data to improve the prediction of missing data values. It produces more reliable estimates than determination of missing data by interpolation or regression.

The MI process consists of three main phases: imputation, analysis, and pooling. In the imputation phase, the available measured

data are used to estimate distribution parameters, which in turn are used to estimate the missing data values. In the analysis phase, each imputed value is analyzed together with the available values by means of a statistical procedure to produce new imputed estimates and—subsequently—missing data values. This process alternates between simulating missing data and parameters until convergence. Repetition of this procedure generates multiple imputations of the missing values. Finally, in the pooling phase, the multiple imputation results are integrated into a single set of results to produce overall estimates and standard errors that reflect missing-data uncertainty. These combined standard errors are useful for statistical significance testing and for drawing of inferential conclusions.

The working of the MI method makes use of two main algorithms, namely expectation maximization (EM) and data augmentation (DA). The procedure of data imputation adopted in this study involves the following steps:

Step 1. Data transformation. First, the data must be transformed to approximately normal before imputation by using transformation functions, such as logit, log, or square root functions. After imputation, the data will be retransformed to their original scale.

Step 2. Imputation using EM. EM uses the maximum likelihood approach to perform the imputation function in the imputation and analysis phases of the MI procedure. This step generates estimates of missing values for the data matrix with the convergence criterion that the maximum relative change in the value of any parameter during the iterative process is less than 10^{-6}.

Step 3. Imputation using DA. With the initial parameter estimates from the EM algorithm serving as the basis, the DA algorithm performs multiple imputations, as explained in Step 2 of the MI procedure. The commonly adopted practice of 10 imputations (*14, 20*) is applied in this study.

Step 4. Synthesis of estimates. The synthesis of estimates averages the multiple estimates of the multiple imputation analysis to obtain the final set of estimates (*17*).

Calculations in the MI procedure used the software NORM (*21*).

TABLE 1 Observed IRI Values of Road Sections Studied

SHRP ID	State	Year	Date of IRI Measurement	IRI (m/km)
28-1802	Mississippi	1	Aug. 1990	0.895
		2	May 1991	1.011
		3	Aug. 1992	1.163
		4	Jan. 1993	1.251
		5	Aug. 1994	1.722
		6	July 1995	2.187
		7	April 1996	2.142
		8	Oct. 1997	1.991
20-1005	Kansas	1	May 1992	2.933
		2	March 1993	2.911
		3	May 1994	2.833
		4	March 1995	2.964
		5	April 1996	2.948
		6	Feb. 1997	3.164
		7	April 1998	3.369
		8	March 1999	3.408
		9	Feb. 2000	3.448
		10	May 2001	1.177
25-1002	Massachusetts	1	Oct. 1989	1.164
		2	Sept. 1990	1.196
		3	July 1991	1.189
		4	Sept. 1992	1.132
		5	Sept. 1993	1.186
		6	Jan. 1994	1.408
		7	Jan. 1995	1.607
		8	Nov. 1996	2.198
		9	June 1997	3.387
		10	June 1998	2.947
		11	July 1999	1.451
		12	June 2000	2.791
		13	April 2001	2.844
		14	Feb. 2002	3.014
		15	Sept. 2003	3.245
		16	April 2004	2.943

NOTE: Aug. = August; Jan. = January; Oct. = October; Feb. = February; Sept. = September; Nov. = November.

ILLUSTRATIVE EXAMPLE: IMPUTATION OF ROUGHNESS DATA

IRI Records in LTPP Database

From the LTPP database that provides measured records of pavement roughness data covering the years from 1989 to 2012, the following three records were extracted for the illustrative analysis of this study (*16*):

- Road Section SHRP ID 28-1802 with 8 years of continuous measured annual IRI data,
- Road Section SHRP ID 20-1005 with 10 years of continuous measured annual IRI data, and
- Road Section SHRP ID 25-1002 with 16 years of continuous measured annual IRI data.

Table 1 shows the measured IRI values and the times of measurements of the IRI records of the three road sections. These IRI data are plotted in Figure 1. Although the annual IRI measurements were not gathered at time intervals of exactly 12 months, they can be considered as time series data for analysis and illustration purposes of the present example to study the effects of missing data.

The three road sections have been selected because their pavement roughness variation trends display distinctly different patterns. Road Sections SHRP ID 28-1802 and ID 20-1005 both had roughness values gradually increasing with time, except that the latter had a sharp drop in roughness value in the last year of the record. The roughness variations of Road Section ID 25-1002 were characterized by two periods of gentle increases (from Years 1 to 7 and from Years 12 to 15), two periods of sharp rises (from Years 7 to 9 and from Years 12 to 15), a period of sharp fall (from Years 9 to 11), and a mild drop in Year 16.

Data Representation

Pavement roughness is expected to increase with the number of years of service because of the impact of traffic loading. However, on the occasions of pavement resurfacing or rehabilitation, roughness would be restored to a lower value. Such maintenance and rehabilitation (M&R) activities are common in road operations. They occurred for all three road sections considered here. As indicated in the LTPP

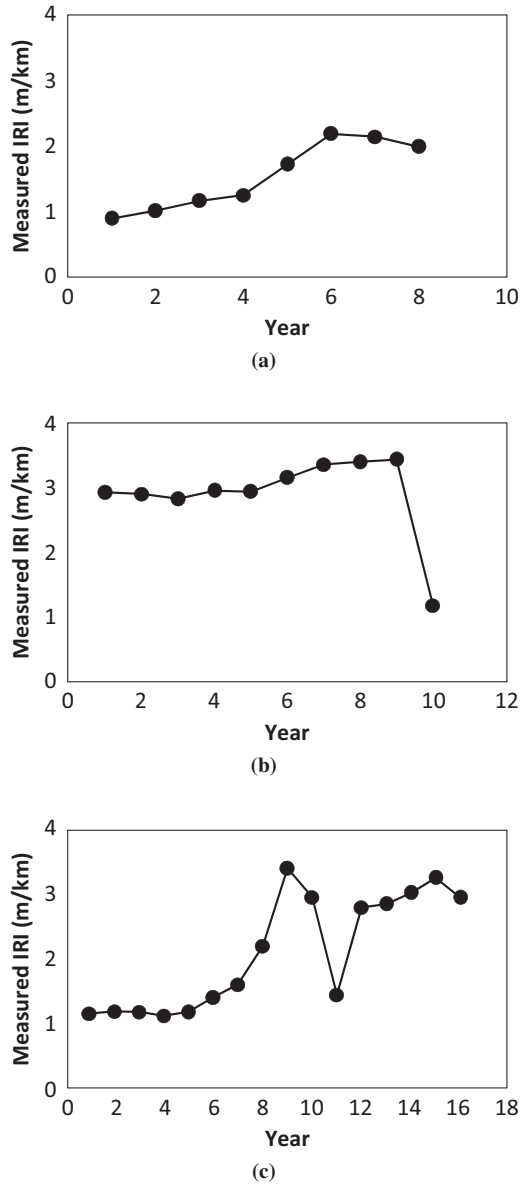

FIGURE 1 Measured IRI data for road sections studied: (*a*) SHRP ID 28-1802, (*b*) SHRP ID 20-1005, and (*c*) SHRP ID 25-1002.

database, for Road Section SHRP ID 28-1802, minor M&R occurred in Years 7 and 8 and resulted in slight decreases in the IRI value. For Road Section ID 20-1005, a minor and a major M&R were performed in Years 5 and 10, respectively. For Road Section SHRP ID 25-1002, the database records indicated a major and a minor M&R in Years 10 and 16, respectively.

In the data imputation analysis, this situation was handled by introducing an M&R dummy variable that would be assigned a value of 1 if an M&R operation occurred in the year of interest and 0 otherwise. Figure 1 shows, for Road Section SHRP ID 25-1002, that, although the LTPP database indicated an M&R operation in Year 11, a drop in the IRI value began in Year 10. The authors suspect that the M&R might have commenced in Year 10 and resulted in the fall of IRI.

Generation of Data Sets with Missing Data

To study the effect of the proportion of missing data and to determine the maximum allowable proportion of missing data, data sets with proportions of missing data ranging from approximately 10% to 90% were created for the three road sections studied. These data sets with missing data were randomly generated from the respective original complete data records. The levels of missing data created for the three road sections studied are as follows:

- SHRP ID 28-1802. Six levels of missing data were created. The percentages of missing data created were 12.5%, 25%, 37.5%, 50%, 62.5%, and 75%.
- SHRP ID 20-1005. Eight levels of missing data were created. The percentages of missing data created were 10%, 20%, 30%, 40%, 50%, 60%, 70%, and 80%.
- SHRP ID 25-1002. Seven levels of missing data were created. The percentages of missing data created were 12.5%, 25%, 37.5%, 50%, 62.5%, 75% and 87.5%.

For each of the three road sections, at each level of missing data, 10 patterns of missing data were randomly created. Figures 2, 3, and 4 show all the patterns of missing data created for Road Sections SHRP ID 28-1802, SHRP ID 20-1005, and SHRP ID 25-1002, respectively.

Analysis of Imputation Results

Error Analysis

Error analysis involves examining the differences between the imputed data and the corresponding actual data values of the original complete data records. As explained earlier, for each roughness record for a road section that is analyzed, 10 patterns of missing data were created for each level of missing data (Figures 2 to 4). And, for each pattern of missing data for a given level of missing data, 10 imputation runs were made by means of the MI technique. Hence, 10 values were imputed for each item of missing data from the 10 imputation runs, and the error of each imputed roughness value was defined as its deviation from the actual roughness value of the original complete data record.

Figure 5 presents three examples of the mean and range of errors of the imputed values against the levels of missing data (i.e., proportions of missing data) for the three road sections studied. Figure 5*a* shows the results of imputation errors for the data sets of the level of missing data with 25% missing data for the roughness data of Road Section SHRP ID 28-1802. At 25%, two items of data were missing per data set (i.e., two items of missing data per pattern of missing data, Figure 2). Figure 5*a* shows two sets of error results for each of the 10 patterns of missing data. Similarly, Figure 5*b* for the roughness data of Road Section SHRP ID 20-1005 shows two sets of error results for each of the 10 patterns of missing data at the 20% level of missing data. The roughness data for Road Section SHRP ID 25-1002 have four sets of error results for each of the 10 patterns of missing data at the level of missing data of 25% as shown in Figure 5*c*.

From the three plots of the errors of the imputed data values shown in Figure 5, the following comments can be made:

1. Road Section SHRP ID 28-1802 has no clear trends of variation among the errors for the 10 patterns. This result is within expectation because the imputed data values were generated through a random process.

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	1	1	0	1	0	0
2	0	1	1	0	1	0
3	1	1	0	1	0	1
4	1	1	1	0	0	1
5	1	0	1	1	1	0
6	1	0	1	0	0	0
7	1	1	0	1	0	0
8	1	1	1	0	1	0

(a)

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	1	1	1.01	0	1	0
2	1	1	0	1	0	0
3	1	1	1	0	0	1
4	1	1	1	0	1	0
5	1	1	0	1	0	1
6	0	1	1	1	0	0
7	1	0	0	1	0	0
8	1	0	1	0	1	0

(b)

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	1	0	1	0	1	0
2	1	1	1	0	0	1
3	0	1	0	1	0	0
4	1	0	1	0	1	0
5	1	1	0	1	0	0
6	1	1	1	0	0	1
7	1	1	1	1	0	0
8	1	1	0	1	1	0

(c)

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	1	0	0	1	0	1
2	1	1	0	1	0	0
3	1	0	1	0	1	0
4	1	1	1	1	0	0
5	1	1	0	0	0	1
6	1	1	1	1	0	0
7	0	1	1	0	1	0
8	1	1	1	0	1	0

(d)

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	0	1	0	0	1	0
2	1	1	0	1	0	0
3	1	0	1	0	1	0
4	1	1	1	1	0	1
5	1	1	0	1	0	0
6	1	1	1	0	1	0
7	1	0	1	1	0	0
8	1	1	1	0	0	1

(e)

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	1	1	0	1	0	0
2	1	1	1	0	0	1
3	1	1	0	1	0	0
4	1	0	1	0	1	0
5	1	1	1	0	1	0
6	1	1	0	1	1	0
7	1	0	1	0	0	1
8	0	1	1	1	0	0

(f)

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	1	1	0	1	0	1
2	1	1	1	0	1	0
3	1	0	1	0	1	0
4	0	1	1	0	0	1
5	1	0	1	1	0	0
6	1	1	0	1	0	0
7	1	1	1	1	0	0
8	1	1	0	0	1	0

(g)

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	1	1	1	0	1	0
2	1	0	1	0	0	1
3	1	1	0	1	0	0
4	1	1	1	0	0	1
5	0	1	1	1	0	0
6	1	0	1	0	1	0
7	1	1	0	1	0	0
8	1	1	0	1	1	0

(h)

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	1	0	0	1	0	0
2	1	1	1	0	1	1
3	1	1	0	0	1	0
4	0	1	1	1	0	0
5	1	1	0	1	0	0
6	1	1	1	0	1	0
7	1	0	1	0	0	1
8	1	1	1	1	0	0

(i)

Year	Percentage of Missing Data					
	12.5	25.0	37.5	50.0	62.5	75.0
1	1	1	1	0	0	1
2	1	1	0	1	0	0
3	1	0	1	0	0	0
4	1	1	1	0	1	0
5	1	1	0	1	1	0
6	1	0	1	1	1	0
7	1	1	0	1	0	0
8	0	1	1	0	0	1

(j)

FIGURE 2 Patterns of missing IRI data created for Road Section SHRP ID 28-1802: (*a*) Pattern 1, (*b*) Pattern 2, (*c*) Pattern 3, (*d*) Pattern 4, (*e*) Pattern 5, (*f*) Pattern 6, (*g*) Pattern 7, (*h*) Pattern 8, (*i*) Pattern 9, and (*j*) Pattern 10 (0 = IRI missing data).

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	1	1	0	1	1	1	1	0
2	1	1	0	0	0	1	0	1
3	1	1	1	0	1	0	0	0
4	1	1	1	0	0	0	1	0
5	0	0	1	1	0	1	0	0
6	1	1	0	1	0	1	0	1
7	1	1	1	0	1	0	0	0
8	1	1	1	1	0	0	0	0
9	1	1	1	1	1	0	1	0
10	1	0	1	1	1	0	0	0

(a)

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	1	1	1	1	0	1	0	0
2	1	0	1	1	0	0	0	0
3	1	1	0	1	1	1	0	0
4	1	1	1	0	0	0	1	0
5	1	1	0	1	1	0	1	0
6	1	1	1	1	0	0	0	0
7	1	1	1	1	1	0	0	0
8	0	1	0	0	1	0	1	1
9	1	0	1	0	1	1	0	0
10	1	1	1	0	0	1	0	1

(b)

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	0	0	1	1	1	0	1	0
2	1	1	0	1	1	0	0	0
3	1	1	0	0	1	0	1	0
4	1	1	1	1	0	1	0	1
5	1	1	1	0	1	0	1	0
6	1	0	1	1	1	0	0	0
7	1	1	0	0	0	1	0	0
8	1	1	1	1	0	1	0	0
9	1	1	1	0	0	0	0	1
10	1	1	1	1	0	1	0	0

(c)

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	1	0	0	0	1	1	0	0
2	1	1	1	1	1	0	0	0
3	0	1	1	0	1	0	0	1
4	1	1	1	1	0	1	0	0
5	1	1	1	1	1	0	0	0
6	1	0	0	0	1	0	1	0
7	1	1	0	0	0	1	0	0
8	1	1	1	1	0	0	1	0
9	1	1	1	1	0	1	0	1
10	1	1	1	1	0	0	1	0

(d)

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	1	1	0	0	0	1	0	1
2	1	1	0	0	1	0	0	0
3	1	0	1	1	0	1	1	1
4	1	1	1	1	1	0	0	0
5	1	1	1	1	0	1	0	0
6	1	1	1	0	1	1	0	0
7	1	1	1	1	0	0	1	0
8	1	0	1	1	1	0	0	0
9	1	1	0	1	0	0	1	0
10	0	1	1	0	1	0	0	0

(e)

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	1	1	1	0	1	0	0	0
2	1	1	1	0	0	0	1	0
3	1	1	1	1	0	1	0	0
4	1	1	1	1	0	0	0	0
5	1	1	0	1	0	1	0	0
6	0	1	1	0	1	0	1	1
7	1	1	1	1	0	1	1	0
8	1	0	0	1	1	0	0	1
9	1	1	1	0	1	1	0	0
10	1	0	0	1	1	0	0	0

(f)

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	1	1	1	1	1	0	0	0
2	1	0	1	1	1	1	1	0
3	1	1	1	1	1	0	0	0
4	1	1	0	0	0	1	0	0
5	1	1	1	0	1	0	1	0
6	1	1	1	0	1	0	1	0
7	0	0	1	1	0	0	0	1
8	1	1	1	0	0	1	0	0
9	1	1	0	1	0	0	0	0
10	1	1	0	1	0	1	0	1

(g)

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	1	1	1	1	0	0	1	0
2	0	1	1	0	0	1	0	0
3	1	1	0	0	0	0	1	0
4	1	0	1	1	1	1	0	1
5	1	0	1	0	1	1	0	1
6	1	1	1	1	0	0	0	0
7	1	1	0	1	1	0	0	0
8	1	1	1	0	1	1	0	0
9	1	1	0	1	0	0	1	0
10	1	1	1	1	1	0	0	0

(h)

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	1	1	1	0	0	0	0	1
2	1	1	1	1	1	1	0	0
3	1	0	1	1	1	0	0	0
4	0	1	0	0	1	0	0	0
5	1	1	0	1	0	0	0	0
6	1	1	0	1	1	1	0	0
7	1	0	1	0	1	0	0	1
8	1	1	1	1	0	1	1	0
9	1	1	1	0	0	1	0	0
10	1	1	1	0	0	1	1	0

(i)

Year	Percentage of Missing Data							
	10	20	30	40	50	60	70	80
1	1	1	1	0	0	0	0	1
2	1	1	1	1	1	1	0	0
3	1	1	1	1	0	1	0	0
4	1	0	0	1	1	0	1	0
5	1	1	1	0	1	0	0	1
6	1	1	1	1	0	1	0	0
7	1	1	1	1	0	1	1	0
8	1	1	0	0	1	0	0	0
9	0	0	1	1	1	0	0	0
10	1	1	0	0	0	0	1	0

(j)

FIGURE 3 Patterns of missing IRI data created for Road Section SHRP ID 20-1005: (*a*) Pattern 1, (*b*) Pattern 2, (*c*) Pattern 3, (*d*) Pattern 4, (*e*) Pattern 5, (*f*) Pattern 6, (*g*) Pattern 7, (*h*) Pattern 8, (*i*) Pattern 9, and (*j*) Pattern 10.

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	1	1	0	1	0	0	1
2	1	1	1	0	1	0	0
3	1	1	1	0	1	0	0
4	1	1	0	0	1	0	0
5	1	0	1	1	0	0	1
6	0	1	0	0	1	0	0
7	1	1	1	0	0	1	0
8	1	0	1	1	0	0	0
9	1	0	1	0	0	1	0
10	1	1	0	0	1	0	0
11	1	0	0	1	0	0	0
12	1	1	0	1	0	1	0
13	1	1	1	1	0	0	0
14	1	1	1	0	1	0	0
15	0	1	1	1	0	0	0
16	1	1	1	1	0	1	0

(a)

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	1	0	1	1	1	0	0
2	1	1	0	0	1	0	0
3	1	1	1	0	0	1	0
4	1	0	1	0	1	0	0
5	0	1	0	1	0	0	0
6	1	1	1	0	1	0	0
7	1	1	0	1	0	0	1
8	1	1	0	1	0	0	1
9	1	1	0	1	1	0	0
10	0	1	1	1	0	0	0
11	1	0	1	0	0	1	0
12	1	1	1	0	0	1	0
13	1	1	0	0	1	0	0
14	1	1	1	1	0	0	0
15	1	0	1	0	0	1	0
16	1	1	1	1	0	0	0

(b)

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	1	0	0	1	0	0	0
2	0	1	1	0	0	1	0
3	1	0	1	1	0	0	1
4	1	1	1	0	1	0	0
5	1	1	0	1	0	1	0
6	0	1	0	0	1	0	0
7	1	1	1	0	0	1	0
8	1	1	1	0	1	0	0
9	1	1	0	0	1	0	0
10	1	1	1	1	0	0	0
11	1	1	0	1	0	1	0
12	1	1	0	1	0	0	0
13	1	1	1	0	1	0	0
14	1	0	1	1	0	0	0
15	1	1	1	1	0	0	1
16	1	0	1	0	1	0	0

(c)

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	1	1	1	0	1	0	0
2	1	1	0	1	0	0	1
3	1	0	1	0	1	0	0
4	1	1	1	0	0	1	0
5	1	1	0	0	1	0	0
6	1	0	1	1	0	0	0
7	1	1	0	0	1	0	0
8	1	1	1	0	1	0	0
9	1	1	0	1	0	0	0
10	1	1	0	1	0	1	0
11	0	0	1	0	0	1	0
12	1	1	1	1	0	0	0
13	1	1	1	1	0	0	1
14	1	0	1	1	0	0	0
15	0	1	0	0	1	0	0
16	1	1	1	1	0	1	0

(d)

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	1	1	0	1	0	0	0
2	1	1	1	1	0	0	0
3	0	0	1	0	1	0	0
4	1	1	1	1	0	0	1
5	1	1	0	1	0	0	0
6	1	1	1	0	0	1	0
7	1	1	1	0	0	1	0
8	1	1	0	0	1	0	0
9	0	1	1	0	1	0	0
10	1	0	0	1	1	0	0
11	1	1	1	0	0	1	0
12	1	1	1	0	0	1	0
13	1	1	1	1	0	0	0
14	1	0	0	1	1	0	0
15	1	1	0	1	0	0	1
16	1	0	1	0	1	0	0

(e)

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	1	1	1	1	0	0	0
2	1	0	1	1	0	1	0
3	1	0	1	0	1	0	0
4	1	0	1	0	0	1	0
5	0	1	1	1	0	0	0
6	1	1	1	0	1	0	0
7	0	1	0	1	0	0	0
8	1	1	0	0	1	0	1
9	1	0	1	1	0	0	0
10	1	1	1	0	1	0	0
11	1	1	0	1	0	1	0
12	1	1	0	0	1	0	0
13	1	1	1	0	1	0	0
14	1	1	1	0	0	1	0
15	1	1	0	1	0	0	0
16	1	1	0	1	0	0	1

(f)

FIGURE 4 Patterns of missing IRI data created for Road Section SHRP ID 25-1002: (*a*) Pattern 1, (*b*) Pattern 2, (*c*) Pattern 3, (*d*) Pattern 4, (*e*) Pattern 5 and (*f*) Pattern 6.

(continued on next page)

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	1	1	0	0	1	0	0
2	1	1	0	1	0	0	0
3	1	1	1	0	0	1	0
4	1	1	1	1	0	0	0
5	0	0	1	0	0	0	1
6	1	0	1	1	1	0	0
7	1	0	0	1	0	0	0
8	0	1	1	0	0	1	0
9	1	1	0	1	1	0	0
10	1	1	1	0	0	1	0
11	1	1	1	1	0	0	0
12	1	0	1	0	0	1	0
13	1	1	1	1	0	0	1
14	1	1	0	0	1	0	0
15	1	1	0	0	1	0	0
16	1	1	1	1	1	0	0

(g)

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	1	0	1	0	1	0	0
2	1	1	0	1	0	0	1
3	1	1	0	1	0	0	0
4	0	1	1	0	0	1	0
5	1	1	1	0	0	1	0
6	1	1	0	0	1	0	0
7	0	1	0	1	0	0	1
8	1	1	1	1	0	0	0
9	1	0	1	0	0	1	0
10	1	1	0	1	1	0	0
11	1	0	1	1	0	0	0
12	1	1	1	0	0	1	0
13	1	0	1	1	0	0	0
14	1	1	1	1	1	0	0
15	1	1	0	0	1	0	0
16	1	1	1	0	1	0	0

(h)

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	1	1	0	1	1	0	0
2	1	1	1	0	0	1	0
3	0	1	0	1	0	0	0
4	1	0	1	0	0	1	0
5	1	1	1	0	1	0	0
6	1	0	1	0	0	1	0
7	1	1	1	0	1	0	0
8	1	1	1	1	0	0	0
9	1	1	1	0	1	0	0
10	1	1	0	1	1	0	0
11	1	0	0	1	0	0	0
12	0	1	0	1	0	0	1
13	1	1	1	1	0	0	0
14	1	0	1	0	1	0	0
15	1	1	1	0	0	1	0
16	1	1	0	1	0	0	1

(i)

Year	Percentage of Missing Data						
	12.5	25.0	37.5	50.0	62.5	75.0	87.5
1	0	0	1	0	0	1	0
2	1	1	0	1	0	0	0
3	1	1	1	0	0	1	0
4	1	0	1	1	0	0	0
5	1	1	1	0	0	1	0
6	1	1	0	1	0	0	1
7	1	0	0	1	0	0	1
8	1	1	1	0	1	0	0
9	1	0	0	1	1	0	0
10	1	1	1	1	0	0	0
11	1	1	1	0	1	0	0
12	1	1	1	1	0	0	0
13	0	1	0	0	1	0	0
14	1	1	1	0	0	1	0
15	1	1	1	0	1	0	0
16	1	1	0	1	1	0	0

(j)

FIGURE 4 *(continued)* **Patterns of missing IRI data created for Road Section SHRP ID 25-1002: (*g*) Pattern 7, (*h*) Pattern 8, (*i*) Pattern 9, and (*j*) Pattern 10.**

2. For Road Section SHRP ID 20-1005, large errors are found for one imputed mean value each for Patterns 1 and 6. These large errors occurred because both patterns contain missing data in Year 10, the year with a sudden drop in roughness value. This observation highlights that having missing data in regions of sharp changes in roughness data introduces large errors when data imputation is applied.

3. For Road Section SHRP ID 25-1002, large errors occurred for one imputed mean value each for Patterns 1, 5, 6, 8, and 10. Each of these patterns has a missing value in either Year 9 or 10. These are the 2 years with a sharp fall of the roughness value. This observation reinforces the observation in the preceding paragraph about larger imputation errors associated with sharp changes in roughness data.

Another error characteristic of interest is the ways in which errors vary with the level of missing data. Figure 6 plots the mean and range of the absolute errors of imputed values against the levels of data missingness (i.e., proportions of missing data) for all the cases of the three road sections studied. The plots appear to suggest in general that the magnitude of imputation error increases with the level of missing data. Further work to analyze more pavement section IRI data is needed to confirm this trend.

Reliability Analysis

The uncertainty involved in the imputation of missing values is reflected in the variations of the multiple imputed values for each missing data value in the example problem. Such variations are seen in the plots in Figures 5 and 6, for which the distributions in the errors of imputed values as well as the variations among the means of different imputation runs, respectively, are depicted.

With the error characteristics presented in Figures 5 and 6, an analysis of statistical reliability of the imputation results can be performed. For the purpose of the present study, a hypothesis test was performed to compare the mean computed value for each missing item of data with the corresponding actual data value of the original complete record. Because each missing item of data had 10 imputed values, Student's *t*-test was employed (*22*). The hypothesis testing considers the following null and alternative hypotheses:

Null hypothesis (H_0). The mean imputed value, which is obtained from the 10 imputation analyses (μ_z), is no different from the actual data value μ_0 from the original data record of the given road section:

H_0:

$\mu_z = \mu_0$

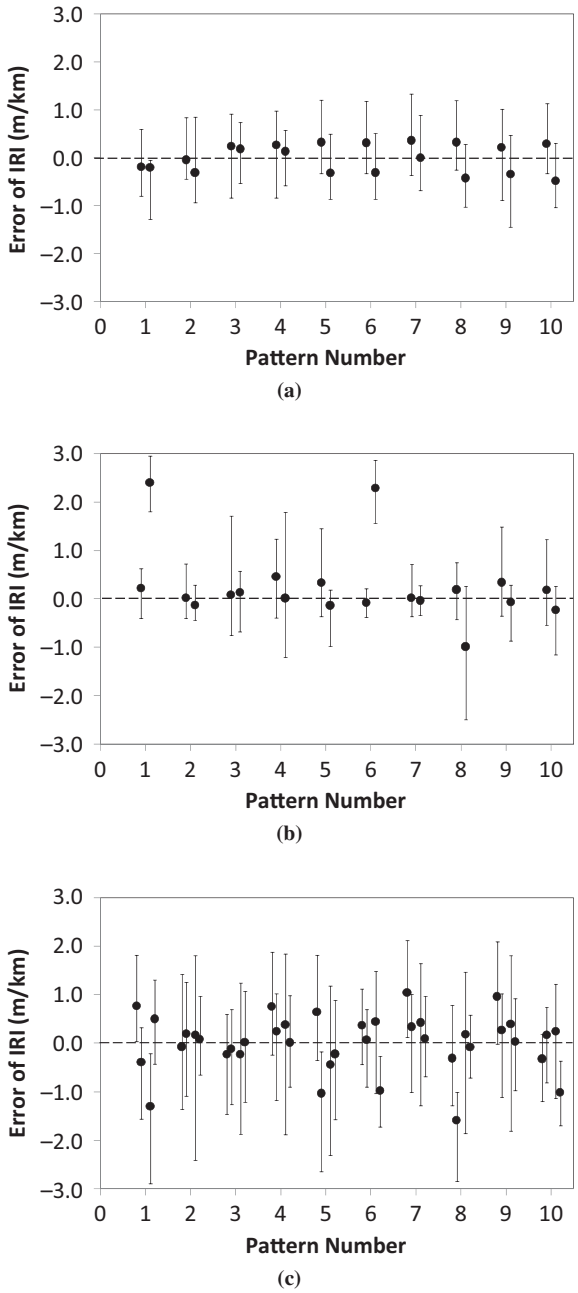

FIGURE 5 Mean values and ranges of imputation results for road sections studied: (*a*) 25% missing data, SHRP ID 28-1802; (*b*) 20% missing data, SHRP ID 20-1005; and (*c*) 25% missing data, SHRP ID 25-1002.

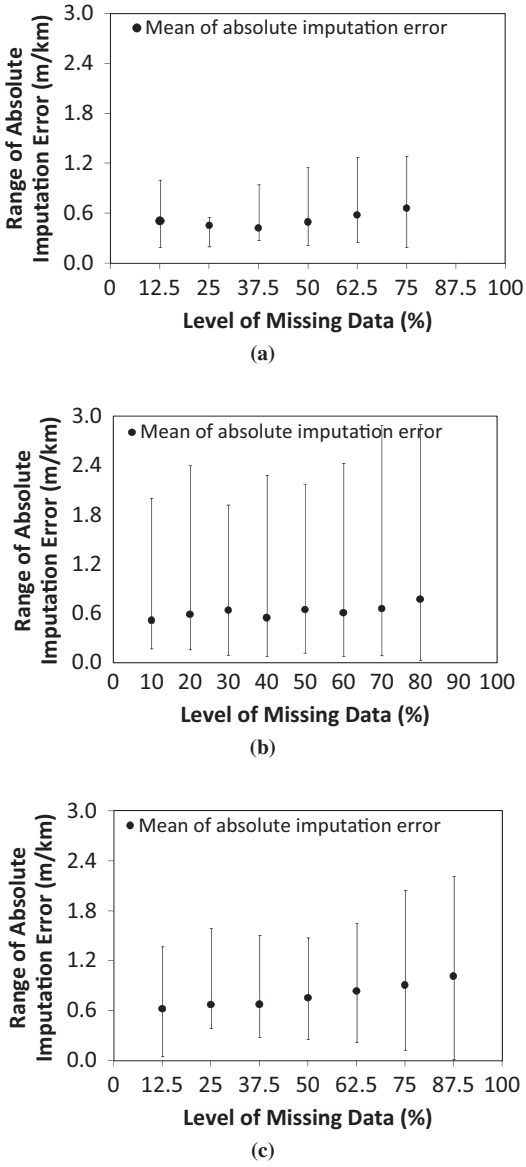

FIGURE 6 Mean errors of imputation data against level of missing data: (*a*) Road Section SHRP ID 28-1802, (*b*) Road Section SHRP ID 20-1005, and (*c*) Road Section SHRP ID 25-1002.

Alternative hypothesis (H_1). The mean imputed value, which is obtained from 10 imputation analyses (μ_z), is different from the actual data value μ_0 from the original data record of the given road section:

H_0:

$$\mu_z \neq \mu_0$$

For each data point in Figure 6, a hypothesis test is performed for a given level of confidence to determine whether the imputed mean value is different from the actual value. Table 2 presents, for a confidence level of 95%, the results of the hypothesis test for all the cases of the three road sections studied. These results are plotted in Figure 7.

From the results in Table 2 and Figure 7, for a permissible error of 20% (i.e., corresponding to the case of 80% "no difference" in Table 2) in the multiple imputation process, the maximum allowable percentage of missing data is 30.3% for Road Section SHRP ID 28-1802, 20% for Road Section SHRP ID 20-1005, and 18.75% for Road Section SHRP ID 25-1002. Thus, setting 25% as the limit of the proportion of missing data appears reasonable for practical application when the records have no abrupt changes in roughness data (i.e., no pavement rehabilitation) and as does applying a limit of 15% when the records involve abrupt changes in roughness data caused by pavement rehabilitation.

The 20% permissible error chosen for illustration was based on the findings of NCHRP Project 20-24 (37B) that network-level IRI data

TABLE 2 Results of Hypothesis Testing of Difference Between Imputed IRI Values of Missing Data and Actual IRI Values

Missing Information (%)	Difference Between Imputed IRI Values and Actual Values at 95% Confidence Interval: Number of Imputations Showing "No Difference in Results"	Number of Imputations Showing "Significant Difference in Results"	Cases Showing "No Difference in Results" (%)
Road Section SHRP ID 28-1802			
12.5	9	1	90.0
25.0	17	3	85.0
37.5	22	8	73.3
50.0	24	16	60.0
62.5	27	23	54.0
75.0	31	29	51.7
Road Section SHRP ID 20-1005			
10	9	1	90.0
20	16	4	80.0
30	21	9	70.0
40	25	15	62.5
50	29	21	58.0
60	30	30	50.0
70	34	36	48.6
80	38	42	47.5
Road Section SHRP ID 25-1002			
12.5	17	3	85
25	30	10	75
37.5	40	20	66.7
50	45	35	56.3
62.5	52	48	52.0
75	57	63	47.5
87.5	53	87	37.9

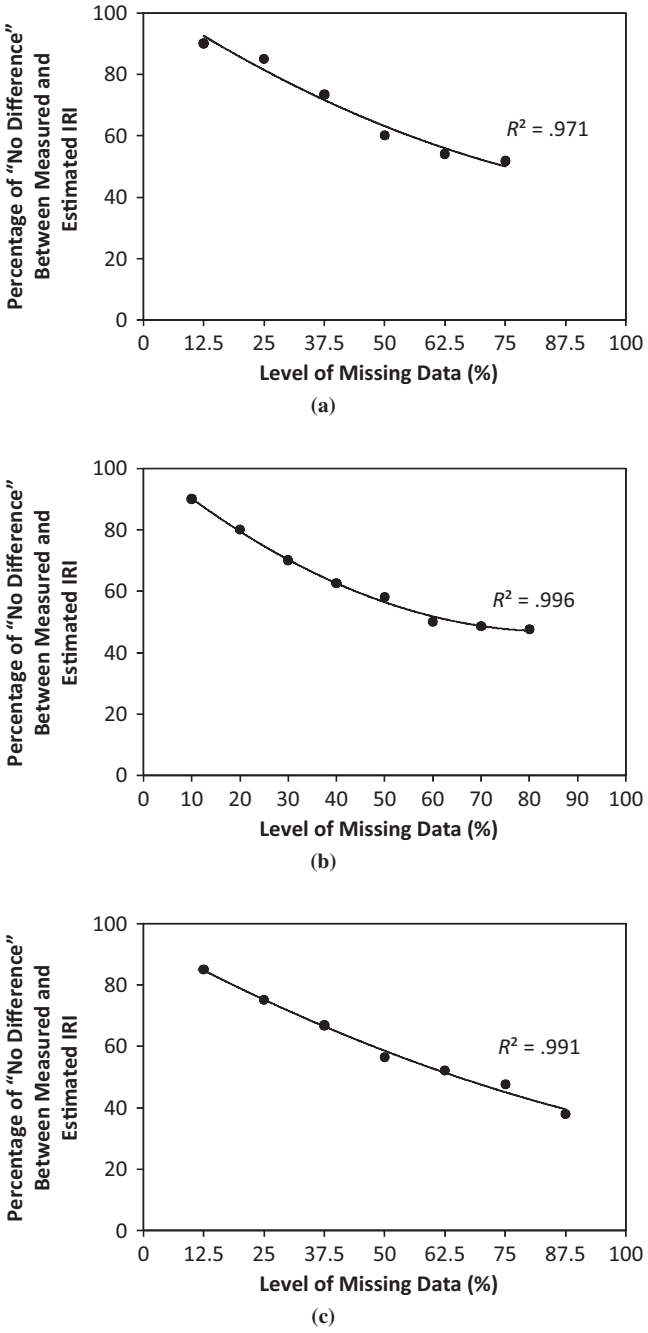

FIGURE 7 Effect of proportion of missing data on imputation results: (*a*) Road Section SHRP ID 28-1802, (*b*) Road Section SHRP ID 20-1005, and (*c*) Road Section SHRP ID 25-1002.

from different states contain baseline measurement error on the order of 15% because of differences in equipment, calibration practices, and variations across operators (*23*). In addition, other possible error sources are possible, including environment-related measurement conditions, variation in measurement speed, and variation of IRI with lateral position in the pavement. Hence, an error of 20% was chosen as the permissible value in the present application.

Overall Comments

The error analysis presented in the preceding sections showed that imputation errors tended to increase with the level of missing data and that abrupt changes in the data of the roughness records caused by pavement resurfacing or rehabilitation would lead to increased errors in the imputation results. As Figure 6 shows, the increased errors from rising levels of missing data are also associated with increased variances of the imputed data. This trend implies that the reliability level of data imputation decreases as the level of missing data increases.

Furthermore, performing pavement rehabilitation within the analysis period, with a resultant abrupt fall in the roughness value, had a significant negative impact on the error magnitude and the reliability of the imputed data. This impact can be expected because rehabilitation caused a discontinuity in the deterioration trend of the roughness data. From the analysis presented, the following recommendations can be made about the maximum proportion of missing data allowable in the application of data imputation in pavement roughness analysis:

1. Allowing up to 20% error in the multiple imputation analysis at a confidence level of 95%, 25% of missing data appears to be a reasonable allowable maximum limit for analysis of time series

data on pavement roughness that contain no rehabilitation within the analysis period. When pavement rehabilitation occurs within the analysis period, the maximum proportion of imputed data should be limited to 15%.

2. Alternatively, a roughness data record that contains pavement rehabilitation operations may be preprocessed before data imputation analysis. This preprocessing will divide the original data record into one or more data records at the year or years of rehabilitation so that each new subdata record will contain time series data on surface roughness beginning after a year of construction–rehabilitation and ending before a year of construction–rehabilitation. In this way, all new subdata records will contain no rehabilitation within the analysis period, and the allowable maximum proportion of missing data can be set to 25% in the data imputation analysis of all subdata records.

CONCLUSIONS

This paper has presented a procedure to evaluate the effect of the level of missing data on the results of data imputation in pavement management analysis. A numerical example using pavement roughness data was presented to illustrate the proposed procedure and to analyze the error and reliability characteristics of imputed data for three road sections. The roughness data of the three road sections were obtained from the LTPP database. From these data records, data sets with different proportions of missing data were randomly generated to study the effect of the level of missing data.

The analysis suggests that the errors of imputed data increased with the level of missing data and that their magnitudes were affected significantly by the effect of pavement rehabilitation. For the three road sections studied, the presence of rehabilitation within the period of the roughness record analyzed caused the mean imputation errors to increase from a range of 0.2 to 0.4 m/km to about 0.3 to 0.7 m/km.

On the basis of the examples, the study proposed maximum allowable proportions of missing data for the application of data imputation in analysis of pavement roughness. Allowing up to 20% error in the multiple imputation analysis at a confidence level of 95%, the study recommends 25% of missing data as a reasonable allowable maximum limit for analysis of time series data on pavement roughness that have no pavement rehabilitation within the analysis period. When pavement rehabilitation occurs within the analysis period, the recommended maximum proportion of imputed data is 15%. These findings were obtained for data that were missing completely at random. For data with systematic errors other than random errors, a separate study is required to investigate the specific problems.

The study also proposed the preprocessing of data records to eliminate the influence of pavement rehabilitation. This preprocessing is achieved by dividing the data record into subrecords, each containing time series data on surface roughness that begin from a year of rehabilitation and end before the next rehabilitation year. Through this process, the maximum allowable limit of 25% missing data can be uniformly applied to the imputation analysis of all data records.

REFERENCES

1. 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.
2. *LTPP Infopave.* FHWA, U.S. Department of Transportation. http://www.infopave.com. Accessed May 20, 2014.
3. 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.
4. Flintsch, G. W., and K. K. McGhee. *NCHRP Synthesis of Highway Practice 401: Quality Management of Pavement Condition Data Collection.* Transportation Research Board of the National Academies, Washington, D.C., 2009.
5. Amado, V., and K. Bernhardt. Expanding the Use of Pavement Condition Data Through Knowledge Discovery in Databases. *Proc., International Conference on Applications of Advanced Technologies in Transportation Engineering,* Cambridge, Mass., 2002, pp. 394–401.
6. Bennett, C. R. Sectioning of Road Data for Pavement. *Proc., 6th International Conference on Managing Pavements,* Queensland, Australia, 2004.
7. Cismondi, F., A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. Sousa, and S. N. Finkelstein. Missing Data in Medical Databases: Impute, Delete or Classify? *Artificial Intelligence in Medicine,* Vol. 58, No. 1, 2013, pp. 63–72.
8. Rubin, D. B., and N. Schenker. Multiple Imputation in Health-Care Databases: An Overview and Some Applications. *Statistics in Medicine,* Vol. 10, No. 4, 1991, pp. 585–598.
9. Saunders, J. A., N. M. Howell, E. Spitznagel, P. Dori, E. K. Proctor, and R. Pescarino. Imputing Missing Data: A Comparison of Methods for Social Work Researchers. *Social Work Research,* Vol. 30, No. 1, 2006, pp. 19–32.
10. King, G., J. Honaker, A. Joseph, and K. Scheve. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review,* Vol. 95, No. 1, 2001, pp. 49–69.
11. Schafer, J. L. Multiple Imputation: A Primer. *Statistical Methods in Medical Research,* Vol. 8, 1999, pp. 3–15.
12. Peng, C. Y. J., M. Harwell, S. M. Liou, and L. H. Ehman. Advances in Missing Data Methods and Implications for Educational Research. In *Real Data Analysis* (S. Sawilowsky, ed.), Greenwich, Conn., 2006, pp. 31–78.
13. Preston, N. J., P. Fayers, S. J. Walters, M. Pilling, G. E. Grande, V. Short, E. Owen-Jones, C. J. Evans, H. Benalia, I. J. Higginson, and C. J. Todd. Recommendations for Managing Missing Data, Attrition and Response Shift in Palliative and End-Of-Life Care Research. *Palliative Medicine,* Vol. 27, No. 10, 2013, pp. 899–907.
14. Little, R. J. A., and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, New York, 1987.
15. Schlomer, G. L., S. Bauman, and N. A. Card. Best Practices for Missing Data Management in Counseling Psychology. *Journal of Counseling Psychology,* Vol. 57, No. 1, 2010, pp. 1–10.
16. *Long-Term Pavement Performance (LTPP) Database.* LTPP DataPave Online. http://www.ltpp-products.com/DataPave. Accessed June 3, 2014.
17. Rubin, D. B. *Multiple Imputation for Nonresponse in Survey.* John Wiley & Sons, New York, 1987.
18. Enders, C. K. *Applied Missing Data Analysis.* Guilford Press, New York, 2010.
19. Farhan, J., and T. F. Fwa. Augmented Stochastic Multiple Imputation Model for Airport Pavement Missing Data Imputation. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2449,* Transportation Research Board of the National Academies, Washington, D.C., 2014, pp. 96–104.
20. Schafer, J. L. *Analysis of Incomplete Multivariate Data.* Chapman & Hall/CRC, Boca Raton, Fla., 1997.
21. Schafer, J. L. *NORM Users Guide: Multiple Imputation of Incomplete Multivariate Data Under a Normal Model,* Version 2. Pennsylvania State University, State College, 1999.
22. Hahn, G., and S. Shapiro. *Statistical Models in Engineering.* John Wiley & Sons, New York, 1967.
23. Harrison, F., and H. A. Park. *NCHRP Report 20-24 (37B): Comparative Performance Measurement: Pavement Smoothness.* Transportation Research Institute, University of Michigan, Ann Arbor, and NCHRP, Transportation Research Board of the National Academies, Washington, D.C., 2008.

The Standing Committee on Pavement Management Systems peer-reviewed this paper.