

# Effect of Proportion of Missing Data on Application of Data Imputation in PMS

by

J. Farhan<sup>1</sup>, B. H. Setiadji<sup>1</sup> and T. F. Fwa<sup>2</sup>

<sup>1</sup>Research Fellow and <sup>2</sup>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:	=	5221	words
Number of tables: 2 (2 x 250)	=	500	words equivalent
Number of figures: 7 (7 x 250)	=	1750	words equivalent
-----			
Total number of words	=	7421	words equivalent

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

July 2014

1  
2  
3 **Effect of Proportion of Missing Data**  
4 **on Application of Data Imputation in PMS**

5 J. Farhan, B. H. Setiadji and T. F. Fwa  
6

7  
8 **Abstract**  
9

10 Missing data are commonly found in pavement condition/performance databases. A common  
11 practice today is to apply statistical imputation methods to replace the missing data with  
12 imputed values. It is thus important for pavement management decision makers to know the  
13 uncertainty and errors involved in the use of datasets with imputed values in their analysis. An  
14 equally important information of practical significance is the maximum allowable proportion  
15 of missing data (i.e. level of data missingness in the pavement condition/performance records)  
16 that will still produce results with acceptable magnitude of error or risk when using imputed  
17 data. This paper proposes a procedure for determining such useful information. A numerical  
18 example analyzing pavement roughness data is presented to demonstrate the procedure through  
19 evaluating the error and reliability characteristics of imputed data. The roughness data of three  
20 road sections were obtained from the LTPP database. From these data records, datasets with  
21 different proportions of missing data were randomly generated to study the effect of level of  
22 data missingness. The analysis shows that the errors of imputed data increased with the level  
23 of data missingness, and their magnitudes are significantly affected by the effect of pavement  
24 rehabilitation. On the application of data imputation in PMS, the study suggests that at 95%  
25 confidence level, 25% of missing data appears to be a reasonable allowable maximum limit for  
26 analyzing pavement roughness time series data not involving rehabilitation within the analysis  
27 period. When pavement rehabilitation occurs within the analysis period, the maximum  
28 proportion of imputed data should be limited to 15%.  
29  
30  
31  
32  
33  
34  
35

36 **Effect of Proportion of Missing Data**  
37 **on Application of Data Imputation in PMS**  
38

39  
40 **INTRODUCTION**  
41

42 Engineering analysis and decision making in a pavement management system (PMS) are data-  
43 driven processes heavily dependent on the quality and accuracy of the data records.  
44 Unfortunately, in practice, the data records in most pavement management systems contain  
45 missing data (1-3). Therefore, missing-data management is an important element in the  
46 engineering analysis and decision making of a pavement management system. According to  
47 an NCHRP Synthesis Report (4), 61% of the pavement agencies in USA included in a survey  
48 used software routine to check for missing data.  
49

50 Since pavement condition and performance data are time-specific information, re-collection of  
51 missing past records through field survey is not possible nor meaningful. Under this situation,  
52 the PMS engineer has the option to discard the records with missing data and proceed with the  
53 remaining records. This is not always desirable as it means making engineering analysis with  
54 a reduced data space, and ignoring some recorded data which could have important  
55 implications to pavement maintenance or traffic operations. A procedure which is increasingly  
56 being adopted today is to apply suitable data imputation techniques to fill up the incomplete  
57 records with imputed values and perform engineering analysis without discarding those records  
58 (4-6).  
59

60 In the application of data imputation methods to manage missing data records in PMS, one  
61 must be aware that the techniques are statistical in nature and uncertainties are involved in the  
62 imputed data values. Knowing the likely magnitudes of the errors involved and the reliability  
63 of the dataset containing imputed data would allow the engineers to make informed decisions  
64 whether to discard the incomplete data records or to proceed with the full set of records made  
65 complete with imputed data. Therefore, a relevant issue is to determine the upper limit of the  
66 proportion of missing data at which filling up the incomplete data records with imputed data  
67 would still provide an accurate representation of the pavement condition. This is the focus of  
68 the present research. Using pavement roughness data from the Long-Term Pavement  
69 Performance Program (LTPP) database, this study examines how different proportions of  
70 missing data would affect the accuracy and reliability of imputed datasets.  
71

72 **SIGNIFICANCE OF STUDY**

73 The theory and principle of statistical quality assurance, in regard to the imputation of missing  
74 data, are well developed and has been applied by researchers and practitioners in a number of  
75 field of studies, notably in the disciplines of medical studies and social sciences (7-10). The  
76 issue of the upper limit threshold for the application of data imputation procedures has also  
77 been addressed by researchers in those disciplines. For instance, Schafer (11) suggested using  
78 statistical data imputation approaches in medical research only when not more than 5 percent  
79 data is missing. On the other hand, in dealing with missing data in public health studies, Bennett  
80 (12) recommended 20 percent missing data as the maximum threshold for the application of  
81 data imputation procedures. However, in their studies of palliative and end-of-life care, Preston  
82 et al. (13) recommended that high rates of attrition or missing data should not be seen as  
83 indicative of poor design and that it is more important to design a clear statistical analysis plan  
84 to account for missing data and attrition.

85

86 Little and Rubin (14) introduced the concept of missingness to highlight the importance of the  
87 influence of the pattern of missing data on (i) the overall bias introduced, and (ii) the proportion  
88 of missing data that is too high for creating a reasonable a “complete” dataset. For example,  
89 in the case that a very high proportion of data (much higher than 20%) were "missing  
90 completely at random", one could still re-create the dataset with imputed data and capture the  
91 essential characteristics of the original data records. Schlomer et al. (15) concurred that the  
92 pattern of data missingness is a major factor of consideration, but stressed that in determining  
93 whether a certain amount of missingness is problematic, one must first determine if the  
94 resultant imputed dataset has adequate statistical power to detect the effects of interest.

95

96 It is clear from past research in various disciplines on the applications of data imputation in  
97 missing-data management that no simple guidelines can be set for the maximum allowable  
98 proportion of missing data across the board covering all fields of studies. The effect of the  
99 proportion of missing data on the quality of analysis using imputed datasets depends on the  
100 nature and characteristics of the data, as well as the pattern of missing data; and statistical  
101 analyses must be performed to provide a fuller assessment of the effect so that the decision  
102 maker can make an informed decision on how to manage the missing data and the way the data  
103 should be used.

104

105 To the knowledge of the authors, in the field of pavement management studies, in regard to the  
106 use of imputed datasets in pavement management analysis, their impact on data quality and  
107 reliability, and the possible bias introduced to the analysis have not been studied. No guidelines  
108 are available concerning the data management procedure necessary to deal with datasets  
109 containing different extents of missing data. As missing data are commonly encountered in  
110 pavement management data records, the availability of the aforementioned information related  
111 to the use of imputed data would have high practical significance. This paper attempts to  
112 provide some information to partially bridge this knowledge gap by analyzing the effect of  
113 missing data in pavement roughness records.

114

## 115 **APPROACH AND METHODOLOGY OF STUDY**

### 116 **Scope of Study**

117 The common types of pavement condition and performance data that are regularly collected in  
118 a typical pavement management system include pavement distress data (such as cracks, ruts,  
119 potholes, depressions, etc), roughness, friction, and structural condition data derived from non-  
120 destructive falling-weight deflectometer testing. Since the nature and characteristics of each  
121 of these types of data are quite different from one another, it is likely that they will be affected  
122 by missing data differently. It would require a major research effort to examine the effects of  
123 missing data on all types of pavement condition/performance data.

124

125 The scope of the present study is limited to the analysis of the effect of missing data in  
126 pavement roughness records. The framework and concept of the proposed analysis will be  
127 described in this section, followed by a demonstration using an example involving actual  
128 pavement roughness data records. Through the analysis of the numerical example, it is  
129 demonstrated that useful informative insight can be gained into the quality of imputed data  
130 obtained, the magnitude of errors involved as the proportion of missing data increases, and the  
131 statistical reliability implications of the imputed dataset as a function of the proportion of  
132 missing data.

133

## 134 **Framework of Analysis**

135

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

142

143 The proposed analysis consists of the following steps:

144 (1) Selection of complete data records -- The Federal Highway Administration's (FHWA)  
145 Long-Term Pavement Performance Program (LTPP) database (*16*) offers a convenient  
146 source for the selection of pavement roughness data records for the present study. The  
147 roughness data are reported in terms of the International Roughness Index (IRI).

148 (2) Creation of datasets having different levels of data missingness and different patterns of  
149 missingness -- To study the effect of the level of data missingness (i.e. proportion of  
150 missing data), at least six equally spaced levels of data missingness were first identified.  
151 Next, for each specified level of data missingness, a random process was employed to  
152 generate a dataset containing the correct number (say  $n$  number) of missing data by  
153 randomly deleting  $n$  data points from the original complete data records. This random  
154 deletion process is repeated another 9 times so as to produce a total of 10 randomly  
155 generated datasets, each with a different patterns of missingness, for each of the 6 or  
156 more levels of data missingness studied.

157 (3) Computation of imputed values for each dataset containing missing data -- For each of  
158 the datasets containing missing data earlier generated in Step 2, apply a suitable data  
159 imputation method to compute a data value for each of the missing data. At the end of  
160 this step, all the datasets with missing data generated in Step 2 would be transformed into  
161 datasets containing imputed data values. That is, there would be 10 datasets containing  
162 imputed data for each level of data missingness. The technique of Multiple Imputation  
163 (MI) was adopted for computing imputed data in this study. The imputed value for each  
164 missing data in each of the 10 generated datasets is obtained as the mean value of 10  
165 imputation runs. The concept and procedure of computation of the MI technique is  
166 explained in the next section.

167 (4) Performing of error and reliability analysis – Using the original complete data records as  
168 the base reference, the errors of the imputed data can be computed and analyzed. The  
169 variation of the errors with the level of data missingness can be examined. The statistical  
170 reliability of the imputed datasets at different levels of data missingness can also be  
171 established by means of hypothesis testing.

172

## 173 **Multiple Imputation (MI) Technique for Data Imputation**

174

175 The most widely used method today in performing data imputation for missing data is the  
176 Multiple Imputation technique first introduced by Rubin (*17*). This method is known to  
177 produce unbiased imputed data and parameter estimates (*14, 17, 18*). The authors have  
178 demonstrated in their earlier work (*19*) that the Multiple Imputation method out-performed the  
179 conventional methods (such as the deletion method, and the substitution methods using mean,  
180 interpolation, or regression) in handling missing pavement condition/performance data, and

181 provides an effective approach to impute missing data required in a pavement management  
182 system.

183 The process of Multiple Imputation consists of three main phases: imputation, analysis and  
184 polling phase. In the imputation phase, the available measured data are used to estimate  
185 distribution parameters, which are then used to estimate the missing data values. In the analysis  
186 phase, each imputed value is analyzed together with the corresponding available ones using  
187 statistical procedure to produce a new imputed value. This iterative process continues until the  
188 imputed value changes very little from one iteration to the next. By repeating this procedure,  
189 multiple imputations of the missing values are generated. Finally, on the pooling phase, the  
190 integration of the multiple imputation results into a single set of result to produce overall  
191 estimates and standard errors that reflect missing-data uncertainty. These combined standard  
192 errors are useful for statistical significance testing and drawing of inferential conclusions.

193  
194 The working of the Multiple Imputation method makes use of two main algorithms, namely  
195 Expectation Maximization (EM) and Data Augmentation (DA). The procedure of data  
196 imputation adopted in the study involves of the following steps:

- 197 • Step I: Data Transformation – Firstly, it is required to transform the data to  
198 approximately normal before imputation using a transformation functions, such as logit,  
199 log or square root functions. After imputation, the data will be transformed back to  
200 their original scale.
- 201 • Step II: Imputation using EM – EM uses the maximum likelihood approach to perform  
202 the imputation function in the “imputation and analysis phase” of the MI procedure.  
203 This step will generate estimates of missing values for the data matrix with the  
204 convergence criterion that the maximum relative parameter change in the value of any  
205 parameter during the iterative process is less than  $10^{-6}$ .
- 206 • Step III: Imputation using DA – With the initial parameter estimates from the EM  
207 algorithm serving as the basis, the DA algorithm carries out multiple imputations as  
208 explained earlier in the “imputation and analysis phase” of the MI procedure. The  
209 commonly adopted practice of 10 imputations (14, 20) is applied in this study.
- 210 • Step IV: Synthesis of Estimates – Average over the multiple estimates of the multiple  
211 imputation analysis to obtain the final set of estimates (17).

212

## 213 **ILLUSTRATIVE EXAMPLE: IMPUTATION OF ROUGHNESS DATA**

214

### 215 **IRI Records in LTPP Database**

216

217 From the LTPP database (16) that provides measured records of pavement roughness data  
218 covering 24 years from 1989 to 2012, the following three records were extracted for the  
219 illustrative analysis of this study:

- 220 • Road Section SHRP ID 28-1802 with 8 years of continuous measured annual IRI  
221 (International Roughness Index) data;
- 222 • Road Section SHRP ID 20-1005 with 10 years of continuous measured annual IRI data;  
223 and
- 224 • Road Section SHRP ID 25-1002 with 16 years of continuous measured annual IRI data.

225

226 Table 1 records the measured IRI values and the corresponding times of measurements of the  
227 IRI records of the three road sections. These IRI data are plotted in Figure 1. Although the  
228 annual IRI measurements were not measured at time intervals of exactly 12 months, they can

229 be considered as time series data for the analysis and illustration purpose of the present example  
230 to study the effects of missing data.

231  
232 The three road sections have been selected because their pavement roughness variation trends  
233 display very distinctly different patterns. Road Sections SHRP ID 28-1802 and ID 20-1005  
234 both had roughness value gradually increased with time, except for the latter there was a sharp  
235 drop in roughness value in the last year of the record. The roughness variations of Road Section  
236 ID 25-1002 were characterized by two periods of gentle increases (from year 1 to 7, and from  
237 year 12 to 15), two periods of sharp rises (from year 7 to 9, and from year 12 to 15), a period  
238 of sharp fall (from year 9 to 11) and a mild drop in year 16.

## 239 **Data Representation**

240  
241  
242 Pavement roughness is expected to increase with the number of years of service due to the  
243 impact of traffic loading. However, in the occasion of pavement re-surfacing or rehabilitation,  
244 the roughness would be restored to a lower value. Such maintenance and rehabilitation (M&R)  
245 activities are common in road operations, they occurred for all three road sections considered  
246 in the present study. As indicated in the LTPP database, for Road Section SHRP ID 28-1802,  
247 minor M&R (maintenance and rehabilitation) took place in years 7 and 8 and resulted in slight  
248 decreases in the IRI value. For Road Section ID 20-1005, a minor and a major M&R were  
249 performed in years 5 and 10 respectively. For Road Section SHRP ID 25-1002, the database  
250 records indicated a major and a minor M&R in years 10 and 16 respectively.

251  
252 In the data imputation analysis, this situation was handled by introducing an M&R dummy  
253 variable. The dummy variable would be assigned a value of 1 if there was an M&R operation  
254 in the year of interest, and 0 otherwise. For the Road Section ID 25-1002, it is noted from  
255 Figure 1 that although the LTPP database indicated an M&R operation in year 11, a drop in the  
256 IRI value started to occur in year 10. It is suspected that part of the M&R might have  
257 commenced in year 10 and resulted in the fall of IRI.

## 258 **Generation of Datasets with Missing Data**

259  
260  
261 To study the effect of the proportion of missing data and determine the maximum allowable  
262 proportion of missing data, datasets with proportions of missing data ranging from  
263 approximately 10 to 90% were created for the three road sections studied. These datasets with  
264 missing data were randomly generated from the respective original complete data records of  
265 the three road sections. The levels of data missingness created for the three road sections  
266 studied are as follows:

- 267 • SHRP ID 28-1802: A total of 6 levels of data missingness was created. The percentages  
268 of missing data created were 12.5%, 25%, 37.5%, 50%, 62.5% and 75%;
- 269 • SHRP ID 20-1005: A total of 8 levels of data missingness was created. The percentages  
270 of missing data created were 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80%;
- 271 • SHRP ID 25-1002: A total of 7 levels of data missingness was created. The percentages  
272 of missing data created were 12.5%, 25%, 37.5%, 50%, 62.5%, 75% and 87.5%.

273  
274 For each of the three road sections, at each level of data missingness, 10 different patterns of  
275 missing data were randomly created. Figures 2, 3 and 4 show all the patterns of missing data  
276 created for Road Sections SHRP ID 28-1802, SHRP ID 20-1005 and SHRP ID 25-1002  
277 respectively.

278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326

## **Analysis of Imputation Results**

### Error Analysis

The 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 road section roughness record analyzed, 10 patterns of missing data were created for each level of data missingness (see Figures 2 to 4); and for each pattern of missing data for a given level of missingness, 10 imputation runs were made using the MI technique. Hence, there were 10 imputed values for each missing data form the 10 imputation runs, and the error of each imputed roughness value is 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 data missisngness (i.e. proportions of missing data) for the three road sections studied. Figure 5(a) shows the results of imputation errors for the datasets of the level of data missingness with 25% missing data for the roughness data of the Road Section SHRP ID 28-1802. At 25%, there were two missing data per dataset (i.e. two missing data per pattern of missing data, see Figure 2). Hence, in Figure 5(a), there are two sets of error results for each of the 10 patterns of missing patters. Similarly, in Figure 5(b) for the roughness data of Road Section SHRP ID 20-1005, there are two sets of error results for each of the 10 patterns of missing patters at the level of missing data of 20%. For the roughness data of Road Section SHRP ID 25-1002, there are four sets of error results for each of the 10 patterns of missing patters at the level of missing data of 25%.

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

- (1) For Road Section SHRP ID 28-1802, there are no clear trends of variation among the errors for the 10 patterns. This is within expectation because the imputed data values were generated through a random process.
- (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 the two patterns both contain a 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 would introduce 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 two years with a sharp fall of the roughness value. This observation reinforces the earlier observation made in the preceding paragraph concerning larger imputation errors associated with sharp changes in roughness data.

Another error characteristic of interest is how errors vary with the level of data missingness. Figure 6 plots the mean and range of the absolute errors of imputed values against the levels of data missisngness (i.e. proportions of missing data) for all the cases of the three road sections studied. The following characteristics can be observed:

- (1) For all three road sections, the magnitude of imputation errors increased with the level of data missingness. Road Section SHRP ID 28-1802 which has no abrupt changes in its roughness data, had the smallest mean imputation errors ranging from about 0.2 to



327 0.4 m/km; while the other two road sections containing abrupt changes in their  
328 roughness data, had larger mean imputation errors ranging from about 0.3 to 0.7 m/km.  
329 (2) The range of the errors was also found to increase with the level of data missingness in  
330 general. Among the three road sections examined, the two road sections with abrupt  
331 changes in roughness data (i.e. SHRP ID 20-1005 and SHRP ID 25-1002) again  
332 displayed significantly larger ranges of variation in the range of the values of imputation  
333 mean.

### 334 Reliability Analysis

335  
336  
337 The uncertainty involved in the imputation of missing values is reflected in the variations of  
338 the multiple imputed values for each missing data value of the example problem. Such  
339 variations are seen in the plots of Figures 5 and 6, where the distributions in the errors of  
340 imputed values as well as the variations among the means of different imputation runs,  
341 respectively, are depicted.

342  
343 With the error characteristics presented in Figures 5 and 6, a statistical reliability analysis of  
344 the imputation results can be performed. For the purpose of the present study, a hypothesis  
345 testing was performed to compare the mean computed value for each missing data with the  
346 corresponding actual data value of the original complete record. Since for each missing data,  
347 there were 10 imputed values, the Student's *t*-test (21) was employed for the test. The  
348 hypothesis testing considers the following null and alternative hypothesis:

349 Null hypothesis ( $H_0$ ): The mean imputed value which is obtained from 10 imputation  
350 analyses,  $\mu_z$ , is no different from the actual data value  $\mu_0$  from the original data record of  
351 the given road section, i.e.

$$352 H_0 : \mu_z = \mu_0$$

353 Alternative hypothesis ( $H_1$ ): The mean imputed value which is obtained from 10 imputation  
354 analyses,  $\mu_z$ , is different from the actual data value  $\mu_0$  from the original data record of the  
355 given road section, i.e.

$$356 H_1 : \mu_z \neq \mu_0$$

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

361  
362 From the results in Table 2 and Figure 7, taking a permissible error of 20% (i.e. corresponding  
363 to the case of 80% "no difference" in Table 2) in the multiple imputation process, the maximum  
364 allowable percentage of missing data is 30.3% for Road Section 28-1802, 20% for Road  
365 Section 20-1005, and 18.75% for Road Section 25-1002. Thus, it appears reasonable for  
366 practical application to set 25% as the limit of the proportion of missing data when there are  
367 no abrupt changes of roughness data (i.e. no pavement rehabilitation) in the data records, and  
368 apply a limit of 15% when the data records involve abrupt changes in roughness data caused  
369 by pavement rehabilitation.

### 370 Overall Comments

371  
372  
373 The error analysis presented in the preceding sections showed that imputation errors increased  
374 with the level of data missingness, and that abrupt changes in the data of the roughness records  
375 brought about by pavement resurfacing or rehabilitation would lead to increased errors in the

376 imputation results. As depicted in the plots of Figures 5 and 6, the increased errors due to  
377 rising levels of data missingness are also associated with increased variances of the imputed  
378 data. This implies that the reliability level of data imputation decreases as the level of data  
379 missingness increases.

380

381 It was also observed that performing pavement rehabilitation within the analysis period,  
382 resulting in an abrupt fall in the roughness value in the data record, had a significant negative  
383 impact on the error magnitude and reliability of the imputed data. This can be expected because  
384 the action of rehabilitation caused a discontinuity in the deterioration trend of the roughness  
385 data. Based on the analysis presented, the following recommendations can be made regarding  
386 the maximum proportion of missing data allowable in the application of data imputation in  
387 pavement roughness analysis:

388 (1) Allowing up to 20% error in the multiple imputation analysis at a confidence level of  
389 95%, 25% of missing data appears to be a reasonable allowable maximum limit for  
390 analyzing pavement roughness time series data not having any pavement  
391 rehabilitation within the analysis period. When pavement rehabilitation occurs within  
392 the analysis period, the maximum proportion of imputed data should be limited to  
393 15%.

394 (2) Alternatively, a pre-processing before data imputation analysis may be performed to  
395 a roughness data record that contains pavement rehabilitation operations. This pre-  
396 processing will break the original data record into one or more data records at the  
397 year(s) of rehabilitation, so that each new sub-data record will contain roughness time  
398 series data beginning after a year of construction/rehabilitation and ending before a  
399 year of construction/rehabilitation. In this way, all new sub-data records will not  
400 contain any rehabilitation within the analysis period, and the allowable maximum  
401 proportion of missing data can be set as 25% for in the data imputation analysis for all  
402 sub-data records.

403

## 404 CONCLUSIONS

405

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

413

414 The analysis shows that the errors of imputed data increased with the level of data missingness,  
415 and their magnitudes are significantly affected by the effect of pavement rehabilitation. For  
416 the three road sections studied, the presence of rehabilitation within the period of the roughness  
417 record analysed caused the mean imputation errors to increase from a range of 0.2 to 0.4 m/km  
418 to about 0.3 to 0.7 m/km.

419

420 Based on the examples analyzed, the study proposed maximum allowable proportions of  
421 missing data for the application of data imputation in pavement roughness analysis. Allowing  
422 up to 20% error in the multiple imputation analysis at a confidence level of 95%, the study  
423 recommends 25% of missing data as a reasonable allowable maximum limit for analyzing  
424 pavement roughness time series data not having any pavement rehabilitation within the analysis

425 period. When pavement rehabilitation occurs within the analysis period, the recommended  
426 maximum proportion of imputed data is 15%.

427

428 The study also proposed performing of pre-processing of data record to eliminate the influence  
429 of pavement rehabilitation. This is achieved by breaking the data record into sub-records, each  
430 containing time series roughness data that begins from a year of rehabilitation and ends before  
431 the next rehabilitation year. By so doing, the maximum allowable limit of 25% missing data  
432 can be uniformly applied to the imputation analysis of all data records.

433

## 434 REFERENCES

435

- 436 1. Amado, V. and Bernhardt, K. L. S. Knowledge Discovery in Pavement Condition Data.  
437 In the 81st Annual Meeting of the Transportation Research Board (TRB), Washington  
438 D.C., 2002.
- 439 2. FHWA. *LTPP Infopave*. <http://www.infopave.com>. Accessed May 20, 2014.
- 440 3. Lindly, J. K., Bell, F. and Sharif U. Specifying Automated Pavement Condition Surveys.  
441 Journal of the Transportation Research Forum, Vol. 44, No. 3, 2005, pp.19-32.1.
- 442 4. National Cooperative Highway Research Program (NCHRP). Quality Management of  
443 Pavement Condition Data Collection. National Cooperative Highway Research Program  
444 Synthesis Report No. 401, Transportation Research Board, Washington, D.C., 2009.
- 445 5. Amado, V. and Bernhardt, K. Expanding the use of pavement condition data through  
446 knowledge discovery in databases. Proc. of the International Conference on Applications  
447 of Advanced Technologies in Transportation Engineering, 2002, pp.394-401.
- 448 6. Bennett, C. R. Sectioning of road data for pavement. Proceedings of the 6th International  
449 Conference on Managing Pavements, Queensland, Australia, 2004.
- 450 7. Cismondi F., Fialho A.S., Vieira S.M., Reti S.R., Sousa J.M. and Finkelstein S.N.  
451 Missing data in medical databases: impute, delete or classify? Artificial Intelligence in  
452 Medicine, 58(1), 2013, pp. 63-72.
- 453 8. Rubin D.B., Schenker N. Multiple imputation in health-care databases: an overview and  
454 some applications. Statistics in Medicine, 10(4), 1991, pp. 585-598.
- 455 9. Saunders J. A., Howell N. M., Spitznagel E., Dori P., Proctor E. K. and Pescarino R.  
456 Imputing Missing Data: A Comparison of Methods for Social Work Researchers. Social  
457 Work Research 30(1), 2006.
- 458 10. King G., Honaker J., Joseph A. and Scheve K. Analyzing incomplete political science  
459 data: An alternative algorithm for multiple imputation. American Political Science  
460 Review, 95(1), 2001, pp. 49-69.
- 461 13. Preston N. J., Fayers P., Walters S. J., Pilling M., Grande G. E., Short V., Owen-Jones  
462 E., Evans C. J., Benalia H., Higginson I. J. and Todd C. J., Recommendations for  
463 managing missing data, attrition and response shift in palliative and end-of-life care  
464 research. Palliative Medicine, 27(10), 2013, pp. 899-907.
- 465 14. Little, R.J.A and Rubin, D. B. Statistical analysis with missing data. John Wiley & Sons,  
466 New York, 1987.
- 467 15. Schlomer G. L., Bauman S., and Card N. A., Best Practices for Missing Data  
468 Management in Counseling Psychology. Journal of Counseling Psychology, 57(1), 2010,  
469 pp. 1-10.
- 470 16. Long Term Pavement Performance (LTPP) Database. "LTPP DataPave Online."  
471 <http://www.ltp-products.com/DataPave>, Accessed on 3 June 2014.
- 472 17. Rubin, D.B. *Multiple Imputation for Nonresponse in Survey*, John Wiley & Sons, New  
473 York, 1987.
- 474 18. Enders, C.K. *Applied Missing Data Analysis*. The Guilford Press, New York, 2010.

- 475 19. Farhan, J. and Fwa T. F. Augmented Stochastic Multiple Imputation Model for Airport  
476 Pavement Missing Data Imputation. Transportation Research Record: Journal of the  
477 Transportation Research Board, No. 2336, 2013, pp. 43-54.
- 478 20. Schafer, J.L. *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC, Florida,  
479 1997.
- 480 21. Hahn, G. and Shapiro, S. *Statistical Models in Engineering*. John Wiley & Sons Inc, New  
481 York, NY, 1967.
- 482
- 483

484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504

### **List of Tables**

Table No.

- 1 Observed IRI Values of Road Sections Studied
- 2 Results of Hypothesis Testing of the Difference between Imputed IRI Values of Missing Data and Actual IRI Values

### **List of Figures**

Figure No.

- 1 Measured IRI Data of Road Sections Studied
- 2 Patterns of Missing IRI Data Created for Road Section SHRP ID 28-1802
- 3 Patterns of Missing IRI Data Created for Road Section SHRP ID 20-1005
- 4 Patterns of Missing IRI Data Created for Road Section SHRP ID 25-1002
- 5 Mean Values and Ranges of Imputation Results for Road Section Studied
- 6 Mean Errors of Imputation Data against Level of Data Missingness
- 7 Effect of Proportion of Missing Data on Imputation Results

**TABLE 1 Observed IRI Values of Road Sections Studied**

SHRP ID	State	Year	Time 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	Jul 1995	2.187
		7	Apr 1996	2.142
		8	Oct 1997	1.991
20-1005	Kansas	1	May 1992	2.933
		2	Mar 1993	2.911
		3	May 1994	2.833
		4	Mar 1995	2.964
		5	Apr 1996	2.948
		6	Feb 1997	3.164
		7	Apr 1998	3.369
		8	Mar 1999	3.408
		9	Feb 2000	3.448
		10	May 2001	1.177
25-1002	Massachusetts	1	Oct 1989	1.164
		2	Sep 1990	1.196
		3	Jul 1991	1.189
		4	Sep 1992	1.132
		5	Sep 1993	1.186
		6	Jan 1994	1.408
		7	Jan 1995	1.607
		8	Nov 1996	2.198
		9	Jun 1997	3.387
		10	Jun 1998	2.947
		11	Jul 1999	1.451
		12	Jun 2000	2.791
		13	Apr 2001	2.844
		14	Feb 2002	3.014
		15	Sep 2003	3.245
		16	Apr 2004	2.943

**TABEL 2 Results of Hypothesis Testing of the Difference between Imputed IRI Values of Missing Data and Actual IRI Values**

(a) Road Section ID 28-1802

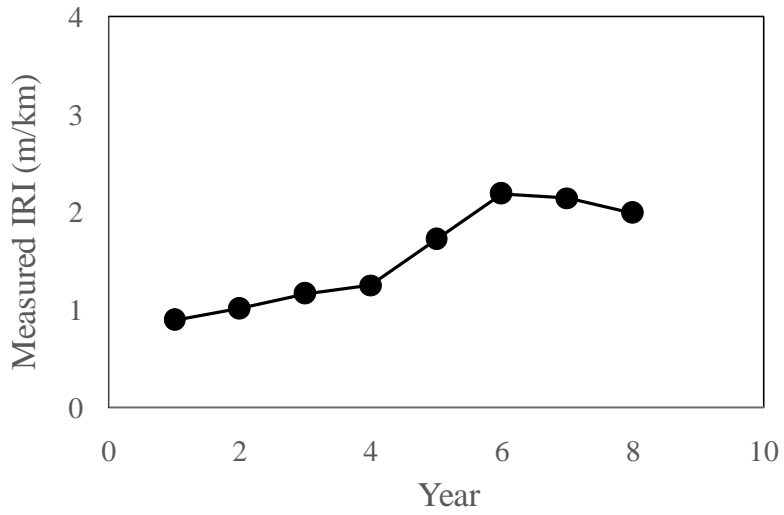
Percentage Missing Data	Difference between Imputed IRI Values and Actual Values at 95% Confidence Interval		
	Number of Imputations Showing “No Difference in Results”	Number of Imputations Showing “Significant Difference in Results”	% Cases Showing “No Difference in Results”
12.5%	9	1	90.0
25.0%	17	3	85.0
37.5%	22	12	73.3
50.0%	24	16	60.0
62.5%	27	23	54.0
75.0%	31	29	51.7

(b) Road Section ID 20-1005

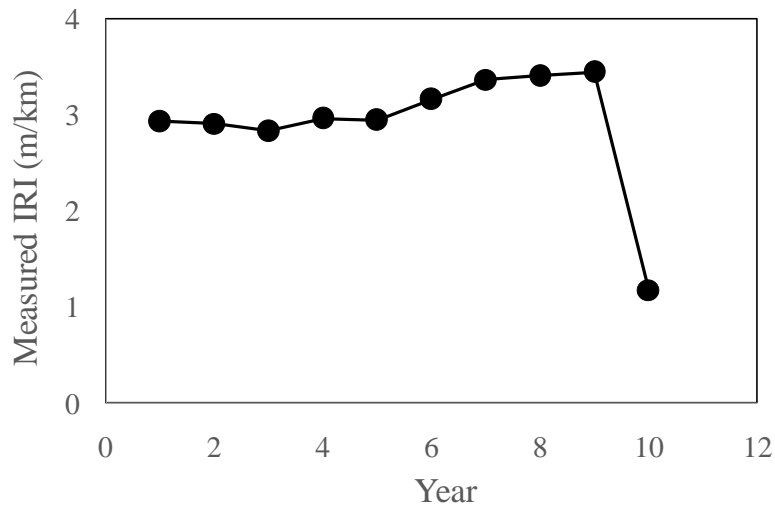
Percentage Missing Data	Difference between Imputed IRI Values and Actual Values at 95% Confidence Interval		
	Number of Imputations Showing “No Difference in Results”	Number of Imputations Showing “Significant Difference in Results”	% Cases Showing “No Difference in Results”
10%	9	1	90.0
20%	16	4	80.0
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

(c) Road Section ID 25-1002

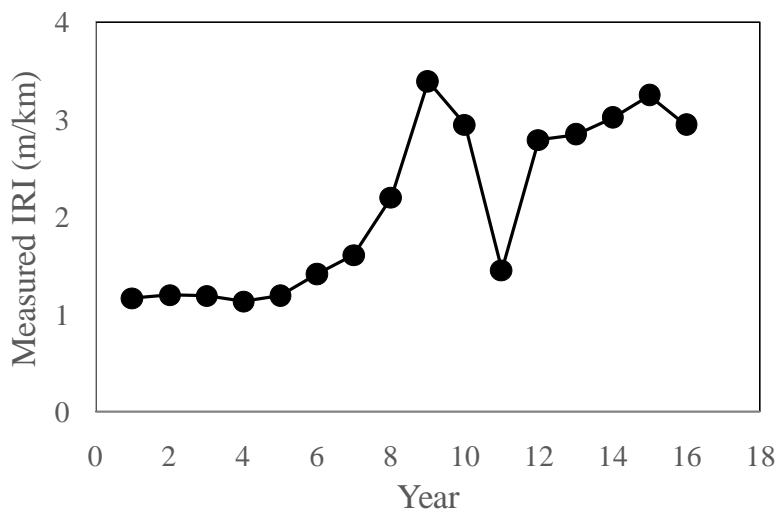
Percentage Missing Data	Difference between Imputed IRI Values and Actual Values at 95% Confidence Interval		
	Number of Imputations Showing “No Difference in Results”	Number of Imputations Showing “Significant Difference in Results”	% Cases Showing “No Difference in Results”
12.5%	17	3	85
25%	30	10	75
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



(a) Road section ID 28-1802



(b) Road section ID 28-1802



(a) Road section ID 25-1002

**FIGURE 1 Measured IRI Data of Road Sections Studied**



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) Pattern 1

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

(a) Pattern 2

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) Pattern 3

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) Pattern 4

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) Pattern 5

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) Pattern 6

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) Pattern 7

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) Pattern 8

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) Pattern 9

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) Pattern 10

Remarks: 0 = IRI missing data

**FIGURE 2 Patterns of Missing IRI Data Created for Road Section SHRP ID 28-1802**

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) Pattern 1

Year	Percentage of Missing Data							
	10%	20%	30%	40%	50%	60%	70%	80%
1	1	1	1	1	0	1	0	0
2	1	1	0	1	1	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) Pattern 2

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) Pattern 3

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) Pattern 4

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) Pattern 5

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) Pattern 6

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) Pattern 7

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) Pattern 8

Year	Percentage of Missing Data							
	10%	20%	30%	40%	50%	60%	70%	80%
1	1	1	1	1	0	0	0	0
2	1	1	1	1	1	0	1	1
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) Pattern 9

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) Pattern 10

**FIGURE 3 Patterns of Missing IRI Data Created for Road Section SHRP ID 20-1005**

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) Pattern 1

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) Pattern 2

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) Pattern 3

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) Pattern 4

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) Pattern 5

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) Pattern 6

**FIGURE 4 Patterns of Missing IRI Data Created for Road Section SHRP ID 25-1002 (continued 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) Pattern 7

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) Pattern 8

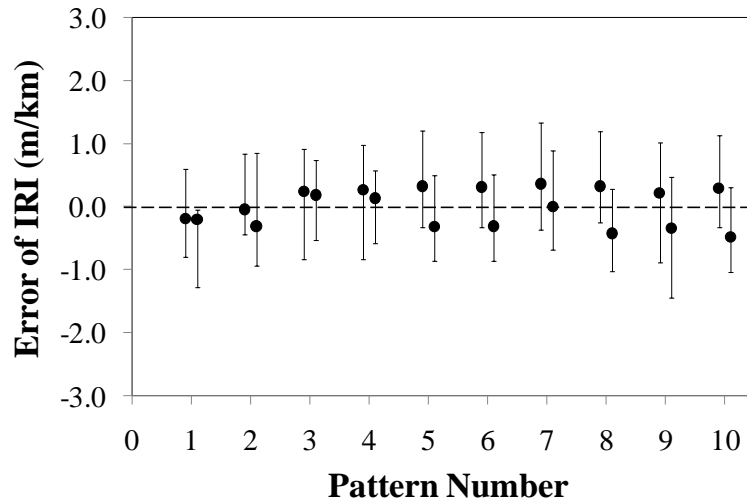
Year	Percentage of Missing Data						
	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%	87.5%
1	1	1	0	1	1	0	0
2	1	1	1	0	0	1	0
3	0	1	0	1	0	0	0
4	1	0	1	0	0	1	0
5	1	1	1	0	1	0	0
6	1	0	1	0	0	1	0
7	1	1	1	0	1	0	0
8	1	1	1	1	0	0	0
9	1	1	1	0	1	0	0
10	1	1	0	1	1	0	0
11	1	0	0	1	0	0	0
12	0	1	0	1	0	0	1
13	1	1	1	1	0	0	0
14	1	0	1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

(i) Pattern 9

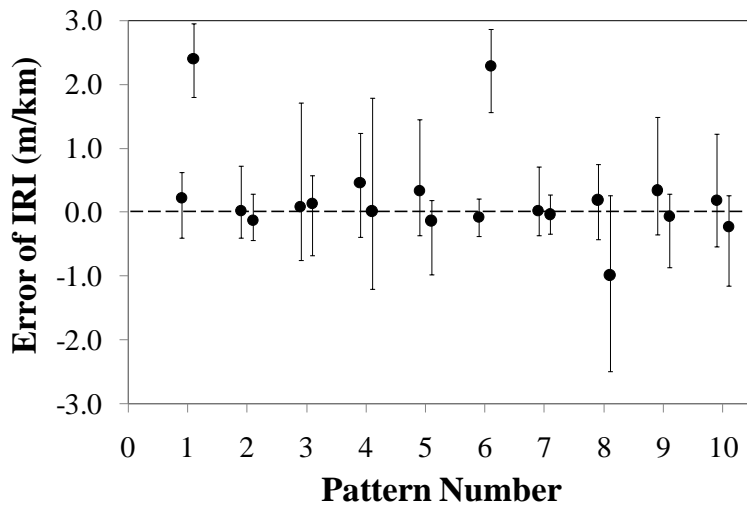
Year	Percentage of Missing Data						
	12.5%	25.0%	37.5%	50.0%	62.5%	75.0%	87.5%
1	0	0	1	0	0	1	0
2	1	1	0	1	0	0	0
3	1	1	1	0	0	1	0
4	1	0	1	1	0	0	0
5	1	1	1	0	0	1	0
6	1	1	0	1	0	0	1
7	1	0	0	1	0	0	1
8	1	1	1	0	1	0	0
9	1	0	0	1	1	0	0
10	1	1	1	1	0	0	0
11	1	1	1	0	1	0	0
12	1	1	1	1	0	0	0
13	0	1	0	0	1	0	0
14	1	1	1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

(j) Pattern 10

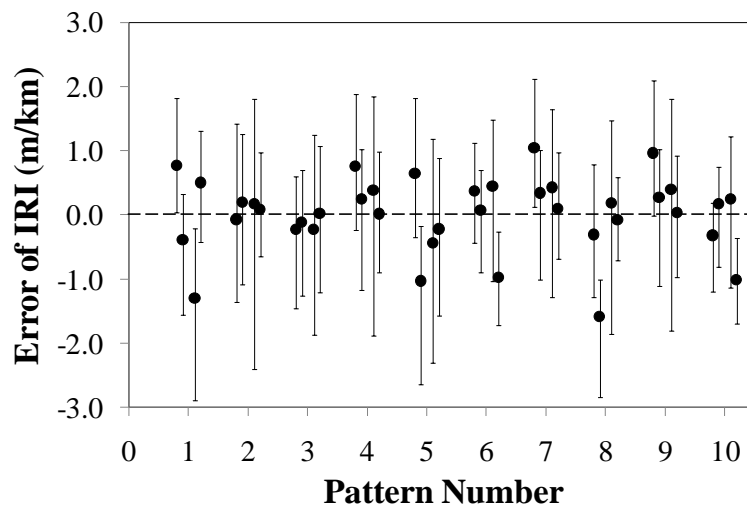
**FIGURE 4 Patterns of Missing IRI Data Created for Road Section SHRP ID 25-1002 (continuation)**



(a) 25% missing data for Road Section ID 28-1802

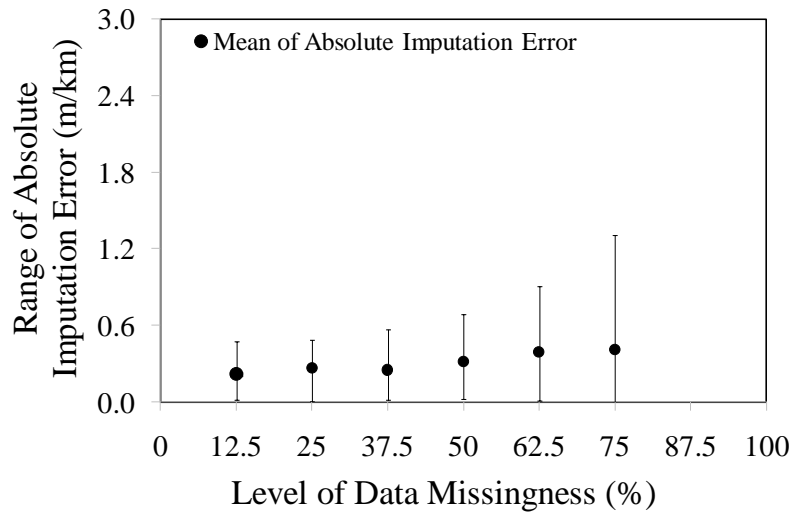


(b) 20% missing data for Road Section ID 20-1005

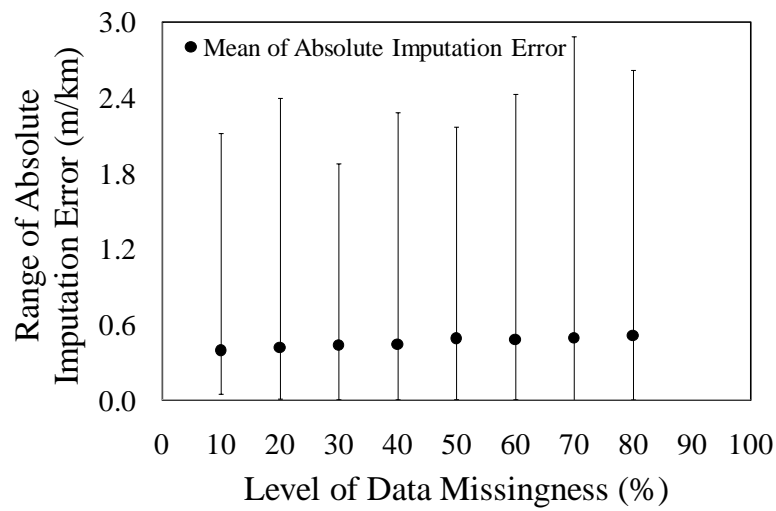


(c) 25% missing data for Road Section ID 25-1002

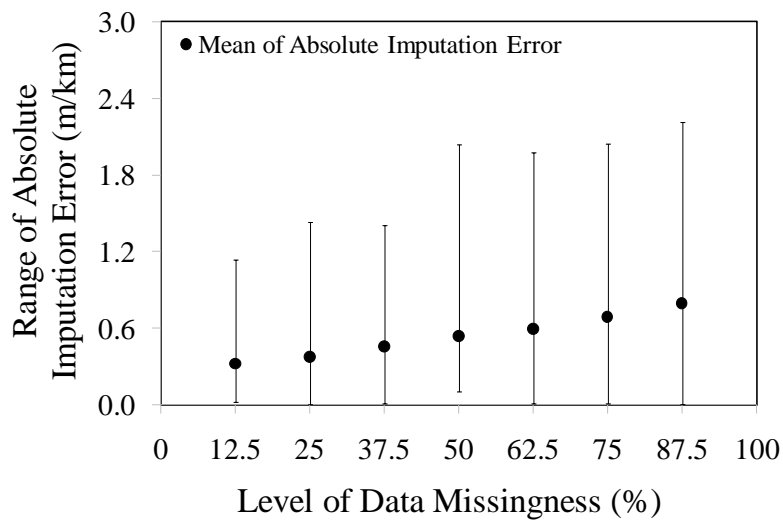
**FIGURE 5 Mean and Ranges of Errors of Imputation Results for Road Sections Studied**



(a) Road section ID 28-1802

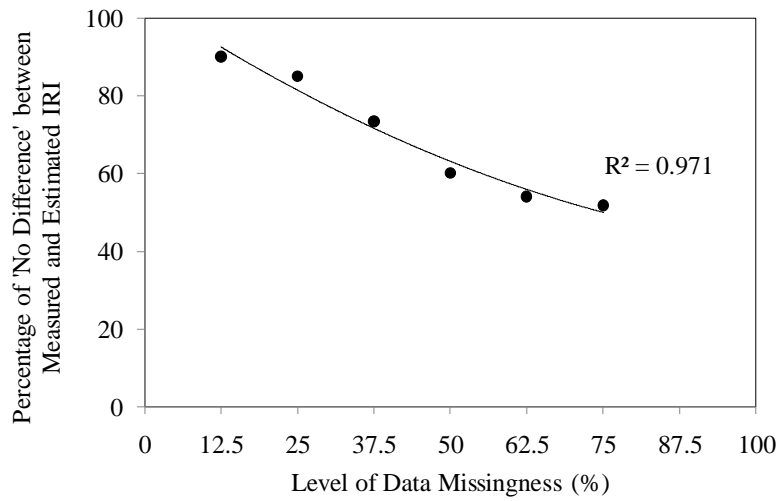


(b) Road section ID 20-1005

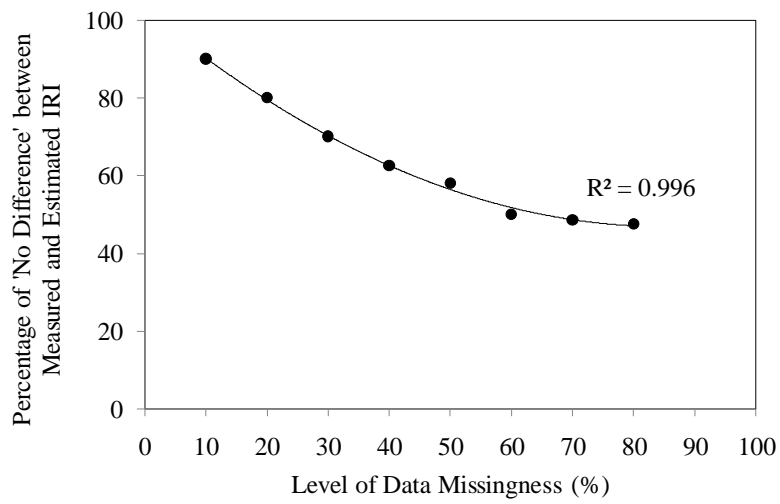


(c) Road section ID 25-1002

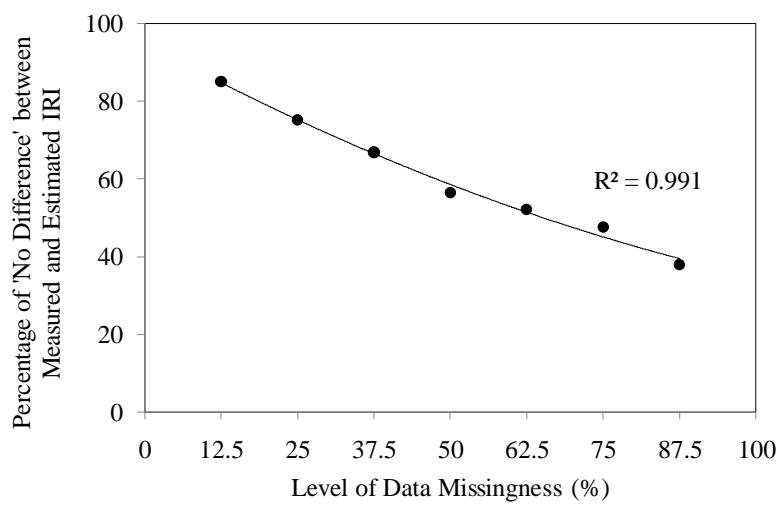
**FIGURE 6 Mean Errors of Imputation Data against Level of Data Missingness**



(a) Road section ID 28-1802



(b) Road section ID 20-1005



(b) Road section ID 25-1002

**FIGURE 7 Effect of Proportion of Missing Data on Imputation Results**

