# Pavement Performance Data Processing Involving Missing Data Records

Farhan JAVED[a], Tien Fang FWA [b]
[a,b] *Department of Civil and Environmental Engineering, National University of Singapore, Singapore, 119260*
[a] *E-mail: farhan@nus.edu.sg*
[b] *E-mail: ceefwatf@nus.edu.sg*

**Abstract**: Pavement performance data are essential for a network level pavement management system and serve as the foundation for an effective decision making process. Therefore, it is essential to ensure a complete pavement condition and performance database for effective decision making process. Until recently, the methods used for analyzing incomplete data have focused on ignoring or removing missing data points, either by deleting those records with incomplete information or by substituting the missing data with some form of estimated mean values. These methods, though simple to implement, at times distort the actual trend of pavement performance, therefore this paper proposes a Multiple Imputation approach to impute missing pavement data. Pavement rut depth data are used in this study for illustration. It is concluded that the proposed Stochastic Multiple Imputation method out-performed the conventional methods in handling missing pavement performance data, and provides an effective approach to impute missing data required in pavement management system.

*Keywords*: Pavement Performance Data, Missing Data, Multiple Imputation, Interpolation, Mean substitution, Regression.

## 1. INTRODUCTION

Any pavement management system is required to have an efficient pavement condition and performance data collection program to its support decision making process. In order to ensure that collected data meet the needs of pavement management decision making process, quality management programs have been developed by several highway agencies (Larson and Forma, 2007; Keleman et al., 2003; NCHRP, 2004). One of the most essential components of data quality management program is its quality assurance process, which includes profiling the data to identify inconsistencies, removal of outliers, and imputation of missing data. Missing data imputation has been widely applied, since highway agencies are heavily relying on data driven applications which requires appropriate treatments to handle empty data cells or missing records in performance database.

Missing data in databases has been one of the most prevalent problems in pavement management systems (Amado and Bernhardt, 2002). According to NCHRP (2009), 61 percent of the highway agencies reported employing software routines to check for missing data elements, and some agencies reported mitigating missing data issues through recollection (Lindly et al., 2005). Zhang and Smadi (2009) listed various data quality checks in Iowa Department of Transportation including missing data. While statistical techniques for imputation of missing data are well developed, their performance to the imputation of pavement management data is unclear. Therefore, this paper proposes a multiple imputation approach to impute missing pavement rut data, and analyses feasibility and applicability of the approach in comparison to existing imputation techniques. The following section presents

a brief overview of the existing imputation procedures which have been implemented in various other fields.

## 2. EXISTING DATA IMPUTATION METHODOLOGIES

Several approaches have been employed for the purpose of dealing with missing data, and range from deletion methods to least square approximation approaches as briefly discussed below.

### 2.1 Deletion Methods

This is by far the most common approach involving neglecting cases with missing data and to run analyses on remaining data. This leads to a loss of reliability as the available sample size for potential analyses is reduced, although it produces unbiased parameter estimates in the case where the data is missing at random. Several works (Allison, 2001; Little and Rubin, 2002; Bennett, 2004) have demonstrated the implications of simply removing cases using the listwise deletion method (LD) on the original data set.

### 2.2 Mean Substitution

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

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

### 2.3 Interpolation using Adjacent Data Points

The missing data are computed by interpolation from the adjacent available data points, which graphically amounts to substituting missing data by connecting with a straight line the point just prior to the missing data with the point just following the missing data. This method assumes a linear correlation in the data, that is, that each observation is to some extent related to and therefore most similar to the previous observation. Yang et al. (2003) applied this approach in forecasting pavement condition rating in Texas. Bennett (2004) suggested this as one of the possible approaches to imputing pavement condition data, and is represented by the following equation in case of three data points ($x_1$,$y_1$), ($x_2$,$y_2$) and ($x_3$,$y_3$),

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

### 2.4 Regression Substitution

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

with the so called noise term $\varepsilon_i$ that adds noise to the linear relationship between the dependent variable and regressors. Thus the model takes the form

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

where $'$ denotes the transpose, so that $x_i'\beta$ is the inner product between vectors $x_i$ and $\beta$.

### 2.5 Expectation Maximization Algorithm

The Expectation Maximization Algorithm (EM) is an iterative regression technique in which the missing variables are regressed on the available data and any additional variables provided as inputs to the algorithm. First, a vector of means and a covariance matrix are calculated using all available data. The means are then imputed for missing values in each variable which serve as a starting value for the imputation. Next, variables with missing values are regressed on all the other available variables. The imputed mean values are then replaced with estimates calculated from the regression equations, and the means and covariances are recalculated. Regression equations and imputations are iteratively calculated, and the process continues until the mean and covariance matrix values converge (Allison, 2009; Little and Rubin, 2002).

### 2.6 Limitation of Existing Methods

A review of the literature indicates that the effectiveness of the data imputation methods relies strongly on the problem domain such as the number of cases, number of variables, and patterns of missing data (Schafer, 1997; Rubin, 1987). Hence, this paper compares existing data imputation approaches in the context of pavement management, and proposes a multiple imputation approach in resolving missing data issue.

## 3. CONCEPT OF MULTIPLE IMPUTATION APPROACH FOR MISSING DATA

Multiple Imputation is a technique in which the missing values are replaced by $m > 1$ plausible values drawn from their predictive distribution. The variation among the $m$ imputations reflects the uncertainty with which the missing values can be predicted from the observed ones. As a result, there are $m$ complete data sets. Rubin (1987) identified that an important limitation of single imputation methods is that "standard variance formulas applied to the filled-in data systematically underestimated the variance of estimates"

In Rubin's method for multiple imputed inference (1987), each of the simulated complete data sets is analyzed by standard statistical methods, and the results (estimates and standard errors) are combined to produce estimates and confidence intervals incorporating missing data uncertainty. The technique is performed using Data Augmentation (DA) Algorithm (Tanner and Wong, 1987), however Expected Maximization Algorithm is considered a preferred approach in establishing initial estimates such as mean and covariance for DA to begin with (Schafer, 1997).

### 3.1 Expectation Maximization Algorithm (EM)

Dempster et al. (1977) published a paper titled Maximum Likelihood from Incomplete Data via the "EM" Algorithm. In this paper they presented an iterative regression technique for calculating descriptive statistics on a data set with missing values in such a way that statistical inferences could still be made from the data. The EM algorithm is a general method for obtaining maximum likelihood estimates of parameters in problems with incomplete data.

Consider an incomplete data matrix with the observed data defined as $Y_{obs}$, missing data as $Y_{mis}$, and a vector of parameters as θ. Hence, complete data, $Y_{com}$, can be defined as $Y_{com} = (Y_{obs}, Y_{mis})$. With the complete data log-likelihood function, $L(\theta) = f(Y_{com}|\theta)$ and the observed data log-likelihood function, $L(\theta) = f(Y_{obs}|\theta)$, the expected complete data log-likelihood function can be defined as,

$$Q(\theta|\theta') = E\{ln[f(Y_{com}|\theta)]Y_{obs}, \theta'\} \quad (4)$$

The EM algorithm begins with some value of $\theta$ and alternates between two steps (Ripley, 1996) as follows:

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

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

The increase in the log-likelihood function $L(\theta)$ is observed with each iteration of the EM algorithm until convergence (Dempster, 1977), and the rate of convergence is proportional to the amount of unobserved or missing information in a data matrix (Fraley, 1999).

### 3.2 Data Augmentation Algorithm (DA)

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

Data augmentation (DA) process is similar in nature to that of EM algorithm i.e. an iterative process which alternately fills in the missing data while crafting inferences about the unknown parameters, however in contrast to the EM algorithm; this is performed in a stochastic manner (Schafer, 1998). A random imputation of missing data under assumed values of the parameters is performed by DA, followed by estimating of new parameters from a Bayesian posterior distribution based on the observed and imputed data. Beginning at some value of $\theta$, each iteration of the DA algorithm alternates between two steps as follows:

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

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

This process of alternately imputing and establishing missing data and parameters respectively creates a Markov chain that finally converges in distribution (Schafer, 1998).

## 4. PROPOSED MULTIPLE IMPUTATION PROCEDURE FOR PAVEMENT MISSING DATA

The rut data of 4 pavement segments, with measurements taken at 10m intervals, were considered for the purpose of illustration, and records were randomly deleted to create a pattern of missing completely at random (MCAR) data at the rate of 45%. Missing completely at random is representative of a scenario where the missing observations simply represent a random sample from within all observations in the dataset. Since no structural association exists between missing and observed data, missing values do not alter the original distributional relationships between variables. A graphical representation of the collected rut data missing data pattern is depicted in Fig. 1 and Table 1 respectively.

The basic step in the Multiple Imputation method is to create values to be substituted for the missing data, therefore a need arises to identify some model which will allow to create

imputes based on auxiliary or other variables in the data set. Under the multivariate normal imputation model, the imputation of an observation is based on regressing a variable, with missing data, on the other variables in the dataset. Since regression method is used to impute the values for the missing data, the imputation model is selected to be rich enough to preserve the associations or relationships among variables. For instance, the rut data for the left wheelpath are included as an auxiliary variable to impute missing rut data for the right wheelpath and vice versa. The model would take the form as follows,

$$r_i = \alpha l_i + \beta x_i + \gamma + s_{rlx}\varepsilon_i \quad (5)$$

where $r$ and $l$ represent the right and left wheelpath rut variables respectively, $x$ is the location of measurement from any reference point, $\varepsilon$ is a random draw that follows the Normal (0,1) distribution, $s_{rlx}$ is the square root of mean square error, $\alpha$, $\beta$, and $\gamma$ are the calibration constants, and $i$ represents the number of data points in the dataset concerned.

Following the imputation procedure just described, the imputed value will contain a random error component. Each time imputation is performed a slightly different result will be obtained, followed by estimating of new parameters from a Bayesian posterior distribution based on the observed and imputed data (Schafer and Rubin, 1998). The procedure of the Multiple Imputation method adopted in the study involves the following steps:

*Step I: Data Transformation* – Transform the data for all variables to approximately normal before imputation using a logit, log or square root transformation function. Next, transform back to their original scale after imputation. The logit or logistic transformation (Hill and Lewicki, 1992) is defined as:

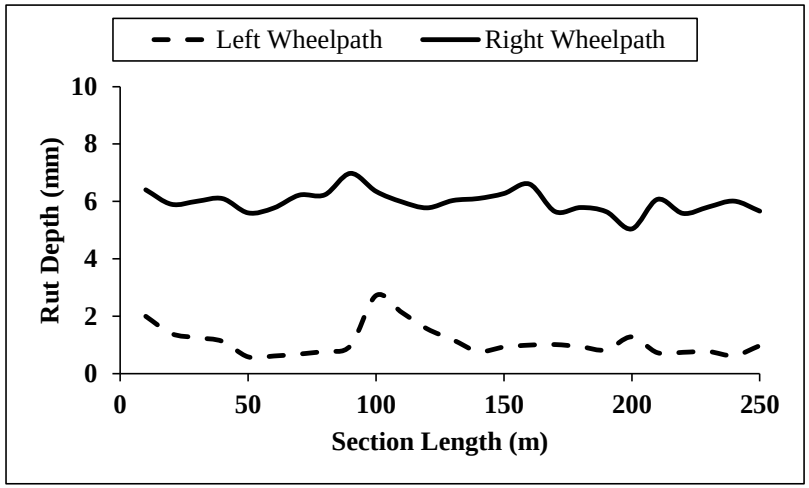
$$\log it(p) = \log\left(\frac{p}{1-p}\right) \quad (6)$$

where $p$ stands for probability or proportion. In the case of elevation profile, a constant value to the data prior to applying the log transformation can be added in order to handle negative values.

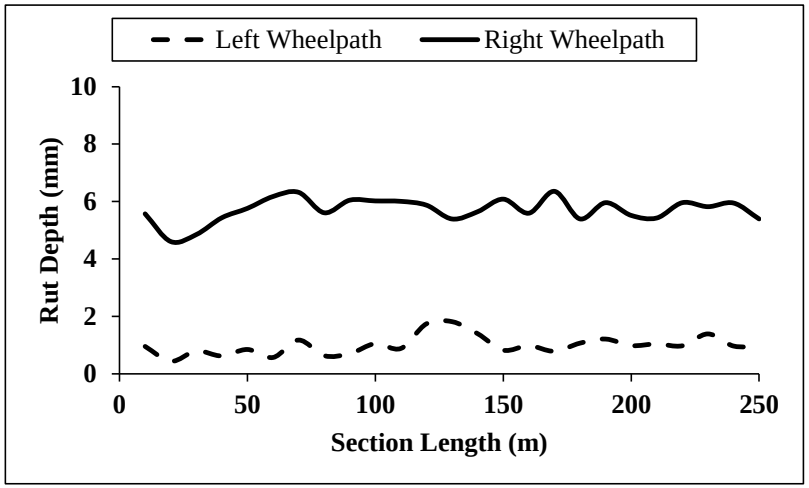
*Step II: Imputation using EM* – Generate estimates of missing values for the data matrix using the EM algorithm with the convergence criterion that the maximum relative parameter change in the value of any parameter during iterative process is less than 0.0001.

*Step III: Imputation using DA* – With the initial parameter estimates from the EM algorithm serving as the basis for the DA algorithm, generate imputed data and new parameter estimates, as explained in the preceding section. The commonly adopted practice of 10 imputations (Little and Rubin, 1987; Schafer, 1997) is applied in this study.

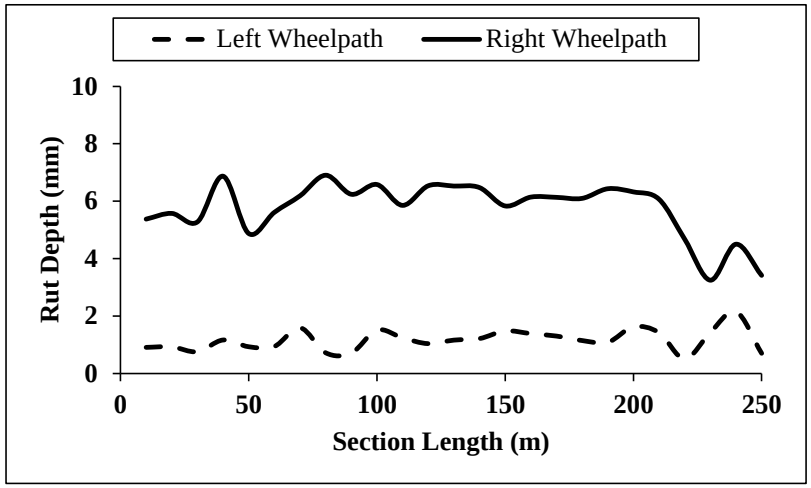
*Step IV: Synthesis of Estimates* – Average over the multiple estimates to obtain the final set of estimates (Rubin, 1987).



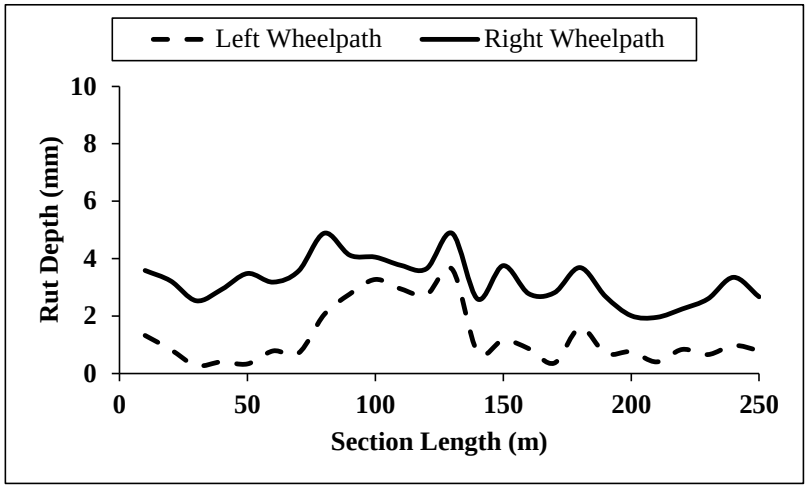

**(a) Rut data for pavement segment 1**




**(b) Rut data for pavement segment 2**




**(c) Rut data for pavement segment 3**




**(d) Rut data for pavement segment 4**

Figure 1. Pavement rut data for various segments considered.

Table 1. State of missing rut data and pattern in pavement database

| Pavement Segment | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | 2 | | 3 | | 4 | |
| Rut Depth (mm) | | Rut Depth (mm) | | Rut Depth (mm) | | Rut Depth (mm) | |
| Left | Right | Left | Right | Left | Right | Left | Right |
| 1.995 | 6.404 | 0.952 | 5.571 | 0.908 | 5.376 | 1.322 | 3.586 |
| | | | | | 5.575 | | 3.221 |
| | | | 4.849 | 0.758 | 5.279 | 0.275 | |
| 1.122 | 6.098 | 0.621 | 5.436 | | | | |
| | 5.6 | 0.846 | | | | 0.332 | 3.482 |
| | | | 6.175 | 0.942 | | 0.784 | 3.181 |
| | | | | 1.593 | 6.187 | | 3.57 |
| | 6.231 | 0.63 | 5.608 | 0.72 | | 2.053 | 4.884 |
| 0.969 | | 0.7 | | 0.709 | | | |
| | | | | | 6.581 | | 4.052 |
| 2.136 | 5.988 | 0.883 | | 1.239 | 5.854 | | |
| 1.558 | | | 5.872 | | | 2.746 | 3.649 |
| | | | | | | 3.632 | |
| | 6.102 | 1.396 | | 1.214 | | | |
| | | 0.823 | 6.083 | 1.481 | 5.834 | 1.158 | 3.755 |
| 0.992 | 6.604 | 0.977 | 5.593 | 1.388 | | 0.862 | |
| | 5.645 | | 6.355 | | 6.134 | | |
| 0.935 | 5.787 | | 5.394 | 1.146 | 6.098 | 1.591 | 3.687 |
| | | 1.209 | | | | 0.698 | 2.678 |
| | | | 5.516 | 1.622 | | | 2.012 |
| 0.727 | | | 5.426 | | | 0.403 | |
| | 5.582 | 0.975 | | 0.508 | | | 2.247 |
| | | 1.388 | 5.82 | 1.429 | 3.248 | | 2.597 |
| | | | 5.948 | 2.15 | 4.502 | 0.974 | |
| 0.973 | 5.662 | 0.967 | 5.393 | 0.702 | 3.416 | 0.797 | 2.67 |

## 5. EVALUATION OF EXISTING IMPUTATION STRATEGIES AGAINST THE PROPOSED MULTIPLE IMPUTATION APPROACH

In this section, the performance of the proposed Multiple Imputation approach, in estimating missing pavement rut data, is compared against the following three existing imputation methods: (i) the substitution by mean method, (ii) the substitution by linear interpolation/extrapolation using adjacent points, and (iii) the substitution by regression method.

### 5.1 Evaluation Concept

There are a number of measures (Armstrong and Collopy, 1992) that can be used to assess the imputation capability of existing and proposed method such as root mean square error (RMSE), based on squared errors, as shown in Eq. (*4*). In addition, measures based on absolute error such as the mean absolute percentage error (MAPE) displayed in Eq. (*5*) are predominantly employed. Since MAPE is to some extent scale dependent such as when

evaluating very low values or integers (i.e. a value of one or two), the size of the measure can be easily inflated. Therefore, it is best to use a combination of measures to evaluate the accuracy of imputation. Both RMSE and MAPE, as in Eq. (*4*) and (*5*) respectively, are employed to compare the accuracy of missing data imputation using different approaches.

$$RMSE = \sqrt{\sum_{t=1}^{n} (O_t - I_t)^2 \Big/ n} \tag{7}$$

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{O_t - I_t}{O_t} \right| \tag{8}$$

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

Table 2. Estimated pavement rut data using Multiple Imputation

| **Pavement Segment** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **1** | | **2** | | **3** | | **4** | |
| **Rut Depth (mm)** | | **Rut Depth (mm)** | | **Rut Depth (mm)** | | **Rut Depth (mm)** | |
| **Left** | **Right** | **Left** | **Right** | **Left** | **Right** | **Left** | **Right** |
| 1.995 | 6.404 | 0.952 | 5.571 | 0.908 | 5.376 | 1.322 | 3.586 |
| 1.123 | 6.397 | 0.424 | 4.578 | 0.906 | 5.575 | 0.895 | 3.221 |
| 1.313 | 6.065 | 0.717 | 4.849 | 0.758 | 5.279 | 0.275 | 3.239 |
| 1.122 | 6.098 | 0.621 | 5.436 | 0.938 | 5.785 | 0.433 | 3.231 |
| 0.589 | 5.6 | 0.846 | 5.46 | 0.986 | 4.121 | 0.332 | 3.482 |
| 0.731 | 5.971 | 0.769 | 6.175 | 0.942 | 5.718 | 0.784 | 3.181 |
| 0.614 | 5.734 | 1.105 | 6.352 | 1.593 | 6.187 | 0.566 | 3.57 |
| 0.739 | 6.231 | 0.63 | 5.608 | 0.72 | 6.541 | 2.053 | 4.884 |
| 0.969 | 6.769 | 0.7 | 5.666 | 0.709 | 6.741 | 2.531 | 3.996 |
| 2.368 | 6.099 | 0.928 | 5.576 | 1.638 | 6.581 | 3.769 | 4.052 |
| 2.136 | 5.988 | 0.883 | 5.652 | 1.239 | 5.854 | 2.815 | 4.382 |
| 1.558 | 6.057 | 1.91 | 5.872 | 1.093 | 6.24 | 2.746 | 3.649 |
| 1.307 | 6.031 | 1.801 | 5.351 | 1.277 | 6.317 | 3.632 | 4.117 |
| 0.554 | 6.102 | 1.396 | 5.542 | 1.214 | 6.421 | 0.777 | 2.899 |
| 1.103 | 5.753 | 0.823 | 6.083 | 1.481 | 5.834 | 1.158 | 3.755 |
| 0.992 | 6.604 | 0.977 | 5.593 | 1.388 | 6.794 | 0.862 | 2.885 |
| 1.195 | 5.645 | 0.85 | 6.355 | 1.235 | 6.134 | 0.349 | 3.105 |
| 0.935 | 5.787 | 1.208 | 5.394 | 1.146 | 6.098 | 1.591 | 3.687 |
| 1.122 | 5.783 | 1.209 | 5.669 | 1.106 | 6.04 | 0.698 | 2.678 |
| 1.525 | 5.037 | 0.99 | 5.516 | 1.622 | 6.905 | 0.416 | 2.012 |
| 0.727 | 5.811 | 1.089 | 5.426 | 1.34 | 6.61 | 0.403 | 1.691 |
| 0.956 | 5.582 | 0.975 | 6.068 | 0.508 | 4.135 | 0.829 | 2.247 |
| 0.952 | 5.659 | 1.388 | 5.82 | 1.429 | 3.248 | 0.631 | 2.597 |
| 0.568 | 5.729 | 1.122 | 5.948 | 2.15 | 4.502 | 0.974 | 3.461 |
| 0.973 | 5.662 | 0.967 | 5.393 | 0.702 | 3.416 | 0.797 | 2.67 |

Table 3. Estimated pavement rut data using Interpolation

| Pavement Segment | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | 2 | | 3 | | 4 | |
| Rut Depth (mm) | | Rut Depth (mm) | | Rut Depth (mm) | | Rut Depth (mm) | |
| Left | Right | Left | Right | Left | Right | Left | Right |
| 1.995 | 6.404 | 0.952 | 5.571 | 0.908 | 5.376 | 1.322 | 3.586 |
| 1.704 | 6.302 | 0.842 | 5.210 | 0.833 | 5.575 | 0.799 | 3.221 |
| 1.413 | 6.200 | 0.731 | 4.849 | 0.758 | 5.279 | 0.275 | 3.308 |
| 1.122 | 6.098 | 0.621 | 5.436 | 0.819 | 5.506 | 0.304 | 3.395 |
| 1.091 | 5.600 | 0.846 | 5.806 | 0.881 | 5.733 | 0.332 | 3.482 |
| 1.061 | 5.810 | 0.774 | 6.175 | 0.942 | 5.960 | 0.784 | 3.181 |
| 1.030 | 6.021 | 0.702 | 5.892 | 1.593 | 6.187 | 1.419 | 3.570 |
| 1.000 | 6.231 | 0.630 | 5.608 | 0.720 | 6.318 | 2.053 | 4.884 |
| 0.969 | 6.150 | 0.700 | 5.674 | 0.709 | 6.450 | 2.226 | 4.468 |
| 1.553 | 6.069 | 0.792 | 5.740 | 0.974 | 6.581 | 2.400 | 4.052 |
| 2.136 | 5.988 | 0.883 | 5.806 | 1.239 | 5.854 | 2.573 | 3.851 |
| 1.558 | 6.026 | 1.054 | 5.872 | 1.231 | 5.849 | 2.746 | 3.649 |
| 1.417 | 6.064 | 1.225 | 5.942 | 1.222 | 5.844 | 3.632 | 3.684 |
| 1.275 | 6.102 | 1.396 | 6.013 | 1.214 | 5.839 | 2.395 | 3.720 |
| 1.134 | 6.353 | 0.823 | 6.083 | 1.481 | 5.834 | 1.158 | 3.755 |
| 0.992 | 6.604 | 0.977 | 5.593 | 1.388 | 5.984 | 0.862 | 3.732 |
| 0.964 | 5.645 | 1.054 | 6.355 | 1.267 | 6.134 | 1.227 | 3.710 |
| 0.935 | 5.787 | 1.132 | 5.394 | 1.146 | 6.098 | 1.591 | 3.687 |
| 0.866 | 5.736 | 1.209 | 5.455 | 1.384 | 5.528 | 0.698 | 2.678 |
| 0.796 | 5.685 | 1.131 | 5.516 | 1.622 | 4.958 | 0.551 | 2.012 |
| 0.727 | 5.633 | 1.053 | 5.426 | 1.065 | 4.388 | 0.403 | 2.130 |
| 0.789 | 5.582 | 0.975 | 5.623 | 0.508 | 3.818 | 0.593 | 2.247 |
| 0.850 | 5.609 | 1.388 | 5.820 | 1.429 | 3.248 | 0.784 | 2.597 |
| 0.912 | 5.635 | 1.178 | 5.948 | 2.150 | 4.502 | 0.974 | 2.634 |
| 0.973 | 5.662 | 0.967 | 5.393 | 0.702 | 3.416 | 0.797 | 2.670 |

Table 4. Estimated pavement rut data using Regression substitution

| Pavement Segment | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | 2 | | 3 | | 4 | |
| Rut Depth (mm) | | Rut Depth (mm) | | Rut Depth (mm) | | Rut Depth (mm) | |
| Left | Right | Left | Right | Left | Right | Left | Right |
| 1.995 | 6.404 | 0.952 | 5.571 | 0.908 | 5.376 | 1.322 | 3.586 |
| 1.687 | 6.194 | 0.743 | 5.559 | 0.936 | 5.575 | 1.320 | 3.221 |
| 1.649 | 6.173 | 0.762 | 4.849 | 0.758 | 5.279 | 0.275 | 3.533 |
| 1.122 | 6.098 | 0.621 | 5.436 | 0.974 | 5.968 | 1.309 | 3.777 |
| 0.572 | 5.600 | 0.846 | 5.586 | 0.992 | 5.899 | 0.332 | 3.482 |
| 1.534 | 6.112 | 0.817 | 6.175 | 0.942 | 5.829 | 0.784 | 3.181 |
| 1.496 | 6.092 | 0.836 | 5.604 | 1.593 | 6.187 | 1.293 | 3.570 |
| 1.458 | 6.231 | 0.630 | 5.608 | 0.720 | 5.689 | 2.053 | 4.884 |
| 0.969 | 6.051 | 0.700 | 5.622 | 0.709 | 5.620 | 1.282 | 3.500 |
| 1.382 | 6.030 | 0.891 | 5.631 | 1.085 | 6.581 | 1.277 | 4.052 |
| 2.136 | 5.988 | 0.883 | 5.640 | 1.239 | 5.854 | 1.271 | 3.388 |
| 1.558 | 5.990 | 0.928 | 5.872 | 1.122 | 5.410 | 2.746 | 3.649 |
| 1.267 | 5.969 | 0.947 | 5.658 | 1.141 | 5.340 | 3.632 | 3.277 |
| 0.729 | 6.102 | 1.396 | 5.667 | 1.214 | 5.271 | 1.255 | 3.222 |
| 1.191 | 5.929 | 0.823 | 6.083 | 1.481 | 5.834 | 1.158 | 3.755 |
| 0.992 | 6.604 | 0.977 | 5.593 | 1.388 | 5.131 | 0.862 | 3.111 |
| 1.115 | 5.645 | 1.021 | 6.355 | 1.215 | 6.134 | 0.562 | 3.055 |
| 0.935 | 5.787 | 1.040 | 5.394 | 1.146 | 6.098 | 1.591 | 3.687 |
| 1.039 | 5.847 | 1.209 | 5.713 | 1.252 | 4.922 | 0.698 | 2.678 |
| 1.001 | 5.827 | 1.077 | 5.516 | 1.622 | 5.852 | 1.223 | 2.012 |
| 0.727 | 5.806 | 1.095 | 5.426 | 1.289 | 4.782 | 0.403 | 2.833 |
| 0.924 | 5.582 | 0.975 | 5.740 | 0.508 | 4.712 | 1.212 | 2.247 |
| 0.886 | 5.765 | 1.388 | 5.820 | 1.429 | 3.248 | 1.206 | 2.597 |
| 0.848 | 5.745 | 1.151 | 5.948 | 2.150 | 4.502 | 0.974 | 2.666 |
| 0.973 | 5.662 | 0.967 | 5.393 | 0.702 | 3.416 | 0.797 | 2.670 |

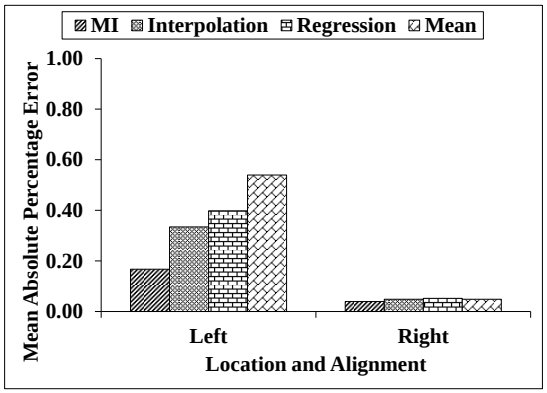
Table 5. Estimated pavement rut data using Mean substitution

| Pavement Segment | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | | 2 | | 3 | | 4 | |
| Rut Depth (mm) | | Rut Depth (mm) | | Rut Depth (mm) | | Rut Depth (mm) | |
| Left | Right | Left | Right | Left | Right | Left | Right |
| 1.995 | 6.404 | 0.952 | 5.571 | 0.908 | 5.376 | 1.322 | 3.586 |
| 1.267 | 5.973 | 0.951 | 5.669 | 1.157 | 5.575 | 1.259 | 3.221 |
| 1.267 | 5.973 | 0.951 | 4.849 | 0.758 | 5.279 | 0.275 | 3.285 |
| 1.122 | 6.098 | 0.621 | 5.436 | 1.157 | 5.340 | 1.259 | 3.285 |
| 1.267 | 5.600 | 0.846 | 5.669 | 1.157 | 5.340 | 0.332 | 3.482 |
| 1.267 | 5.973 | 0.951 | 6.175 | 0.942 | 5.340 | 0.784 | 3.181 |
| 1.267 | 5.973 | 0.951 | 5.669 | 1.593 | 6.187 | 1.259 | 3.570 |
| 1.267 | 6.231 | 0.630 | 5.608 | 0.720 | 5.340 | 2.053 | 4.884 |
| 0.969 | 5.973 | 0.700 | 5.669 | 0.709 | 5.340 | 1.259 | 3.285 |
| 1.267 | 5.973 | 0.951 | 5.669 | 1.157 | 6.581 | 1.259 | 4.052 |
| 2.136 | 5.988 | 0.883 | 5.669 | 1.239 | 5.854 | 1.259 | 3.285 |
| 1.558 | 5.973 | 0.951 | 5.872 | 1.157 | 5.340 | 2.746 | 3.649 |
| 1.267 | 5.973 | 0.951 | 5.669 | 1.157 | 5.340 | 3.632 | 3.285 |
| 1.267 | 6.102 | 1.396 | 5.669 | 1.214 | 5.340 | 1.259 | 3.285 |
| 1.267 | 5.973 | 0.823 | 6.083 | 1.481 | 5.834 | 1.158 | 3.755 |
| 0.992 | 6.604 | 0.977 | 5.593 | 1.388 | 5.340 | 0.862 | 3.285 |
| 1.267 | 5.645 | 0.951 | 6.355 | 1.157 | 6.134 | 1.259 | 3.285 |
| 0.935 | 5.787 | 0.951 | 5.394 | 1.146 | 6.098 | 1.591 | 3.687 |
| 1.267 | 5.973 | 1.209 | 5.669 | 1.157 | 5.340 | 0.698 | 2.678 |
| 1.267 | 5.973 | 0.951 | 5.516 | 1.622 | 5.340 | 1.259 | 2.012 |
| 0.727 | 5.973 | 0.951 | 5.426 | 1.157 | 5.340 | 0.403 | 3.285 |
| 1.267 | 5.582 | 0.975 | 5.669 | 0.508 | 5.340 | 1.259 | 2.247 |
| 1.267 | 5.973 | 1.388 | 5.820 | 1.429 | 3.248 | 1.259 | 2.597 |
| 1.267 | 5.973 | 0.951 | 5.948 | 2.150 | 4.502 | 0.974 | 3.285 |
| 0.973 | 5.662 | 0.967 | 5.393 | 0.702 | 3.416 | 0.797 | 2.670 |

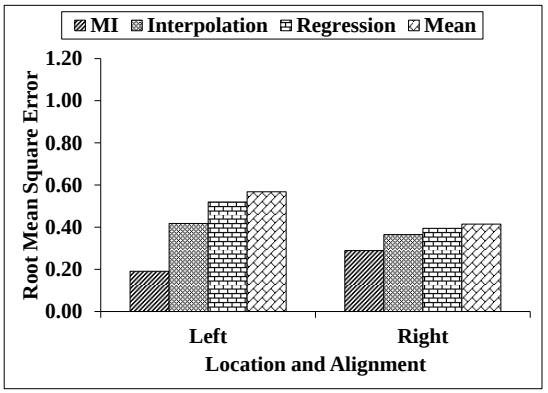
## 5.2 Comparison of Imputation Results

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

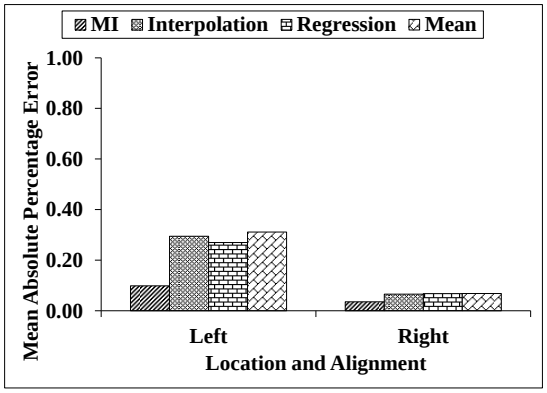
The relative quality of data imputation from the four methods are assessed using mean absolute error in percentage (MAPE) and root mean square error (RMSE), as presented in Fig. 2. As can be seen from the figure, the mean substitution method largely resulted in imputed values with the highest amount of deviations from the observed values, followed by the regression substitution method and the interpolation method. The stochastic Multiple Imputation method, proposed in this study, yielded the smallest errors for rut data for right and left wheelpath.



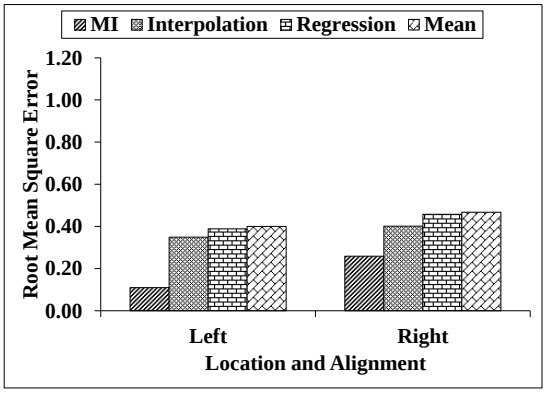

**(a) MAPE of imputed segment 1 rut data**



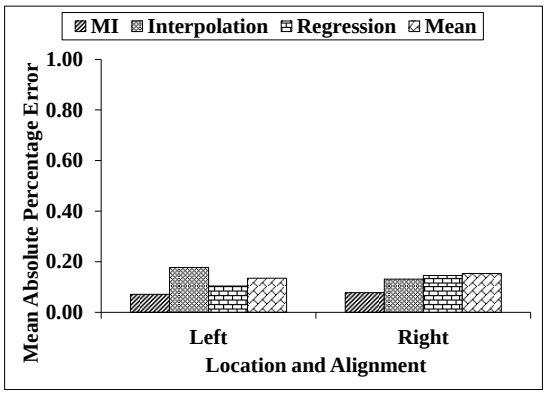

**(b) RMSE of imputed segment 1 rut data**



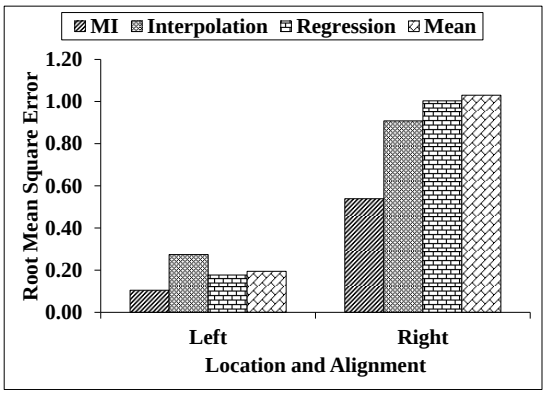

**(c) MAPE of imputed segment 2 rut data**



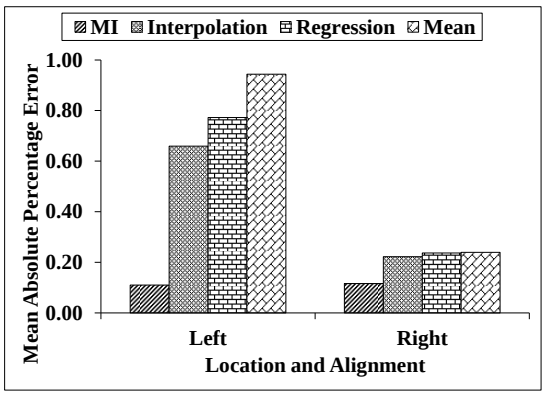

**(d) RMSE of imputed segment 2 rut data**



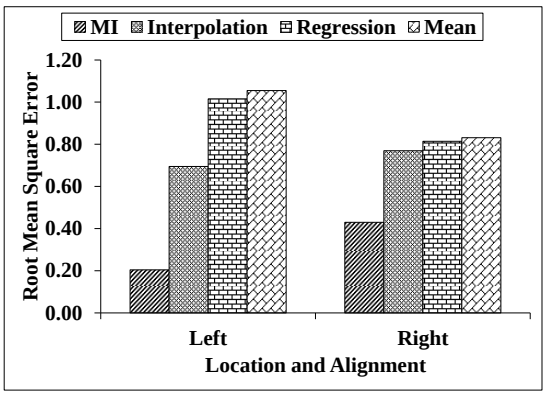

**(e) MAPE of imputed segment 3 rut data**




**(f) RMSE of imputed segment 3 rut data**




**(g) MAPE of imputed segment 4 rut data**




**(h) RMSE of imputed segment 4 rut data**

Figure 2. Appraisal of accuracy of various imputation methods for missing pavement rut data.

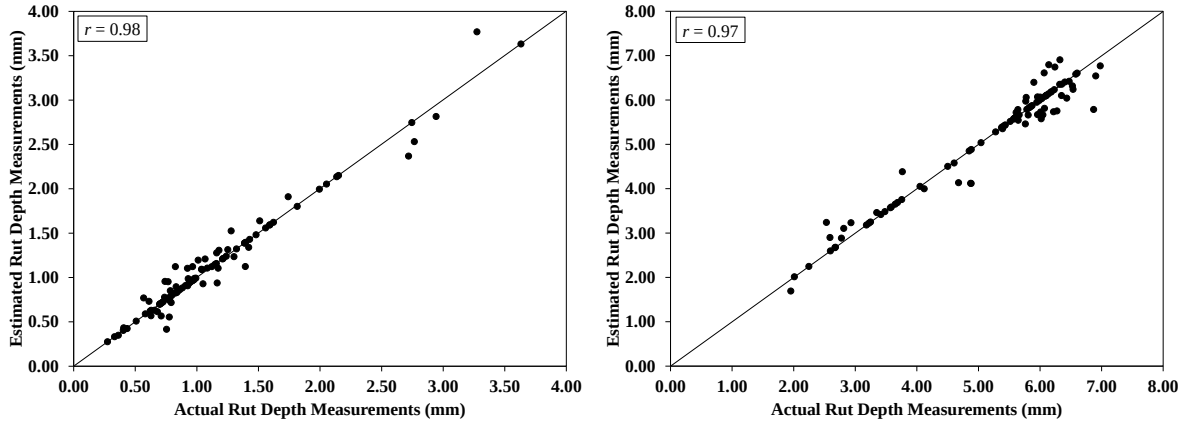
The benefits of employing auxiliary variables in the analysis of the proposed Multiple Imputation approach can be seen from Fig. 3. For instance, left wheelpath served as an auxiliary variable, for imputing rut data of right wheelpath and vice versa, in segment 1 resulting in a reduced MAPE or RMSE value for rut data imputed using Multiple Imputation in comparison to Regression substitution as shown in Fig. 3(a) and 3(b).

The quantitative assessment of the imputation performance of the proposed approach against existing methodologies, in imputing missing pavement rut data, can be measured using the Pearson correlation coefficients *r* (Neter et al., 1990), which reflects the degree of a linear relationship between any two sets of results evaluated as follows,

$$r = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2} \times \sqrt{n(\sum y_i^2) - (\sum y_i)^2}} \qquad (9)$$

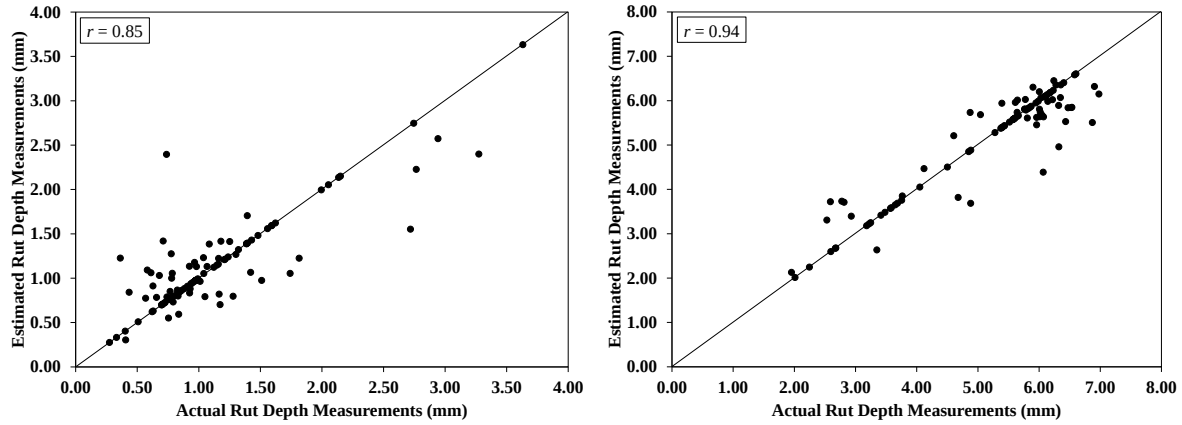
where $x_i$ = value from observation $i$ on variable $X$, $y_i$ = value from observation $i$ on variable $Y$, $n$ = number of values in each data set, $i = 1,\ldots,n$.

The degree of correlation between actual and estimated rut depth measurements, obtained by imputing missing data using the proposed Multiple Imputation approach, for left and right wheelpath is 0.98 and 0.97 respectively. The correlation between actual and estimated rut depth measurements, obtained using substitution by mean, interpolation, and regression, for left wheelpath is 0.85, 0.72, and 0.69 respectively, while that of right wheelpath is 0.94, 0.93, and 0.92 respectively. The correlation results are consistent with the MAPE and RMSE results, in the preceding section, signifying superior and robust performance of the proposed MI approach for missing data imputation of pavement rut data.



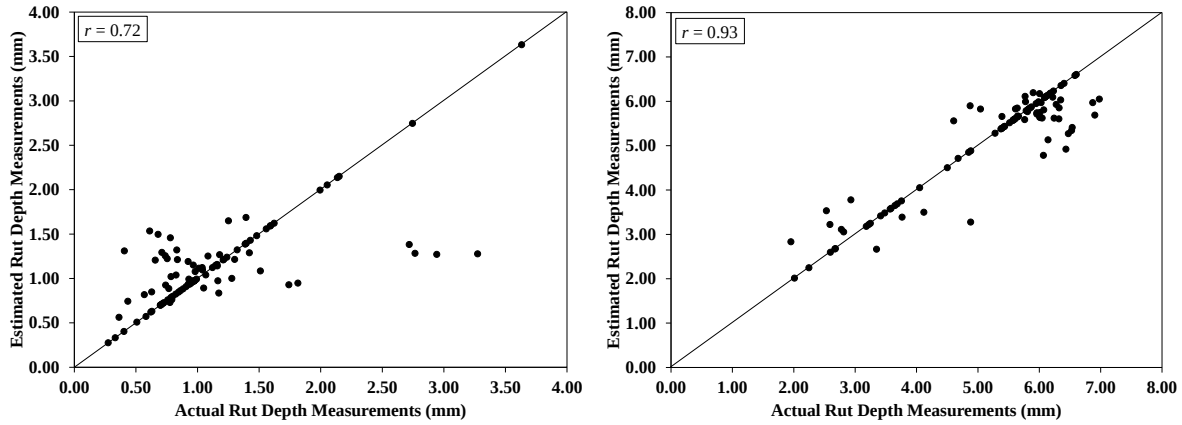

**(i) Left Wheelpath** **(ii) Right Wheelpath**

**(a) Substitution by Multiple Imputation**



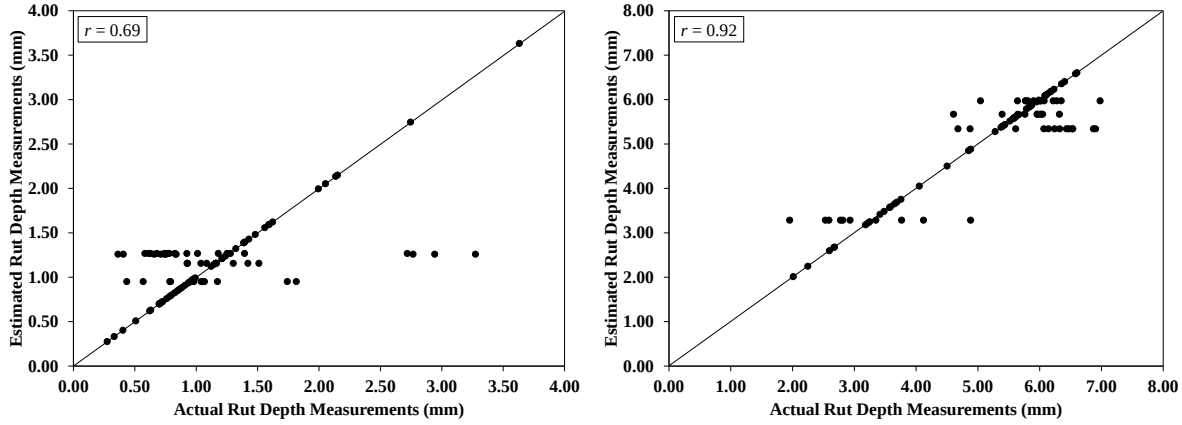

**(i) Left Wheelpath** **(ii) Right Wheelpath**

**(b) Substitution by Interpolation**




**(i) Left Wheelpath** **(ii) Right Wheelpath**

**(c) Substitution by Regression**




**(i) Left Wheelpath** **(ii) Right Wheelpath**

**(d) Substitution by Mean**

Figure 3. Scatter plots between imputed and actual data for all segments.

## 6. CONCLUSIONS

This paper has proposed a Multiple Imputation approach to impute missing rut data in

pavement performance database of a pavement management system. The quality of the imputed data values by the proposed approach was assessed against values obtained using conventional methods, including the listwise deletion method, the mean substitution method, the interpolation method, and the regression substitution method. For illustration, rut data of 4 pavement segment were considered, and were randomly deleted to create different patterns of datasets with missing data. The applicability and relative quality of the proposed approach, in handling missing data was analyzed in comparison to the three existing imputation techniques. The effectiveness of employing auxiliary variables in the pavement rut data imputation models is also demonstrated. The proposed stochastic Multiple Imputation method yielded the smallest errors for the rut data. The mean substitution method resulted in imputed values with the highest amount of deviations from the observed values, followed by the regression substitution method and the interpolation method. Therefore, it is concluded that the proposed Stochastic Multiple Imputation method out-performed the conventional methods in handling missing pavement rut data, and thus providing an effective approach to impute missing data required in a pavement management system.

## REFERENCES


Allison P. D. (2001) *Missing Data*. Sage Publications, Inc., Thousand Oaks, CA.

Amado, V. and Bernhardt, K. L. S. (2002) Knowledge Discovery in Pavement Condition Data. In the 81st Annual Meeting of the Transportation Research Board (TRB), Washington DC.

Armstrong, J. S. and Collopy, F. (1992) Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons. *International Journal of Forecasting*, Vol. 8, 69–80.

Bennett, C. R. (2004) Sectioning of Road Data for Pavement. In the 6th International Conference on Managing Pavements, Queensland, Australia.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Vol. 39, No. 1, 1-38.

Fraley, C. (1999) On Computing the Largest Fraction of Missing Information for the EM Algorithm and the Worst Linear Function for Data Augmentation. *Computational Statistics & Data Analysis*, Vol. 31, 13–26.

Hill, T. and Lewicki, P. (2006) *Statistics: Methods and Applications*, Statsoft, Inc., 652pp

Keleman, M., Henry, S. and Farrokhyar, A. (2003) Pavement Management Manual. Colorado Department of Transportation, Denver, CO.

Larson, C. D. and Forma, E. H. (2007) Application of Analytic Hierarchy Process to Select Project Scope for Video Logging and Pavement Condition Data Collection. *Transportation Research Record*, No. 1990, Transportation Research Board, Washington, DC., 40-47.

Lindly, J. K., Bell, F. and Sharif, U. (2005) Specifying Automated Pavement Condition Surveys. *Journal of Transportation Research Forum*, Vol. 44, No. 3, 19-32.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. 2nd edition, John Wiley, New York.

National Cooperative Highway Research Program (NCHRP) (2004). Automated Pavement Distress Collection Techniques. National Cooperative Highway Research Program Synthesis Report No. 334, Transportation Research Board, Washington, DC.

National Cooperative Highway Research Program (NCHRP) (2009) Quality

Management of Pavement Condition Data Collection. National Cooperative Highway Research Program Synthesis Report No. 401, Transportation Research Board, Washington, DC.

Neter, J., Wasserman, W. and Kutner, M. H. (1990) *Applied linear models: regression, analysis of variance, and experimental designs*. Homewood, Illinois; Richard D. Irwin, Inc., 38-44, 62–104.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, Cambridge.

Rubin, D. B. (1987) *Multiple Imputation for Survey Nonresponse*. Wiley, New York.

Schafer, J. L. and Olsen, M. K. (1998) Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behavioral Research*, Vol. 33, 545–571.

Schafer, J. L. and Rubin, D. B. (1998) Multiple Imputation for Missing Data Problems, Short course presented at Joint Statistical Meetings, Dallas, TX, August.

Tanner, M. A and Wong, W. H. (1987) The Calculation of Posterior Distributions by Data Augmentation. *Journal of American Statistical Association*, Vol. 82, 528–550.

Yang, J., Lu, J. J. and Gunaratne, M. (2003) Application of Neural Network Models for Forecasting of Pavement Crack Index and Pavement Condition Rating. *Transportation Research Record*, No. 1699, Transportation Research Board, Washington, DC, 3-12.

Zhang, L. and Smadi, O. G. (2009) What is Missing in Quality Control of Contracted Pavement Distress Data Collection? In the 90th Annual Meeting of Transportation Research Board, Washington, DC.