Systematic Statistical Approach to Populate Missing Performance Data in Pavement Management Systems

Mazin M. Al-Zou'bi, Ph.D.[1]; Carlos M. Chang, Ph.D., P.E., M.ASCE[2]; Soheil Nazarian, Ph.D., P.E., F.ASCE[3]; and Vladik Kreinovich, Ph.D.[4]

Abstract: Transportation agencies use pavement management systems (PMS) for their maintenance and rehabilitation planning, programming, and budgeting. PMS is used to make decisions regarding when maintenance and rehabilitation should be applied. To support these decisions, it is important to have reliable data on pavement conditions and accurate performance models for predicting pavement condition. The data on pavement condition typically come from regular field surveys resulting in distress, condition, and ride scores. PMS data sets are often incomplete (for some locations and some years) as a result of operational limitations reducing the predictive power of the performance models. Model-free and model-based replacement techniques for estimating missing data points have been designed and successfully used in other application areas like statistics, economics, marketing, medicine, psychometrics, and political science. It is therefore reasonable to apply these methods to the PMS databases. Statistical techniques are assembled and used in a robust approach to systematically analyze the effect of applying these techniques to rebuild missing performance data. As a case study, continuous reinforced concrete pavement (CRCP) sections were selected to test the proposed statistical systematic approach from a pavement management information system (PMIS) maintained by the Texas Department of Transportation (TxDOT). A major effect was observed in the results of predicting the distress scores when applying the developed approach. **DOI: 10.1061/(ASCE)IS.1943-555X.0000247.**

Author keywords: Pavement management systems (PMS); Continuous reinforced concrete pavement (CRCP); Pavement management information system (PMIS); Pavement performance; Pavement performance models; Missing data techniques.

Introduction

Pavement management systems (PMS) are used by transportation agencies for planning, programming, and budgeting short and long-term treatment needs required for preserving pavement networks. Relevant data for developing the pavement performance models for PMS come from field surveys that are conducted regularly to assess the structural and functional conditions of the pavement network.

Pavements are complex physical structures that respond to the influence of numerous environmental, subsurface, and load-related variables and their interactions. Subsequently, the task of predicting the multifaceted responses of pavements to the series of interrelated variables is complex and must be addressed by using a number of assumptions and simplifications.

The accurate prediction of pavement performance is important at all management levels for the efficient preservation of road infrastructure. At the network level, pavement performance prediction is essential for rational budget and resource allocation. At the programming level, pavement performance prediction is needed for adequate activity planning and project prioritization. At the project level, it is required in establishing and designing the necessary corrective maintenance and rehabilitation actions.

In practice, data sets that are maintained in the pavement management systems are often incomplete, affecting the accuracy of pavement performance predictions. The crucial question discussed in this paper is how to improve the accuracy of pavement performance predictions with limited data. This paper presents a systematic approach, using statistical techniques, to handle missing data properly and efficiently for predicting pavement performance. The aim of this approach is to improve pavement performance predictions. An approach to evaluate the effectiveness of various statistical methods in estimating missing values in pavement condition data sets is developed as part of the approach. This approach is validated using the Texas Department of Transportation (TxDOT) continuous reinforced concrete pavements (CRCP) database. CRCP, which is the highest-quality longest-lasting pavement, is the main type of pavement used for road segments with high-volume traffic such as interstate highways. The maintenance of road segments with this type of pavement is of high priority. CRCP has been built in more than 35 states and is used to construct roadways, airport runways, railway tracks, and warehouse floors. The Texas road network length is 195,300 lane miles, and approximately 30% of interstate highways is composed of CRCP.

Historical Overview of Missing Data Estimation Methods

Missing data is a common problem in most research studies. Yet, no commonly agreed upon solution for addressing this problem exists. Consequently, researchers have developed a wide variety of techniques for handling missing data. However, no single technique is without pitfalls. Thus, researchers facing a missing data problem should thoroughly investigate the sources of the missing data,

[1]Research Assistant, Dept. of Civil Engineering, Univ. of Texas at El Paso, El Paso, TX 79968 (corresponding author). E-mail: mmalzoubi@miners.utep.edu

[2]Assistant Professor, Dept. of Civil Engineering, Univ. of Texas at El Paso, El Paso, TX 79968.

[3]Professor, Dept. of Civil Engineering, Univ. of Texas at El Paso, El Paso, TX 79968.

[4]Professor, Dept. of Computer Science, Univ. of Texas at El Paso, El Paso, TX 79968.

Note. This manuscript was submitted on April 14, 2013; approved on December 22, 2014; published online on March 24, 2015. Discussion period open until August 24, 2015; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Infrastructure Systems***, © ASCE, ISSN 1076-0342/04015002(8)/$25.00.

and the options to handle missing data. Nevertheless, most standard statistical methods have been designed to analyze data sets but with no missing data.

There are two alternatives to address missing data: (1) Delete those cases that have missing data; or (2) fill in or input the missing values with estimated values (Anderson et al. 1983). Statistical techniques are used to fill in the missing values in the data sets. The recovery of the sample size and statistical power is a motivational factor in imputing values. A brief historical overview of missing data techniques is summarized, and then the topic of imputation techniques to address the missing data problem is presented.

Before the 1970s, missing data problems were addressed through editing, in which missing variables were logically inferred from additional data that had been observed. The literature on the statistical analysis of data sets with missing data has grown since the early 1970s, spurred by advances in computer technology that made previously laborious numerical calculations a simple matter. A method for inference from incomplete data was only developed in 1976. Immediately afterward, Dempster, Laird, and Rubin developed the expectation maximization (EM) algorithm in 1977 that resulted in the use of the maximum likelihood (ML) methods for missing data estimation.

The general location model of Olkin and Tate (1961) and extensions introduced by Krzanowski (1982) form the basis for methods. Maximum likelihood estimation with incomplete data is achieved by an application of the EM algorithm (Dempster et al. 1977). Special cases of the algorithm include Orchard and Woodbury's (1972) algorithm for incomplete normal samples, Fuchs's (1982) algorithms for log linear modeling of partially classified contingency tables, and Day's (1969) algorithm for multivariate normal mixtures. Applications include (1) imputation of missing values; (2) logistic regression and discriminate analysis with missing predictors and unclassified observations; (3) linear regression with missing continuous and categorical predictors; and (4) parametric cluster analysis with incomplete data.

According to Dempster and Rubin (1983), the idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into a pleasurable state of believing that the data are complete after all. It is dangerous because it lumps together situations in which the problem is sufficiently minor that it can be legitimately handled in this way, and situations in which standard estimators applied to the real imputed data have substantial biases.

According to Raymond (1986), "The most widely used estimation technique is probably the mean substitution method." Because the means are replacing the missing values, the variances and covariance would be downwardly biased (Little and Rubin 1987). Additionally, the confidence interval may not be as precise as expected (Little 1988). In the regression imputation, the imputed data would preserve deviation from the mean and the shape of the distribution (Little 1988). Thus, the imputed data "will not attenuate correlations as much as mean substitution," according to Roth (1994).

Missing data has received no coverage in PMS research. In contrast, certain fields such as marketing (Kaufman 1988), organizational behavior (Roth et al. 1999), economics, statistics, and psychometrics have paid more attention to the issue. Tests and interval estimates from small samples with missing data have had limited development. Very little work has been done on diagnostic tests concerning the validity of models when data are missing and incomplete, or on the robustness of estimates derived from the proposed models (Marwala 2009).

From the literature review, it is concluded that missing data problems are ubiquitous in many practical applications; to address these problems, special statistical missing data techniques have been developed in the past. Although these techniques have had many successful applications in different areas of science and engineering, they have not been structured for their use in pavement management. In this study, a number of promising missing data rebuilding techniques are compared to address the pavement performance prediction problem faced in pavement management. Recommendations on the most efficient techniques to use for pavement management will lead to more accurate predictions of pavement performance scores.

Distress Score As a Measure of Pavement Performance

To predict the future pavement performance, it is necessary to know the current pavement condition, and the future deterioration rate. To estimate how fast the pavement deteriorates, it is desirable to know the pavement condition history.

The structural and material condition of pavement is determined by exhibited distress types (e.g., cracking, faulting, spalling), the severity of these distress types, and the density of these distress types (i.e., extent of occurrence in the surveyed pavement area) (Livneh 1994) (Shahin et al. 1980). To evaluate the pavement condition, TxDOT has combined these characteristics into a single distress index known as the distress score (DS) since the late 1980s. Distress scores range from 1 to 100, with 1 representing the worst pavement and 100 representing a pavement section with no distresses (Stampley and Miller 1995b).

TxDOT uses a *sigmoidal* expression to predict the distress deterioration characteristics (X) as a function of the age of the pavement

$$X = X_0 - \alpha e^{-[(\rho/\mathrm{Age}_i)^{\beta}]} \tag{1}$$

where X_0 = largest possible value; α = maximum loss factor; β = slope factor that determines how fast the road segment deteriorates; and ρ = prolongation factor that controls the location of the distress X curve's inflection point.

The values of the parameters α, β, and ρ are determined based on a statistical analysis to obtain the best fit for the observed data. For predicting the distress score, Eq. (1) takes the form shown in Eq. (2)

$$\mathrm{DS} = 100 - \alpha e^{-[(\rho/\mathrm{Age}_i)^{\beta}]} \tag{2}$$

The general shape of the distress score–pavement performance model is illustrated in Fig. 1.

Distress scores can be clustered in very good, good, fair, poor, and very poor condition categories (Table 1).

Populating Missing Pavement Performance Data

As mentioned previously, PMS data sets are often incomplete, and the pavement performance historical data are missing. The problem of missing data is found in different practical applications; PMS databases frequently face missing data within their data sets in different modules and at various levels. PMS data may be missing from the database because it could not be rated, measured, collected, saved, or managed correctly, which makes the development of pavement performance models difficult.

Missing data creates various problems in PMS applications and affects pavement performance predictions, treatment selection, and budget estimates. The difficulty of modeling pavement performance with missing data also relies on the fact that in many cases maintenance and rehabilitation (M&R) records are also not available. As a result, it is difficult to identify when a sudden

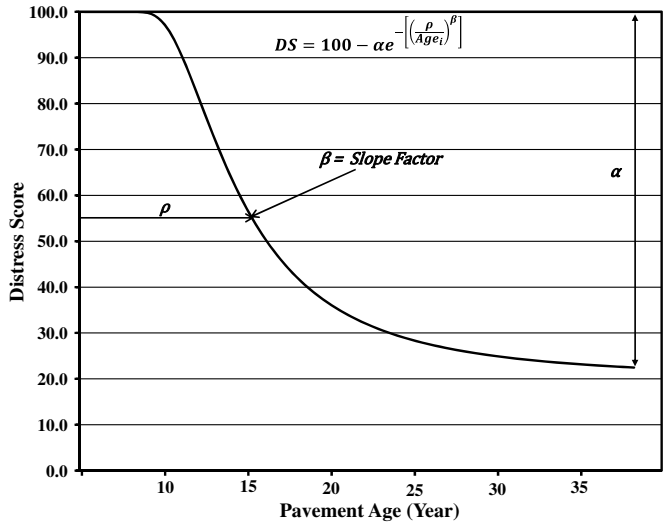

Fig. 1. General shape of distress score pavement performance model (reprinted from Al-Zou'bi 2013)

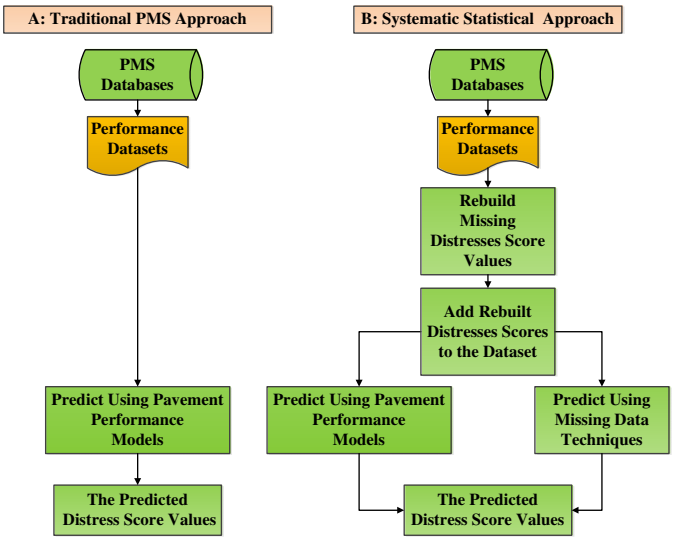

Fig. 2. Schematic of traditional versus systematic statistical approach to predict distress scores

Table 1. PMIS Distress Score Categories

Classification	Distress score
Very good	90–100
Good	80–89
Fair	70–79
Poor	60–69
Very poor	1–59

increase in condition is the result of M&R intervention or random noise.

In 2008, Chu and Durango proposed intervention analysis and state-space specifications of auto-regressive moving averages for the development of pavement deterioration and inspection models. It was concluded from this study that this approach could be used in situations in which there are extensive time series of data (Chu and Durango 2008). In 2010, Hong and Prozzi presented a roughness nonlinear pavement deterioration model to account for the heterogeneity of pavement performance data. This approach also requires time-series data, and a Bayesian approach is used to obtain individual-level parameters for the modeling (Hong and Prozzi 2010). Most recently, Aguiar-Moya proposed in 2011 a performance modeling technique to capture the deterioration rate in performance models without filtering points with maintenance intervention. The method is able to provide the probability of maintenance intervention for each observation based on a Bayesian-based algorithm that can distinguish maintenance or no maintenance (Gao et al. 2011).

In many situations, the accuracy is improved by first rebuilding the missing data points and then using the rebuilt data sets for the predictions (Tsikriktsis 2005). As an alternative to the methods that require large time-series data sets, a systematic statistical approach is described in this paper to improve pavement performance predictions. This approach complements traditional methods by filling in the missing data in the historical data sets before modeling. Then, future pavement performance predictions are treated as missing data points to be rebuilt.

Fig. 2 shows a schematic diagram of predicting the distress score by using the traditional PMS approach and by using the systematic statistical approach to populate (rebuild) missing pavement performance data. In the traditional approach, Eq. (2) is used to predict distress scores without rebuilding any missing values in the data-set history.

In the proposed systematic statistical approach, first, the missing data points are rebuilt by using known missing data techniques, and then, the prediction is performed based both on the observed data points and on the rebuilt data points for the purpose of comparison. Two methods are compared when predicting distress scores: a method that uses Eq. (2) and a method that uses missing data techniques. In the first prediction method, nonlinear regression techniques are used to find α, β, and ρ parameters for which this model provides the best fit, and then used to predict the distress score at future year(s). In the second prediction method, a missing data technique is used both for rebuilding the missing data points and for predicting distress scores.

The hypothesis of the study is that rebuilding the missing data should improve the prediction of pavement performance. It is also hypothesized that certain statistical techniques should be more efficient than others in rebuilding the missing data, depending on pavement age, pavement condition, and rate of deterioration.

To test these hypotheses, statistical techniques are used to rebuild the pavement performance missing data. The prediction efficiency of different statistical techniques is then quantified. A nonparametric Mann-Whitney statistical test (Naser et al. 2012) is used to check the statistical significance of the prediction improvement.

Missing Data Techniques Applied to Predict Pavement Performance

To explain different possible ways of using missing data techniques to predict pavement performance, it is necessary to discuss different known statistical techniques. The final selection of the method for populating missing data depends on the field of application. The following three possible approaches to the handling of missing data have been proposed (Tsikriktsis 2005):

1. Delete the records that contain the missing data;
2. Use model-free methods to fill in the missing data with estimated values; and
3. Model the distribution of existing data and then fill in the missing data with estimates.

The first approach is currently used by the departments of transportation (DOTs) when predicting pavement performance.

Model-Free Replacement Techniques

To replace the missing value, one or several available values of the same quantity are used to rebuild the missing values. Model-free replacement techniques differ in how the corresponding values are selected and processed to rebuild an estimate for the missing data point: *case substitution* techniques use values from the same data set, *subgroup substitution* techniques use values from different data sets, whereas *total substitution* techniques use all available data.

This research uses case substitution and subgroup substitution techniques. The case substitution only uses data from multiple years for a given pavement segment, whereas the subgroup substitution uses data from multiple road segments of a given year. The subgroup substitution uses all of the road segments in the road network, and not only the nearby road segments.

In this study, the following model-free techniques were used: Case substitution techniques in which the missing value is replaced by using (1) the mean of four nearby distress score values (two before and two after); (2) the median of four nearby distress score values; or (3) moving average (replacing the missing value with the last available data point); and subgroup substitution techniques in which the missing value is replaced by (4) mean; (5) median; (6) maximum; or (7) minimum of the distress scores corresponding to the road segments of the same year. These techniques were evaluated because of their reliability in rebuilding the missing data points (Tsikriktsis 2005). Out of all these techniques, only the moving average can be used for prediction, because all other model-free techniques would require additional data scores.

For some of these missing data techniques, the rebuilt value is sometimes larger than the previously available value. However, for a pavement section to which no treatment was applied, the distress score can only deteriorate (i.e., decrease) with age. Therefore, in this study, in situations in which the rebuilt value that is estimated by a missing data technique is larger than the previous available value, the rebuilt value is instead set to be equal to this previous value.

Model-Based Replacement Techniques

In addition to model-free missing data techniques, researchers also use model-based techniques. In each of these techniques, a statistical model is fixed. The parameters of the corresponding model are obtained from available data, and then missing values are rebuilt based on these parameters. The parameters are estimated based either on the whole data series, or only on data from a limited time period (DeSarboand Rao 1986). A model in which different values of the parameters are used to describe different time periods is known as a *spline*. The most commonly used are cubic splines. In missing data techniques, it is usually assumed that the observed data are a sample drawn from a multivariate normal distribution.

In this study, the following model-based replacement techniques were used: (1) Linear interpolation based on two nearby distress score values (one before and one after); (2) linear regression (based on the whole history of the given road segment); (3) cubic regression (also based on the whole history); and (4) cubic spline based on four nearby distress score values (two before and two after). These techniques were selected because of their reliability in rebuilding the missing data points (Tsikriktsis 2005). Each of these techniques can also be used for prediction, as an alternative to Eq. (2).

Case Study

A case study was conducted to test the efficiency of the missing data approach in predicting the distress scores. The pavement data were taken from the TxDOT database. TxDOT divides Texas into 25 districts, out of which 23 have continuous reinforced concrete pavement. This case study used all of the distress records related to CRCP pavements in the 23 districts from 1993 to 2010.

Ideally, each 0.80 km (0.50-mi) pavement section should have 18 distress scores corresponding to all 18 fiscal years from 1993 to 2010. In practice, some of these data points are missing. Because accurate predictions are only possible when there are sufficiently many data points, this study only considers road segments in which at least 10 distress scores out of 18 possible are available. In addition, this study only considers road segments that exhibit at least 10% distress score deterioration from 1993 to 2010. Overall, there are 491 pavement sections that fit these requirements.

The following procedure is used to test how efficient are the missing data techniques in predicting pavement performance. For each selected road segment, the actual distress score value, A, corresponding to the last year, Y, is deleted, and the distress score values in the previous years are used to predict that distress score. For example, if the last data point extends up to year 2010, the 2010 distress score value is deleted and then the data from all available previous years $1993, 1994, \ldots, 2009$ are used to predict this distress score, A (i.e., the value DS_{2010}).

To simulate the situation with missing data points, one, two, or three previous years are selected at random and the distress scores corresponding to these selected years are deleted. For example, if all of the data corresponding to the years 1993 through 2009 are available, and one year (1999) is selected at random, then the values from $1993, 1994, \ldots, 1998$, and from years $2000, 2001, \ldots, 2009$ are selected to predict the distress score at year 2010.

The predicted value P for the distress score at year Y is then obtained using the remaining data points. In the previous example, the data from years $1993, 1994, \ldots, 1998, 2000, 2001, \ldots, 2009$ is used to predict the distress score at year 2010. (For some road segments, the value P predicted by using the Eq. (2) is below 1 or above 100; approximately 15% of such segments were excluded from the analyses.) In general, this prediction is somewhat different from the actual value A; the difference $|P - A|$ is taken as a measure of accuracy of this prediction.

The following statistical missing data techniques are used to rebuild the deleted distress score data points at year 1999:

1. Mean of nearby points;
2. Median of nearby points;
3. Moving average;
4. Subgroup of mean, median, maximum, and minimum substitution;
5. Linear interpolation;
6. Linear trend at point;
7. Cubic spline fitting; and
8. Cubic spline based on four data points.

After the missing data points are rebuilt with the statistical techniques, then either the nonlinear regression and the same missing data technique is used to compute a predicted value R of the distress score at year Y. This prediction R is based both on the original data points (in the previous example, data from years 1993, $1994, \ldots, 1998, 2000, 2001, \ldots., 2009$) and on the rebuilt distress score value (corresponding to year 1999).

This new prediction R is accurate $|R - A|$, in comparison with the prediction error $|P - A|$ of the original PMS prediction method. The smaller the prediction error, the more efficient the prediction method will be.

Evaluation of the Effectiveness of the Statistical Methods in Estimating Missing Values

This study uses a nonparametric Mann-Whitney test to check whether the prediction improvement efficiency of the distress score is statistically significant. The absolute improvement is considered to be statistically significant if the significance value of the Mann-Whitney test does not exceed 0.05 (i.e., $P =< 0.05$).

The Mann-Whitney test checks whether the median value of the accuracy $|R - A|$ of the new prediction method is statistically significantly different from the median of the accuracy $|P - A|$ of the original PMS predictions. The null hypothesis is that the medians of the two samples of accuracy values are equal. The alternative hypothesis is that the medians are different. The medians were used because the distress score distributions are not normal (Naser et al. 2012), and for general (not normal) distributions, the median is known to be a more robust characteristic than the mean.

This test was used to compare the median prediction accuracies obtained without rebuilding data with the median prediction accuracies obtained by applying different missing data techniques to populate (rebuild) missing data sets, as described previously. For each technique, the Mann-Whitney test was performed for the three cases of one, two, and three years missing data, respectively.

Results of Testing

The testing was performed for each *case* consisting of a road segment and a combination of one, two, or three missing data points from that road segment. For each prediction method, and for each number of missing data points, the median value of the prediction accuracy $|R - A|$ over all of the corresponding cases was taken as a measure of the method's prediction efficiency. The results corresponding to one, two, and three missing data points are given in Tables 2–4. Each table lists, for each method, the median value m of the prediction accuracy $|R - A|$ and the p-value corresponding to the Mann-Whitney test. To quantify the efficiency of each data missing technique, a percentage decrease in the median is also listed. This percentage is computed as the ratio in Eq. (3)

$$\text{Percentage Decrease in the Median of } |R - A| = \frac{(M - M_{\text{PMS}})}{(M)} \times 100\% \quad (3)$$

where M = median accuracy of the systematic DS prediction method; and M_{PMS} = median accuracy of the traditional DS prediction method. Tables 2–4 also list the mean accuracy, the quartiles of the accuracy distribution, and the number of cases in which the new prediction method was more accurate, of the same accuracy, or less accurate than the current PMS prediction.

For cases with one missing data point, the moving average was the most efficient method, showing the best median accuracy and leading to 34% improvement in prediction accuracy, whereas the two missing data points resulted in a 12% improvement. For three missing data points, the moving average method is again the most efficient, showing 20% improvement in prediction accuracy. All of these improvements are statistically significant ($p < 0.05$).

In addition to the moving average, the only other rebuilding missing data technique that leads to statistically significant prediction improvements in all three situations (with one, two, and three missing data points) is the subgroup substitution minimum method; however, for this method, for one and three missing data points, the improvement is much smaller than for the moving average: 13% for one missing data point and 4% for three missing data points.

Comparing the Results of Pavement Prediction Approaches

The comparison between the traditional and the proposed prediction approaches is illustrated in Fig. 3. For this pavement section, the PMS contains the distress scores for all the years from 1993 to 2005, except for the years 1995 and 2003. To compare the different approaches, the distress scores for 2001 and 2005 were deleted from the historical records of a pavement section, and the remaining distress score values were used to predict the 2005 value. This comparison was done to evaluate whether the approach developed in this study would be more effective than the traditional approach in predicting the distress score. This is important to analyze, to

Table 2. Median of Prediction Accuracy and Statistical Significance of the Improvement in Predicted Distress Scores: Cases with One Missing Data Point (Number of Cases: 1,232)

		Prediction accuracy $\|R - A\|$						Number of cases		
Prediction method	Missing data technique (MDT)	Median of prediction accuracy $\|R - A\|$	p-value	Percentage decrease in the median of $\|R - A\|$	Mean of $\|R - A\|$	1st quart	3rd quart	More accurate	Same accuracy	Less accurate
Eq. (2)	Traditional PMS method	6.07	—	0.00	12.00	0.94	15.72	0	1,232	0
	Mean of nearby points	5.12	0.04	15.70	11.76	0.55	15.17	689	49	494
	Median of nearby points	5.32	0.03	12.33	11.77	0.56	15.15	678	54	500
	Linear interpolation	5.32	0.03	12.33	11.77	0.56	15.15	678	54	500
	Linear regression	5.14	0.11	15.40	11.84	0.76	15.95	683	36	513
	Moving average	5.39	0.01	11.16	11.77	0.12	15.04	656	52	524
	Cubic regression	5.39	0.07	11.19	11.91	0.55	16.59	692	36	504
	Cubic spline	5.32	0.03	12.29	11.81	0.41	15.16	680	52	500
	Subgroup substitutions mean	5.64	0.03	7.06	11.78	0.39	15.00	667	54	511
	Subgroup substitutions median	5.59	0.03	7.97	11.77	0.18	15.00	662	53	517
	Subgroup substitutions maximum	5.75	0.03	5.23	11.79	0.20	14.97	667	53	512
	Subgroup substitutions minimum	5.25	0.02	13.46	11.77	0.18	15.00	656	51	525
MDT	Linear regression	12.12	0.00	−99.61	17.24	4.76	25.28	164	0	1,068
	Moving average	4.00	0.00	34.10	11.75	0.00	15.00	650	6	576
	Cubic regression	6.15	0.00	−1.28	12.46	1.93	16.52	528	0	704
	Cubic spline	13.00	0.00	−114.16	19.14	5.50	28.71	433	0	799

Table 3. Median of Prediction Accuracy and Statistical Significance of the Improvement in Predicted Distress Scores: Cases with Two Missing Data Points (Number of Cases: 4,144)

		Prediction accuracy $\|R-A\|$						Number of cases		
Prediction method	Missing data technique	Median of prediction accuracy $\|R-A\|$	p-value	Percentage decrease in the median of $\|R-A\|$	Mean of $\|R-A\|$	1st quart	3rd quart	More accurate	Same accuracy	Less accurate
Eq. (2)	Traditional PMS method	4.56	—	0.00	11.44	0.52	15.34	0	4,144	0
	Mean of nearby points	4.52	0.08	0.93	11.30	0.50	14.64	1,866	120	2,158
	Median of nearby points	4.00	0.01	12.35	11.33	0.19	14.43	1,861	128	2,155
	Linear interpolation	4.00	0.01	12.35	11.33	0.19	14.43	1,861	128	2,155
	Linear regression	5.00	0.49	−9.56	11.38	0.95	15.04	1,839	77	2,228
	Moving average	4.00	0.00	12.35	11.33	0.05	14.95	1,726	124	2,294
	Cubic regression	4.46	0.42	2.27	11.45	0.60	15.00	1,953	80	2,111
	Cubic spline	4.33	0.01	5.19	11.39	0.27	14.95	1,853	119	2,172
	Subgroup substitutions mean	4.28	0.00	6.15	11.34	0.12	14.43	1,771	130	2,243
	Subgroup substitutions median	4.06	0.00	11.11	11.38	0.06	14.52	1,816	127	2,201
	Subgroup substitutions maximum	4.26	0.00	6.62	11.38	0.06	14.50	1,823	127	2,194
	Subgroup substitutions minimum	4.00	0.00	12.35	11.30	0.05	14.53	1,766	121	2,257
MDT	Linear regression	10.76	0.00	−135.69	16.16	4.41	24.40	466	0	3,678
	Moving average	4.00	0.00	12.35	10.82	0.00	15.00	532	14	3,598
	Cubic regression	5.40	0.00	−18.32	11.45	1.91	15.19	1,785	0	2,359
	Cubic spline	12.00	0.00	−162.94	18.22	5.50	25.00	1,101	0	3,043

Table 4. Median of Prediction Accuracy and Statistical Significance of the Improvement in Predicted Distress Scores: Cases with Three Missing Data Points (Number of Cases: 7,554)

		Prediction accuracy $\|R-A\|$						Number of cases		
Prediction method	Missing data technique (MDT)	Median of prediction accuracy $\|R-A\|$	p-value	Percentage decrease in the median of $\|R-A\|$	Mean of $\|R-A\|$	1st quart	3rd quart	More accurate	Same accuracy	Less accurate
Eq. (2)	Traditional PMS method	6.22	—	0.00	12.15	0.65	18.65	0	7,554	0
	Mean of nearby points	6.51	0.23	−4.60	12.20	1.35	17.90	3,704	86	3,764
	Median of nearby points	6.37	0.20	−2.46	12.33	0.72	17.96	3,642	101	3,811
	Linear interpolation	6.37	0.2	−2.46	12.33	0.72	17.96	3,642	101	3,811
	Linear regression	6.74	0.00	−8.37	12.16	1.85	17.56	3,616	60	3,878
	Moving average	6.40	0.00	−2.95	12.33	0.06	17.75	3,448	100	4,006
	Cubic regression	6.39	0.03	−2.76	12.20	1.42	18.32	3,840	66	3,648
	Cubic spline	6.58	0.4	−5.86	12.29	0.65	18.56	3,712	94	3,748
	Subgroup substitutions mean	6.94	0.52	−11.67	12.39	0.76	17.00	3,577	103	3,874
	Subgroup substitutions median	7.05	0.17	−13.40	12.55	0.27	18.64	3,609	101	3,844
	Subgroup substitutions maximum	6.72	0.14	−8.09	12.52	0.49	19.18	3,686	106	3,762
	Subgroup substitutions minimum	6.00	0.00	3.52	12.28	0.06	17.55	3,415	88	4,051
MDT	Linear regression	13.23	0.00	−112.71	17.29	4.84	26.12	896	0	6,658
	Moving average	5.00	0.00	19.60	11.54	0.00	16.00	904	24	6,626
	Cubic regression	6.48	0.00	−4.18	11.87	2.17	17.12	3,388	0	4,166
	Cubic spline	13.33	0.00	−114.39	19.10	7.00	26.88	2,032	0	5,522

evaluate whether further implementation of the method would improve the accuracy of pavement performance predictions.

In the traditional approach, only the remaining distress scores were used, from years 1993 to 1994, 1996–2000, 2002, and 2004. The resulting prediction is denoted by P. In the proposed approach, first, the 2001 value is rebuilt; then, both the available distress scores and the rebuilt 2001 distress score are used to predict the 2005 value. The resulting prediction is denoted by R. The value R predicted by using the new approach is much closer to the actual value A of the 2005 distress score than the value P predicted by using the traditional approach.

The results of the illustrative example show the importance of the systematic prediction accuracy for the maintenance and rehabilitation plan, and for need estimates in the decision-making process. In this example, A of the 2005 distress score is 71, P is 67, and R is 73. The distress score predicted by using the traditional approach is classified as poor. However, the distress score predicted by using the new approach is classified as fair, which conforms to the classification of the actual PMS distress score value A. This example confirms the reliability of the systematic approach in predicting the distress score values.

Refined Testing

Up until now, different missing data techniques were tested and compared on all CRCP road segments, and it was shown that

Fig. 3. Example of predicting distress score: CRCP pavement, section number 3,598, Houston District (reprinted from Al-Zou'bi 2013)

the best technique leads to 17–19% improvement in prediction accuracy. In this test, the same rebuilding missing data techniques were used to rebuild all missing data points for the road segments. A natural hypothesis is that by using different missing data techniques for different missing data points, it may be possible to achieve an even better improvement in prediction accuracy.

To test the dependence of the efficiency on the condition of missing point, the medians of prediction accuracies corresponding to the cases with one missing data point were recomputed on five subsets of the original set of cases, e.g., the subset in which missing points are in very good condition, and the subset in which missing points are in good condition. The results of this recomputation are presented in Table 5.

For all five classes, the moving average technique improved the median accuracy. In all of the classes except for the road segments with a poor distress score, this improvement is statistically significant. For pavements with poor distress scores, none of the missing data techniques lead to statistically significant improvement, as the sample of such road segments is too small. This is not a serious problem, because when one of the observed conditions is poor or very poor, the road segment clearly needs repair; the critical problem of predicting pavement performance is when the observed conditions are very good, good, or fair. For classes corresponding to very good, good, and fair DS, the moving average is the only missing data technique that leads to a statistically significant prediction improvement. Based on this analysis, it is recommended to use the moving average technique.

Conclusions

To achieve the best improvement in pavement prediction quality, this paper proposes that pavement engineers apply missing data techniques to populate missing data points before predicting the future distress scores. These statistical data techniques to populate missing data leads to a significant improvement in predicting the distress score. The authors recommend first using statistical techniques to rebuild the missing data points, and then use both original data points and the newly rebuilt points to predict future pavement performance.

Table 5. Median Prediction Accuracy and Statistical Significance of the Improvement in Predicted Distress Scores, Based on Five Distress Score Categories

Prediction method	Missing data technique (MDT)	Very good			Good			Fair			Poor			Very poor		
		Median prediction accuracy $\lvert R-A\rvert$	p-value	Percentage decrease in the median of $\lvert R-A\rvert$	Median of prediction accuracy $\lvert R-A\rvert$	p-value	Percentage decrease in the median of $\lvert R-A\rvert$	Median of prediction accuracy $\lvert R-A\rvert$	p-value	Percentage decrease in the median of $\lvert R-A\rvert$	Median of prediction accuracy $\lvert R-A\rvert$	p-value	Percentage decrease in the median of $\lvert R-A\rvert$	Median of prediction accuracy $\lvert R-A\rvert$	p-value	Percentage decrease in the median of $\lvert R-A\rvert$
Eq. (2)	Classical PMS method	6.87	—	0.00	2.06	—	0.00	3.25	—	0.00	2.66	—	0.00	4.67	—	0.00
	Mean of nearby points	6.48	0.21	5.79	1.68	0.08	18.26	3.03	0.63	6.75	1.83	0.27	31.08	3.91	0.08	16
	Median of nearby points	6.95	0.20	−1.16	1.38	0.07	33.18	2.64	0.53	18.73	1.00	0.24	62.42	4.00	0.09	14
	Linear interpolation	6.95	0.20	−1.16	1.38	0.07	33.18	2.64	0.53	18.73	1.00	0.24	62.42	4.00	0.09	14
	Linear regression	6.80	0.42	1.06	1.48	0.07	28.40	3.74	0.66	−15.09	1.87	0.30	29.64	3.79	0.06	19
	Moving average	6.97	0.10	−1.41	1.91	0.11	7.51	3.17	0.46	2.38	2.00	0.38	24.83	3.85	0.08	18
	Cubic regression	6.66	0.29	3.13	1.63	0.07	21.11	2.76	0.62	15.12	2.10	0.36	21.02	3.37	0.11	28
	Cubic spline	6.97	0.16	−1.36	1.34	0.06	34.87	2.74	0.60	15.60	1.00	0.24	62.42	4.00	0.19	14
	Subgroup substitutions mean	6.97	0.17	−1.41	1.50	0.09	27.10	3.35	0.64	−3.11	1.00	0.28	62.42	4.00	0.07	14
	Subgroup substitutions median	6.97	0.13	−1.41	1.91	0.22	7.51	3.62	0.74	−11.44	0.96	0.15	63.85	4.00	0.09	14
	Subgroup substitutions maximum	6.97	0.14	−1.41	1.91	0.26	7.51	3.92	0.71	−20.66	1.00	0.27	62.42	4.00	0.11	14
	Subgroup substitutions minimum	6.97	0.12	−1.41	1.91	0.12	7.51	3.53	0.52	−8.49	1.74	0.37	34.72	4.00	0.09	14
MDT	Linear regression	13.50	0.00	−96.47	4.91	0.00	−138.36	7.72	0.00	−137.31	7.51	0.00	−182.38	8.39	0.02	−80
	Moving average	5.00	0.00	27.26	0.00	0.00	100.00	2.00	0.04	38.48	2.00	0.10	24.83	2.00	0.01	57
	Cubic regression	7.14	0.00	−3.89	1.34	0.93	34.91	3.35	0.41	−3.12	1.70	0.94	36.13	5.23	0.62	−12
	Cubic spline	13.00	0.00	−89.13	11.00	0.00	−433.71	10.00	0.00	−207.58	7.00	0.01	−163.09	13.33	0.00	−185

J. Infrastruct. Syst.

After applying the nonparametric Mann-Whitney test to compare the efficiency of different missing data techniques, the authors recommend using the moving average both to rebuild the missing data points and to predict the distress scores. It is observed that, depending on the number of missing data points, the proposed method will have a major effect on the results when compared with the traditional model, leading to statistically significant ($p < 0.05$) improvement, and when compared with traditional PMS, ranging from 12 to 34%. Rebuilding of the missing data with the moving average, before estimating the distress model parameters and then using the imputed data set and using the revised model structure to predict the distress score, leads to better results than the traditional approach.

There are other missing data techniques that lead to statistically significant improvement in accuracy prediction. For example, the subgroup substitution minimum method improves the accuracy for the cases of one, two, and three missing data points. However, the improvement using these methods is much smaller than for the moving average; for example, the improvement generated by the subgroup substitution minimum ranges only from 4 to 13%. Similar conclusions were obtained when the analysis was performed separately on cases in which the missing data point corresponds to a certain DS condition (e.g., very good, good, fair).

The results of this study confirm the hypothesis that the use of statistical techniques to rebuild missing data improves the accuracy of predicting pavement performance. More accurate prediction of distress scores will provide better information for treatment selection in the pavement management process. As shown in the case study results, the moving average method is in general terms more accurate than the other statistical methods, and it is recommended for populating the missing data. However, the systematic approach involves testing different statistical techniques to handle special cases with particular missing data characteristics, and the results of this comparison need to be interpreted using engineering judgment to make the final decision. An example of special missing data cases is when the data series do not follow a gradual and uninterrupted deterioration trend that may occur as a result of missing maintenance and rehabilitation records. In this situation, the data series may need to be divided using expert knowledge to conduct independent analyses.

Acknowledgments

This research was conducted at the Center for Transportation Infrastructure Systems (CTIS) at the University of Texas at El Paso (UTEP) using pavement management data provided by the Texas Department of Transportation.

References

Al-Zou'bi, M. M. (2013). "A systematic approach to manage missing data in pavement management systems." Ph.D. dissertation, Civil Engineering, Univ. of Texas at El Paso, El Paso, TX.

Anderson, A., Basilevsky, A., and Hum, D. (1983). "Measurement: Theory and techniques." *Handbook of survey research*, P. Rossi, J. Wright, and A. Anderson, eds., Academic Press, New York, 244–251.

Day, N. E. (1969). "Estimating the components of a mixture of normal distributions." *Biometrika*, 56(3), 463–474.

Chu, C., and Durango, C. P. (2008). "Incorporating maintenance effectiveness in the estimation of dynamic infrastructure performance models." *Comput.-Aided Civ. Infrastruct. Eng.*, 23(3), 174–188.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum likelihood for incomplete data via the EM algorithm." *J. R. Stat. Soc. B*, 39, 1–38.

Dempster, A. P., and Rubin, D. B. (1983). "Introduction." *Incomplete data in sample surveys (volume 2): Theory and bibliography*, W. G. Madow, I. Olkin, and D. B. Rubin, eds., Academic Press, New York, 3–10.

DeSarbo, W. S., and Rao, V. R. (1986). "A constrained unfolding model for product positioning analysis." *Marketing Sci.*, 5(1), 1–19.

Fuchs, C. (1982). "Maximum likelihood estimation and model selection in contingency tables with missing data." *J. Am. Stat. Assoc.*, 77, 270–278.

Gao, L., Aguiar-Moya, J., and Zhang, Z. (2011). "Performance modeling of infrastructure condition data with maintenance intervention." *Transportation Research Record 2225-12*, Transportation Research Board, Washington, DC, 109–116.

Hong, F., and Prozzi, J. (2010). "Roughness model accounting for heterogeneity based on in-service pavement performance data." *J. Transp. Eng.*, 10.1061/(ASCE)0733-947X(2010)136:3(205), 205–213.

Kaufman, C. J. (1988). "The application of logical imputation to household measurement." *J. Market Res. Soc.*, 30(4), 453–466.

Krzanowski, W.J. (1982). "Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis-testing approach." *Biometrics*, 38, 991–1002.

Little, R. (1988). "Missing data adjustments in large surveys." *J. Bus. Econ. Stat.*, 6(3), 296–297.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data*, Wiley, New York.

Livneh, M. (1994). "Repeatability and reproducibility of manual pavement distress survey methods." *Proc., 3rd Int. Conf. on Managing Pavements*, Transportation Research Board National Council, Washington, DC.

Marwala, T. (2009). *Computational intelligence for missing data imputation, estimation, and management*, Information Science Reference (an imprint of IGI Global), New York.

Naser, G. H., et al. (2012). "Evaluation and development of pavement scores, performance models SND need estimates for the TXDOT pavement management information system-final report." Texas Dept. of Transportation, Texas Transportation Institute, Austin, TX.

Olkin, I., and Tate, R. F. (1961). "Multivariate correlation models with mixed discrete and continuous variables." *Anal. Math. Stat.*, 448–465.

Orchard, T., and Woodbury, M. A. (1972). "A missing information principle: Theory and applications." *Proc., Sixth Berkeley Symp. on Mathematical Statistics and Probability*, University of California Press, Berkeley, CA, 697–715.

Raymond, M. R. (1986). "Missing data in evaluation research." *Valuation Health Prof.* 9(4), 395–420.

Roth, P. L. (1994). "Missing data: A conceptual review for applied psychologists." *Personnel Psychol.*, 47(3), 537–560.

Roth, P. L., Switzer, F. S., and Switzer, D. M. (1999). "Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques." *Organiz. Res. Methods*, 2(3), 211–232.

Shahin, M. Y., Darter, M. I., and Kohn, S. D. (1980). "Condition evaluation of jointed concrete airfield pavement." *Transp. Eng. J.*, 106(4), 381–399.

Stampley, B. E., Smith, R. E., Scullion, T., and Miller, B. (1995a). *Pavement management information system concepts, equations, and analysis models*, Texas Transportation Institute, Texas A&M Univ. System, College Station, TX.

Stampley, B. E., Smith, R. E., Scullion, T., and Miller, B. (1995b). "Pavement management information system concepts, equations, and analysis of pavements." *Research Rep. 1989-1*, Texas Transportation Institute, Texas A&M University System, College Station, TX.

Tsikriktsis, N. (2005). "A review of techniques for treating missing data in OM survey research." *J. Oper. Manage.*, 24(1), 53–62.