Managing Missing Pavement Performance Data in Pavement Management System

Farhan Javed, Ph.D.

North Carolina Department of Transportation (NCDOT)

T. F. Fwa, Professor

National University of Singapore (NUS)

Introduction

- The basic requirement of a pavement management system is to have an efficient pavement condition and performance data collection program to its support decision making process.

- However, missing data in databases has been one of the most prevalent problems in PMS (4).

- Highway agencies have developed data quality management programs (1, 2, 3) entailing procedures and guidelines for managing the quality of pavement data collection activities in terms of quality control and assurance.

Introduction (Cont'd)

- According to NCHRP (5), 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 (6).

- A quality assurance program, developed by the Colorado Department of Transportation (CDOT), checks for duplicate records, missing segments, incorrect highway limits, missing or incorrect highways, incorrect pavement type, and incorrect raw distress values (2).

Introduction (Cont'd)

- MoDOT five-year condition data reported that only the 1999 PSR data were fully complete, and the 1998 PSR data, being the second most complete dataset, had only records of 54 percent of the data (4).

- While the principles of statistical quality assurance in terms of imputation of missing data are well developed, their application and performance to the imputation of pavement management data is ambiguous.

Objectives

- This presentation describes a Multiple Imputation (MI) approach to address the missing pavement condition data issue.

- An analysis on the feasibility and applicability of the approach in comparison to existing imputation techniques is presented.

- The imputation methods examined in this study are:
  - Listwise deletion,
  - Mean substitution,
  - Linear interpolation, and
  - Regression substitution.

Existing Data Imputation Practices

Listwise Deletion

- This is by far the most common approach involving neglecting cases with missing data and to run analyses on remaining data.

Year	Transverse Crack Density (Crack/mi)		Average IRI (in/mi)	Rutting (in)		Avg. Longitudinal Cracking (ft)	Avg. Alligator Cracking (ft^2)	
	Lane 1 (outside)	Lane 2 (inside)		R-RUT	L-RUT			
1980	2.75	2.13	48.18	0.17	0.15	1.26	106.17	
1981	4.01		48.74		0.16	1.62		
1982	4.75	2.22	49.18	0.18	0.20	2.91	161.80	
1983	4.84	2.27	49.37	0.19	0.20	2.95	162.11	
1984	6.2		49.45	0.20		3.20	163.79	
1985	6.21	2.37	49.65	0.20	0.20	3.41	166.03	
1986	6.58	2.39	49.78	0.21	0.21	3.54	179.45	

- This leads to a loss of reliability as the available sample size for potential analyses is reduced

Existing Data Imputation Practices

Mean Substitution

- 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.

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

Existing Data Imputation Practices

Mean Substitution

Year	Transverse Crack Density (Crack/mi)		Average IRI (in/mi)	Rutting (in)		Avg. Longitudinal Cracking (ft)	Avg. Alligator Cracking (ft^2)
	Lane 1 (outside)	Lane 2 (inside)		R-RUT	L-RUT		
1980	2.75	2.13	48.18	0.17	0.15	1.26	106.17
1981	**4.01**	2.27	**48.74**	0.19	**0.16**	**1.62**	156.55
1982	4.75	2.22	49.18	0.18	0.20	2.91	161.80
1983	4.84	2.27	49.37	0.19	0.20	2.95	162.11
1984	**6.2**	2.27	**49.45**	**0.20**	0.18	**3.20**	**163.79**
1985	6.21	2.37	49.65	0.20	0.20	3.41	166.03
1986	6.58	2.39	49.78	0.21	0.21	3.54	179.45

Existing Data Imputation Practices

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 a straight line the point just prior to the missing data with the point just following the missing data.
- Yang et al. (11) applied this approach in forecasting pavement condition rating in Texas.
- This 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),

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

Existing Data Imputation Practices

Regression Substitution

- This approach involves fitting a least-squares regression line to the data on the basis of available information
- The missing data are replaced by the values predicted by this regression line.
- Thus the model takes the following form,

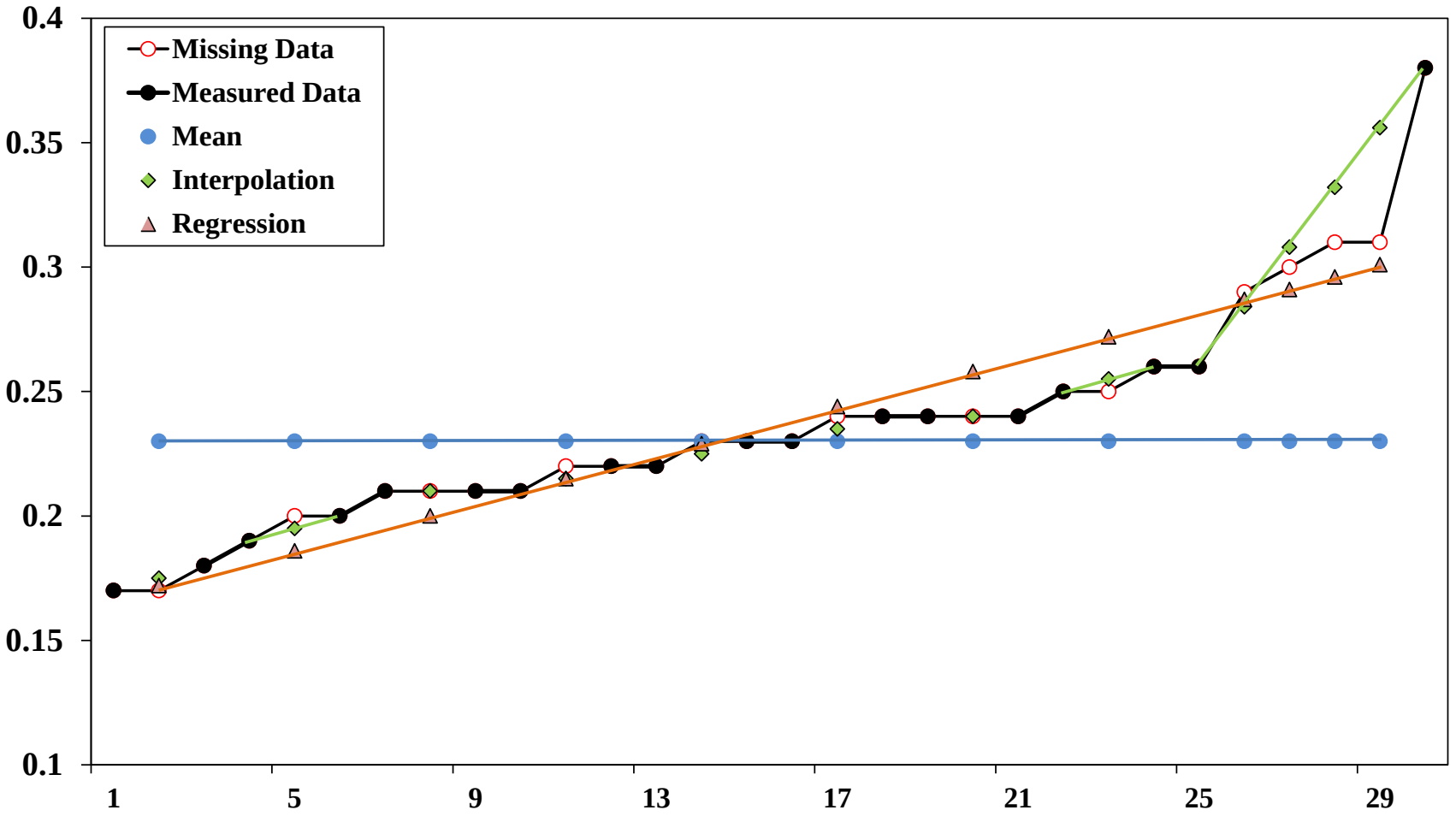
$$y = X\beta + \varepsilon$$

Existing Data Imputation Practices

Limitations

- Mean Substitution
  - Distorts the actual distribution of collected data. Statistically, it produces an imputed dataset with a lower variance than the variance in the original dataset.
- Substitution by Interpolation
  - A form of conditional mean imputation that estimates a missing value only from the available data values immediately before and after it.
- Regression Substitution
  - Fails to account for the variability inherent in the original dataset

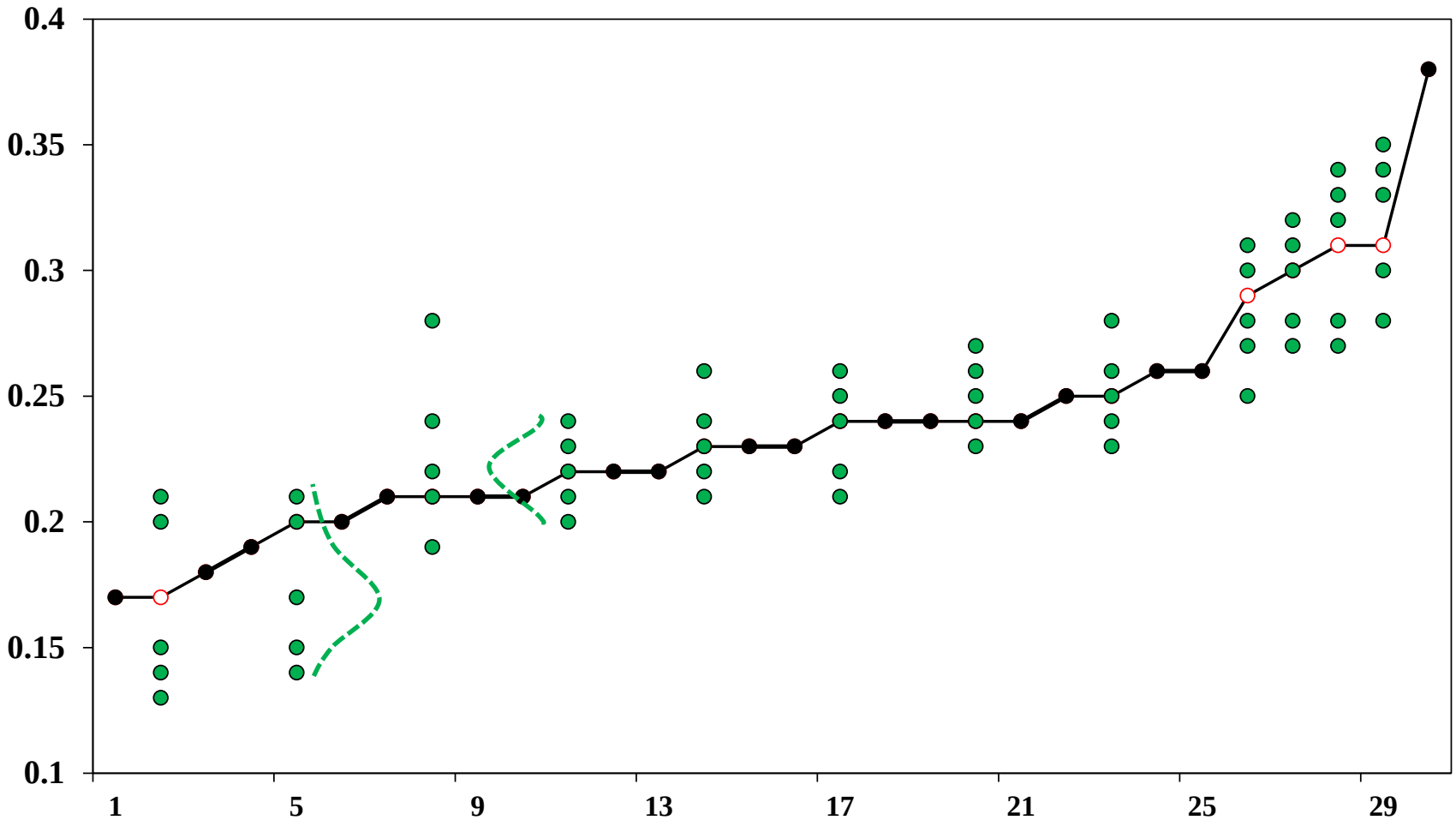
Existing Data Imputation Practices



Proposed Multiple Imputation Approach

- 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 instead of using a point estimate as the imputed value.
- The results are combined to produce overall estimates.
- The technique is performed using Data Augmentation (DA) algorithm (15), however Expectation Maximization algorithm is considered a preferred approach in establishing initial estimates.

Proposed Multiple Imputation Approach



Data Augmentation (DA) Algorithm

- The Data Augmentation Algorithm requires starting values for the mean and covariance matrix.
- Data Augmentation makes use of the concept of Multiple Imputation.
- 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.
- 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 (20).

Data Augmentation (DA) Algorithm

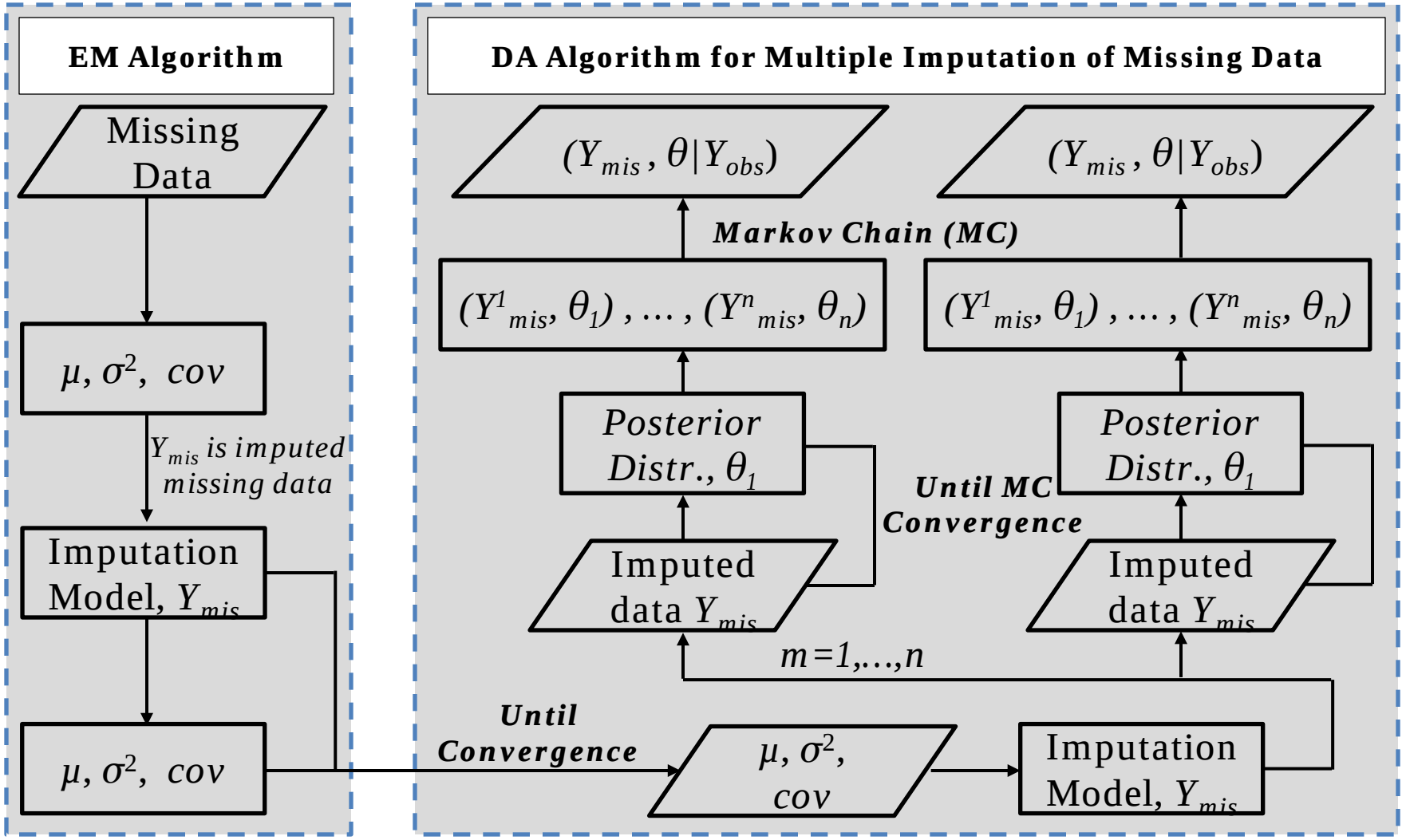
- Beginning at some value of θ, each iteration of the DA algorithm alternates between two steps (20) as follows:

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

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 (20).

Data Augmentation (DA) Algorithm



Implementation Framework

- *Step I: Data Transformation*

  Transform data to approximately normal using a logit, log or square root transformation function.

- *Step II: Imputation using EM*

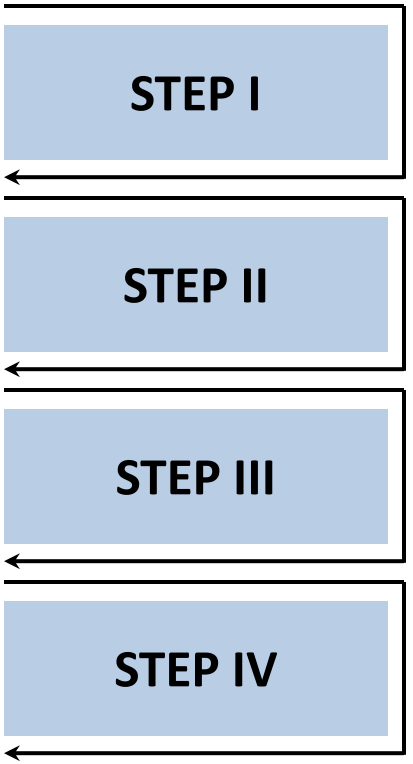
  Generate estimates of missing values for the data matrix using the EM algorithm

- *Step III: Imputation using DA*

  Initial parameter estimates from the EM algorithm, generate imputed data and new parameter estimates

- *Step IV: Synthesis of Estimates*

  Average over the multiple estimates to obtain the final set of estimates



Illustrative Example

- A typical highway pavement condition survey database was used in this study to assess the performance of each imputation approach.

- The database include seven time-series:
  - Transverse crack density in crack per mile,
  - average IRI in inch per mile,
  - rut in inch (left and right wheelpath),
  - average longitudinal cracking in feet, and
  - average alligator cracking in square feet.

Illustrative Example

Year	Transverse Crack Density (Crack/mi)		Average IRI (in/mi)	Rutting (in)		Avg. Longitudinal Cracking (ft)	Avg. Alligator Cracking (ft^2)
	Lane 1 (outside)	Lane 2 (inside)		R-RUT	L-RUT		
1980	2.75	2.13	48.18	0.17	0.15	1.26	106.17
1981	4.01	2.15	48.74	0.17	0.16	1.62	161.72
1982	4.75	2.22	49.18	0.18	0.20	2.91	161.80
1983	4.84	2.27	49.37	0.19	0.20	2.95	162.11
1984	6.2	2.35	49.45	0.20	0.20	3.20	163.79
1985	6.21	2.37	49.65	0.20	0.20	3.41	166.03
1986	6.58	2.39	49.78	0.21	0.21	3.54	179.45
1987	7.09	2.42	50.23	0.21	0.21	3.65	188.04
1988	7.94	2.49	50.44	0.21	0.21	4.85	197.98
1989	7.98	2.53	51.01	0.21	0.21	5.07	203.78
2005	14.71	2.81	54.79	0.29	0.26	9.3	266.81
2006	14.96	2.88	55.53	0.30	0.26	9.52	268.80
2007	16.03	2.88	55.61	0.31	0.26	9.59	282.58
2008	16.69	2.92	56.16	0.31	0.26	9.66	290.97
2009	17.76	3.05	57.46	0.38	0.28	10.16	345.47

Illustrative Example

- In order to draw comparison between various data imputation techniques described earlier, data points are artificially removed from the pavement performance data set.
- Once the missing data has been imputed, real values are compared against imputed values to assess robustness.
- The assessment of the imputation capability is measured using the mean absolute error (MAE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|f_i - y_i\right| = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right|$$

where, f_i = imputed value, y_i = actual value, n = number of observations

Illustrative Example

- Given the normally distributed data, the Expected Maximization algorithm was employed with the convergence criteria set to 0.0001.

- The EM algorithm converged after 15 iterations for all the given pavement performance parameters. The EM estimates serve as starting values for the Data Augmentation (DA) process.

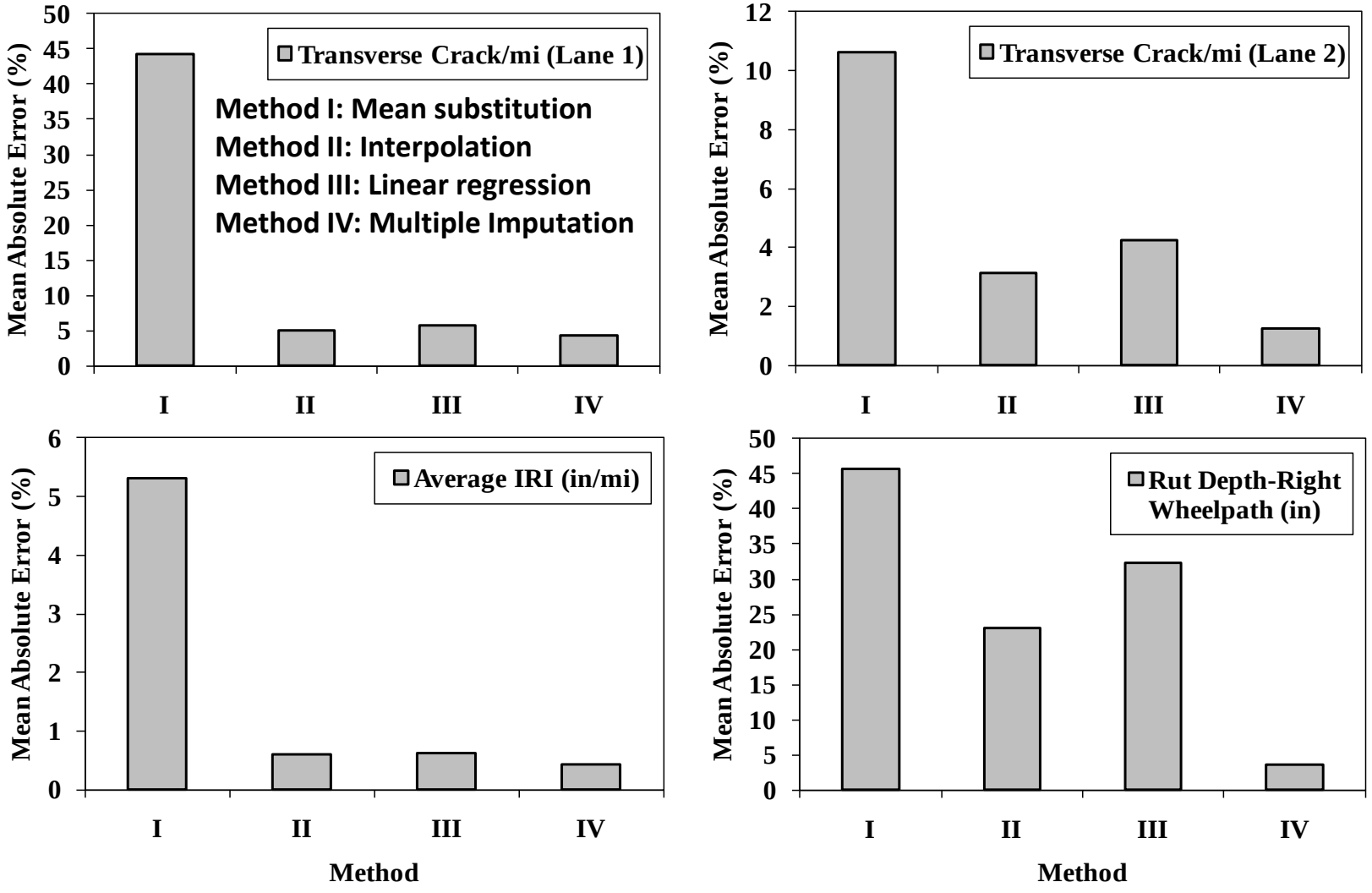
- Since the convergence behavior of DA is the same as EM algorithm, the number of iterations needed for the convergence of the DA algorithm is more or less the same.

Illustrative Example

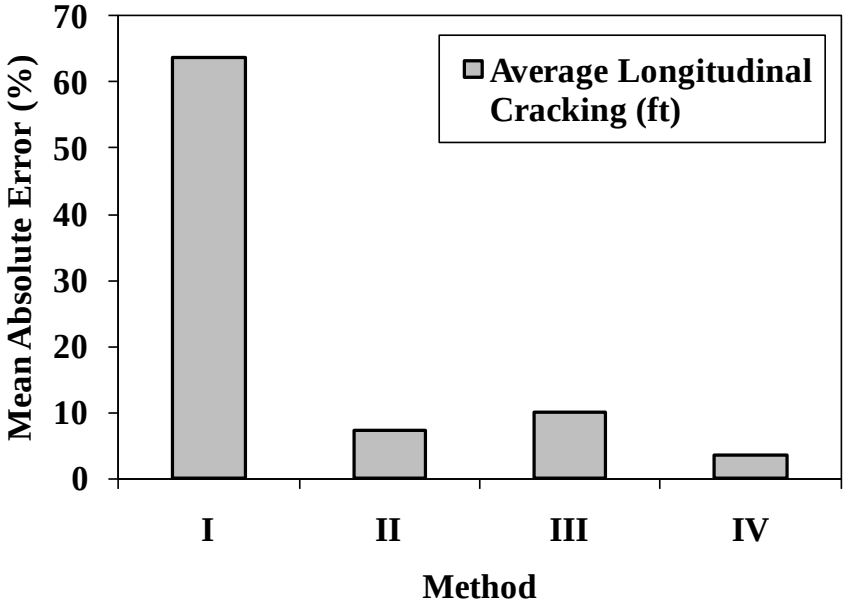
- The performance of the proposed approach is compared against the following methods
  - Method I: Imputation by mean substitution
  - Method II: Imputation by interpolation
  - Method III: Imputation by linear regression
  - Method IV: Imputation by Multiple Imputation

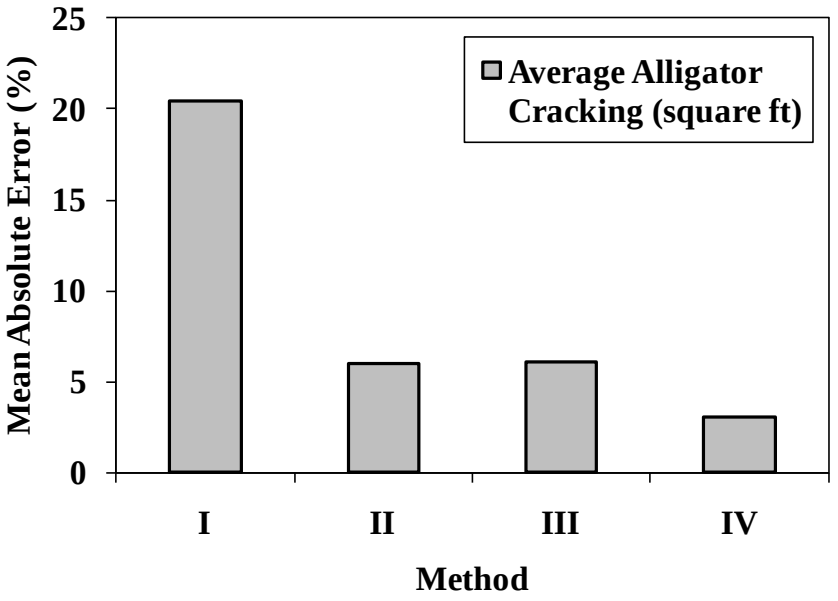
Imputation Method	Transverse Crack Density (%)		Average IRI (%)	Rut Depth (%)		Average Longitudinal Cracking (%)	Average Alligator Cracking (%)
	Lane 1 (outside)	Lane 2 (inside)		Right Rut	Left Rut		
Mean Substitution	44.354	10.648	5.322	45.756	52.821	63.615	20.442
Linear Interpolation	4.982	3.120	0.607	23.116	22.334	7.191	5.987
Regression Substitution	5.653	4.266	0.614	32.259	36.262	10.054	6.059
Multiple Imputation	4.297	1.261	0.433	3.583	3.182	3.651	3.026

Illustrative Example



Illustrative Example





Note: Method I refers to Mean Substitution

Method II refers to Interpolation

Method III refers to Regression Substitution

Method IV refers to Proposed Multiple Imputation

Summary

- A Multiple Imputation approach is presented to address the missing data in pavement condition and performance database of a pavement management system.
- The rationale and applicability of the proposed approach was explained.
- The quality of the imputed data values by the proposed approach was assessed against values obtained using common conventional methods.
- It was found that the proposed approach is superior to other imputation approaches, and produced smaller deviation from actual values from the analysis using cross-validation technique.

Summary

- The aggregated mean absolute error of the imputed values, across several distresses, using Method I is 34.70%, while Method III results in 13.59% followed by 9.62% and 2.77% using Method II and Method IV respectively.

- The mean substitution method resulted in the highest amount of deviations of the imputed values from the actual values, followed by the regression substitution method and the interpolation method.

- However, the Multiple Imputation method proposed in this study yielded the smallest errors for all distress types.

References

1. Larson, C. D. and Forma, E. H. Application of Analytic Hierarchy Process to Select Project Scope for Video Logging and Pavement Condition Data Collection, In Transportation Research Record: Journal of the Transportation Research Board, No. 1990, Transportation Research Board, Washington, D.C., 2007, pp. 40-47.
2. Keleman, M., Henry, S. and Farrokhyar, A. Pavement Management Manual. Colorado Department of Transportation, Denver, CO., 2003.
3. National Cooperative Highway Research Program (NCHRP). Automated Pavement Distress Collection Techniques. National Cooperative Highway Research Program Synthesis Report No. 334, Transportation Research Board, Washington, D.C., 2004.
4. Amado, V. and Bernhardt, K. L. S. Knowledge Discovery in Pavement Condition Data. In the 81st Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2002.
5. National Cooperative Highway Research Program (NCHRP). Quality Management of Pavement Condition Data Collection. National Cooperative Highway Research Program Synthesis Report No. 401, Transportation Research Board, Washington, D.C., 2009.
6. Lindly, J. K., Bell, F. and Sharif U. Specifying Automated Pavement Condition Surveys. Journal of the Transportation Research Forum, Vol. 44, No. 3, 2005, pp. 19-32.
7. Zhang and Smadi. “What is Missing in Quality Control of Contracted Pavement Distress Data Collection?” In the 90th Annual Meeting of Transportation Research Board, Washington, D.C., 2009.
8. Allison P. D. Missing Data. Sage Publications, Inc., Thousand Oaks, CA., 2001.
9. Little, R. J. A. and Rubin, D. B. Statistical Analysis with Missing Data. 2nd edition, John Wiley, New York, 2002.
10. Bennett, C. R. Sectioning of Road Data for Pavement. In the 6th International Conference on Managing Pavements, Queensland, Australia, 2004.
11. Yang, J., Lu, J. J. and Gunaratne, M. Application of Neural Network Models for Forecasting of Pavement Crack Index and Pavement Condition Rating. In Transportation Research Record: Journal of the Transportation Research Board, No. 1699, Transportation Research Board, Washington, D.C., 2003, pp. 3-12.
12. Schafer, J. L. Analysis of Incomplete Multivariate Data. Chapman & Hall, London, 1997.
13. Rubin, D. B. Inference and Missing Data, Biometrika, Vol. 63, No. 3, 1976, pp. 581-592.
14. Rubin, D. B. Multiple Imputation for Survey Nonresponse. Wiley, New York, 1987.
15. Tanner, M. A and Wong, W. H. The Calculation of Posterior Distributions by Data Augmentation. Journal of American Statistical Association, Vol. 82, 1987, pp.528–550.

References

16. Dempster, A. P., Laird, N. M. and Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Vol. 39, No. 1, 1977, pp.1-38.
17. Ripley, B. D. Pattern Recognition and Neural Networks, Cambridge Univ. Press, Cambridge, 1996.
18. Fraley, C. On Computing the Largest Fraction of Missing Information for the EM Algorithm and the Worst Linear Function for Data Augmentation. Computational Statistics & Data Analysis, Vol. 31, 1999. pp. 13–26.
19. Schafer, J. L. and Olsen, M. K. Multivariate Behavioral Research, Vol. 33, 1998, pp. 545–571.
20. Schafer, J. L. and Rubin, D. B. Multiple Imputation for Missing Data Problems, Short course presented at Joint Statistical Meetings, Dallas, TX, August, 1998.
21. Hill, T. and Lewicki, P. Statistics: Methods and Applications, Statsoft, Inc. 2006, pp. 652
22. Anderson, T. W., Darling, D. A. Asymptotic Theory of Certain "Goodness-of-fit" Criteria Based on Stochastic Processes. Annals of Mathematical Statistics, Vol. 23, 1952, pp.193–212.