A Spatial-Bayesian Technique for Imputing Pavement Network Repair Data

Siamak Saliminejad & Nasir G. Gharaibeh*

Zachry Department of Civil Engineering, Texas A&M University, College Station, TX 77843-3136, USA

Abstract: *Pavement construction and repair history is necessary for several pavement management functions such as developing pavement condition prediction models and developing maintenance and rehabilitation (M&R) trigger values based on past repair frequencies. It is often difficult to integrate M&R data with condition data since these data are often stored in disparate heterogeneous databases. This article provides a computational technique for estimating construction and M&R history of a pavement network from the spatiotemporal patterns of its condition data. The technique is founded on Bayesian and spatial statistics and searches pavement condition data in groups of adjacent pavement sections for evidence of repair. The developed technique was applied to a pavement network in Texas and has been found to have a 74% precision and a 95% accuracy in estimating repair history data.*

1 INTRODUCTION

Pavement management is a data-driven process that involves inventory and monitoring data collection (e.g., pavement structure, construction history, traffic, etc.), condition assessment (e.g., aggregation of distress data into condition indexes), condition prediction, and planning of optimal maintenance and rehabilitation (M&R) strategies (Shahin, 2005; Deshpande et al., 2010). Thus, complete and accurate data are essential for a reliable pavement management system (PMS). Unfortunately, the process of preparing and integrating accurate and complete pavement management data is difficult and time-consuming since these data tend to reside in disparate heterogeneous sources. Based on a study of 27 transportation agencies in the United States, Vandervalk-Ostrander et al. (2003) concluded that "most agencies are dealing with disparate data sources in mainframe flat files, redundant data, stovepipe management systems, and functional area barriers." A recent peer exchange of six state and local transportation agencies reported that transportation agencies struggle with numerous problems and challenges in integrating their disparate data (Hall, 2006). Adams (2008) suggested that data integration problems are both technical and organizational since many of the existing information systems and databases were developed separately with one application in mind and using a variety of hardware and software platforms. Thus, it is extremely difficult to locate and assemble information across these many systems (Adams, 2008). The integration of construction history and condition data for a pavement network is one of these challenges. This article is motivated by the need to address this challenge.

This article is concerned with estimating construction and M&R history data from pavement condition data. Although obtaining these data from construction records may be ideal, it may not be practical since construction and condition data for pavement networks are often collected separately and stored separately in legacy databases (FHWA, 2010). As an alternative method, this article provides a computational technique for estimating construction and M&R history of a pavement network from the spatiotemporal patterns of its condition data. The technique is founded on Bayesian and spatial statistics and is implemented in a Geographic Information System (GIS). To demonstrate and validate the developed technique, it was applied to a data set obtained from the Pavement Management Information System (PMIS) of the Texas Department of Transportation (TxDOT).

2 LITERATURE REVIEW

M&R history is necessary for several pavement management functions such as developing pavement

*To whom correspondence should be addressed. E-mail: *ngharaibeh@civil.tamu.edu*.

DOI: 10.1111/j.1467-8667.2012.00762.x

condition prediction models (see, for example, Lytton, 1987 and Wang and Li, 2011), conducting survival analysis, and developing M&R trigger values based on past repair frequency. Several techniques have been used for developing deterministic and probabilistic pavement condition prediction models. In most of these modeling techniques, pavement age and construction history is a key variable. The problem is that even though the M&R data may be available, it is not easily accessible or ready for integration. M&R data are usually gathered by construction personnel and are stored in construction databases (separate from PMS database). For example, at TxDOT, pavement condition data are collected and stored by a statewide data collection unit; whereas M&R data are collected and stored by maintenance and construction units at the district level. Additionally, these databases use incompatible location referencing systems. Integrating these data has always been a major challenge for TxDOT (Zhang et al., 2001). Similar challenges exist in many other highway agencies. For small and new pavement networks, it may be practical to determine the missing M&R information through manual methods such as interviewing construction personnel or manually searching project documents. However, as the network expands and ages, manual solutions turn very time- and labor-consuming.

Researchers have estimated pavement age based on current condition and assumed the performance curves (Lee et al., 1993; Dewan and Smith, 2003; McNinch et al., 2008). Although the utilized performance curve is different in each case, all of these methods estimate construction and M&R history (i.e., pavement age) on a section-by-section basis. Estimating the age of a pavement section solely based on its own condition has some limitation. First, pavement projects often span over multiple pavement sections. Thus, a pavement section that might be in good condition may receive an overlay (for instance) if its adjacent sections are in poor condition and need to be repaired. The second, and perhaps more important limitation, is that pavement condition data are highly noisy. Current collection methods, automated (Ying and Salari, 2010; Lajnef et al., 2011) and manual (e.g., observing video or still images or conducting manual field surveys), introduce variability into the data. This variability can be due to human error, weather conditions during the field survey (e.g., direction and the angle of sunlight, moisture on the pavement surface, and temperature that temporarily heals cracks), and other factors (Prakash et al., 1994; Daleiden and Simpson, 1998; Rada et al., 1998; Smith et al., 1998; Larson et al., 2000; Wolters et al., 2006; Bogus et al., 2010). The technique presented in this article helps to address these limitations by recognizing spatial and temporal patterns in pavement condition data.

Spatial data analysis has been used widely in various areas of civil engineering, such as construction engineering and management (Lee and Adams, 2004; Cheng et al., 2005; Jie and Caldas, 2008; Jia and Wang, 2010), transportation engineering (Li et al., 2007; Vlahogianni et al., 2007; Wang and Kockelman, 2009), water pipeline condition assessment (Adachi and Ellingwood, 2009; de Oliveira et al., 2011), and hydrology (Olivera and Maidment, 1999; Olivera, 2001; Olivera et al., 2006; Cho and Olivera, 2009). In PMS, spatial data analysis has been used to cluster pavement sections based on condition uniformity for project segmentation (Yang et al., 2009) and contracting (Kim et al., 2010) purposes, and for identifying uniform regions for performance modeling purposes (Mishalani and Koutsopoulos, 2002). Many agencies use GIS for generating maps and reports and linking multiple data layers. This article combines spatial data analysis and Bayesian statistics to develop a computational technique for the imputation of missing or inaccessible PMS data (specifically construction and M&R history).

3 DEVELOPED TECHNIQUE

3.1 Rationale for the developed technique

Traditionally, probability is defined for a random event as the relative frequency with which an event occurs in a set of repeated trials. However, in the case of missing M&R history, unknown parameters do not originate from random experiments. Instead, there is uncertainty arising from having insufficient information and not from randomness. As a result, traditional probability concepts are not suitable for addressing this particular problem. To deal with such problems, Bayesian statistics interpret uncertainty as a result of insufficient information and interprets probability as the apparent validity of a hypothesis based on the state of knowledge. The details of Bayesian statistics are well-documented in many textbooks (see, for example, Bolstad, 2007).

To deal with noisy pavement condition data, this method allows for collecting evidence of repair by considering the condition data of a group of adjacent pavement sections, rather than a single one.

3.2 Technique parameters

In this article, we take advantage of observations from close proximity to estimate the probability of repair. To compute the probability that a repair action has been performed on L adjacent pavement sections, as shown in gray color in Figure 1, the following parameters are

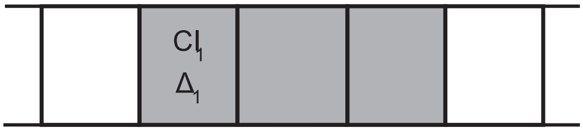

Fig. 1. Data used to calculate the probability that a repair action has been performed on a group of L adjacent pavement sections.

used in conjunction with Bayes' formula:

- Condition of the pavement sections (CI_1 to CI_L) prior to receiving repair action: Pavement sections with a lower condition rating are more likely to receive a repair action than those in better condition.
- Magnitude of increase or decrease in condition rating (Δ_1 to Δ_L): A 20-point increase in a 0–100 condition index is more likely to be a result of repair, and not just data error, in comparison to a rating increase of two points, for instance.
- Number of sections with increased and decreased condition rating (from year to year): The probability that a group of sections indeed received a repair action increases as more sections within that group have improved condition rating. In other words, if the condition ratings of many adjacent sections have improved, it is more probable that these sections indeed received an M&R; compared to another group of adjacent sections where the condition ratings of only a few sections have improved.
- Length of group of adjacent sections (L): Due to practical contract letting considerations, pavement projects tend to have a minimum length. Thus, it is more likely to have a project applied to five 0.5-mile sections, for instance, than to have a pavement project applied to one 0.5-mile section. Note that this length parameter is expressed here in terms of number of adjacent pavement sections (not miles).

The total number of possible projects T with length L in the range of $[L_{\min}, L_{\max}]$ in a roadway that consists of N pavement sections is calculated using Equation (1). As shown in this equation, T increases linearly with N. So, even for large roadway networks, the problem remains tractable.

$$T = \sum_{L=L_{\min}}^{L=L_{\max}} (N - L + 1) = \Delta L \times N - \frac{\Delta L(\Delta L - 1)}{2} \qquad (1)$$

where $\Delta L = L_{\max} - L_{\min}$.

For example, for a hypothetical roadway that consists of 10 pavement sections (see Figure 2), the total number of possible projects with $L = 5$ (i.e., project consists of five pavement sections) is equal to $N - L + 1 = 6$.

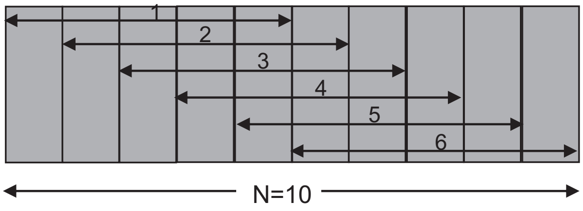

Fig. 2. Six possible boundaries for a project spanning over five sections of a roadway that consists of 10 sections.

To determine the probability that a section was repaired, it is necessary to compute the probabilities of all possible M&R projects which the section is part of. The probability of an M&R project is discussed in the next subsection of this article and is expressed as the probability of repair for a group of contiguous sections.

Finally, all sections are ranked in a decreasing order based on their calculated probability of repair. The highest ranked sections are selected so that the total length of selected sections equals to the average annual portion of the highway network that receives repair. It should be noted that the portion of network that receives repair and the project length range can vary from agency to agency, and thus should be determined based on the agency's past experience. For example, an agency might, on average, repair 10% of its pavement network every year and the project length might range from 2 to 5 miles.

3.3 Probability of repair for a group of contiguous sections

This subsection presents a method for computing the probability that any given contiguous group of pavement sections has received repair in any given year.

By substituting the above-mentioned parameters, Bayes' formula is written as follows:

$$P(R|\mathrm{CI}_1, \ldots, \mathrm{CI}_L, \Delta_1, \ldots \Delta_L, L) = \frac{P(\mathrm{CI}_1, \ldots, \mathrm{CI}_L, \Delta_1, \ldots, \Delta_L|R) \times P(R|L)}{P(\mathrm{CI}_1, \ldots, \mathrm{CI}_L, \Delta_1, \ldots \Delta_L)} \qquad (2)$$

where

R: The event of receiving an M&R action.

CI_i: Condition index of the ith pavement section in the group.

Δ_i: Magnitude of increase or decrease in condition rating for the ith section.

$P(R|\mathrm{CI}_1, \ldots, \mathrm{CI}_L, \Delta_1, \ldots, \Delta_L, L)$: Probability that a group of L sections received an M&R action in a given year, given that the CI and Δ values for all sections in the group and the group length (L) are known.

$P(\mathrm{CI}_1,\ldots,\mathrm{CI}_L,\ \Delta_1,\ldots,\ \Delta_L|R)$: Probability that a group of L sections have CI_1 to CI_L and Δ_1 to Δ_L values, if, in fact, it received an M&R action in that year and it is of length L.

$P(R|L)$: Probability that the group of sections has indeed received an M&R action in a particular year, given that it has a length L. This probability function is constructed based on the total length of the group of pavement sections considered (i.e., without knowing the CI and Δ values of the pavement sections).

$P(\mathrm{CI}_1,\ldots,\mathrm{CI}_L,\ \Delta_1,\ldots,\ \Delta_L)$: Probability that L adjacent pavement sections have condition index values of CI_1 to CI_L and that these values will increase or decrease by Δ_1 to Δ_L in the next year.

To apply Equation (2) to PMS data, a number of checks and transformations are made, as described in the following paragraphs.

Transformation 1: The probability of the product of a number of statements can be expressed by Chain rule (Koch, 2007). Thus, the prior probability of having a repaired group of sections with prerepair condition index of CI_1 to CI_L and postrepair Δ of Δ_1 to Δ_L can be expanded as follows:

$$\begin{aligned}
&P(\mathrm{CI}_1,\ldots,\mathrm{CI}_L,\Delta_1,\ldots\Delta_L|R)\\
&\quad= P(\Delta_L|\mathrm{CI}_1,\ldots\mathrm{CI}_L,\Delta_1,\ldots,\Delta_{L-1},R)\\
&\quad\times P(\Delta_{L-1}|\mathrm{CI}_1,\ldots\mathrm{CI}_L,\Delta_1,\ldots,\Delta_{L-2},R)\\
&\quad\times\cdots\times P(\Delta_1|\mathrm{CI}_1,\ldots\mathrm{CI}_L,R)\\
&\quad\times,P(\mathrm{CI}_L|\mathrm{CI}_1,\ldots\mathrm{CI}_{L-1},R)\\
&\quad\times\cdots\times P(\mathrm{CI}_2|\mathrm{CI}_1,R)\times P(\mathrm{CI}_1|R)
\end{aligned}\tag{3}$$

When a pavement section receives repair action R, the increase or decrease in its CI value (i.e., Δ) depends primarily on the prerepair CI value and repair type. A data set from Texas (see Section 4 of this article) showed that the CI of nearly all repaired sections changes to 100 after an M&R is applied, making $\Delta_i = 100-\mathrm{CI}_i$. For example, if an overlay is applied to a section with an existing CI of 75, Δ will be $100-75=25$ and if an overlay is applied to another section with CI of 90, $\Delta = 100-90=10$. Also, CI_i is independent of CI_j (see Transformation 2) and consequently Δ_i is independent of CI_j since $\Delta_\mathrm{i} = a-\mathrm{CI}_i$, where a is a constant that depends on the repair type. Finally, since CI_i and CI_j are independent, Δ_i and Δ_j would also independent.

It can be shown (Koch, 2007) that if statements A and C_1 to C_k are independent, then the probability that "A is true, given the conditions $C_1,\ldots,C_k$ and B are true" is equal to the probability that "A is true, given that the condition B is true." Mathematically,

$$P(A|\,C_1,\ldots,C_k,B) = P(A|\,B)\tag{4}$$

Based on Equation (4) and the fact that Δ_i is independent of CI_j and Δ_j, it is possible to simplify some terms of Equation (3), as follows:

For $i = 1$ to L:

$$P(\Delta_i|\mathrm{CI}_1,\ldots\mathrm{CI}_L,\Delta_1,\ldots\Delta_{i-1},R) = P(\Delta_i|\mathrm{CI}_i,R)\tag{5}$$

Transformation 2: To be able to justify this transformation, an investigation of possible spatial correlation between the condition ratings of neighboring sections is warranted. There are different methods for representing spatial dependency in geostatistical data. However, the primary tool to investigate spatial dependency is semivariogram as defined in Equation (6) (Schabenberger and Gotway, 2004).

$$\gamma(d_{ij}) = \frac{1}{2}\mathrm{Var}[\mathrm{CI}_i - \mathrm{CI}_j]\tag{6}$$

where d_{ij} is the distance between the geometric centroid of two pavement sections i and j; CI_i and CI_j are the CI values of pavement sections i and i, respectively; and γ is the semiovariagram of pavement sections i and j, which depends only on the distance between sections i and j.

Generally, if there is a spatial dependency in a variable, the semivariogram will have an increasing pattern as the distance increases, ultimately approaching its sill asymptotically. This pattern occurs in a spatially dependent variable since as the distance increases, the variance also increases.

Semivariogram plots were generated for the condition index of randomly selected roadways of different classifications (Farm-to-Market road, State Highway, and Interstate Highway) using condition data from Texas for 2 years (2009 and 2010). Figures 3 and 4 show semivariogram plots for State Highway 79 (SH-79) and Interstate Highway 45 (I-45), as examples (remaining plots are not shown here for brevity). Each point in these plots is computed based on the condition index of a pair of sections. The distance extends from 0.5 mile (representing the distance between the centers of two adjacenet 0.5-mile sections) to 25 miles (representing the distance between the centers of two 0.5-mile sections that are 25 miles apart). Due to the large number of data points, quantiles are shown as bars instead of points. Hollow points in these figures represent outlier data.

The semivariograms shown in Figures 3 and 4 are highly fluctuating and cyclic. This pattern indicates that there is no significant spatial dependency in the condition index data. [Note that the condition index in these plots is TxDOT's 0–100 distress score (DS)]. This finding is not surprising because of the influence of local factors (such as changes in traffic at intersections, variability in subgrade condition, variability in drainage condition, variability in construction quality, etc.) on pavement condition. It should also be mentioned that

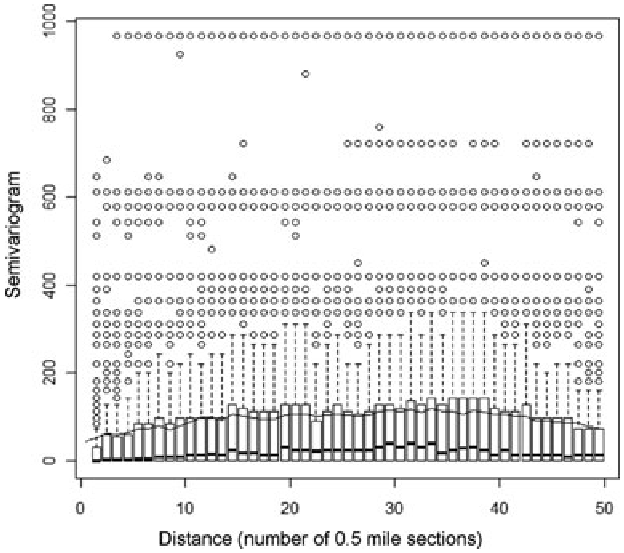

Fig. 3. Condition index semivariogram for Highway SH-79 in 2010.

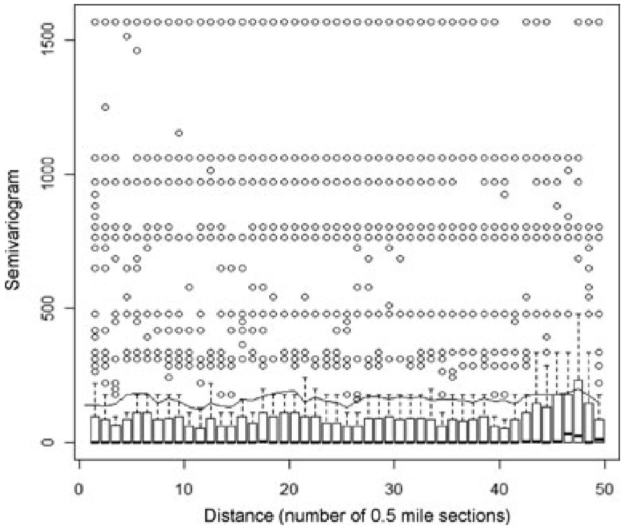

Fig. 4. Condition index semivariogram for Highway I-45 in 2010.

more than 70% of the sections have DS = 100. This leads to similarities in the conditions of sections located far from each other. This is reflected in the cycles and fluctuations of the semivariogram. The CI in this data set (i.e., DS) is biased toward 100 since it does not account for all distress types, severity levels, and density. For example, even if several types of distress exist in the pavement; as long as they are of low density, DS remains at or near 100. Detailed information about this condition index, and several other pavement condition indexes, is provided in Gharaibeh et al. (2010).

Since no strong spatial correlation exists between the CI of pavement sections in the network (i.e., sections are spatially independent in terms of CI values), it is possible to make the following simplifications:

For $i = 1$ to L:

$$P(\mathrm{CI}_i|\mathrm{CI}_1, \ldots, \mathrm{CI}_{i-1}, R) = P(\mathrm{CI}_i|R) \tag{7}$$

After replacing Equation (3) with the simplified terms shown in Equations (5) and (7), the final priori probability can be computed as follows:

$$P(\mathrm{CI}_1, \ldots, \mathrm{CI}_L, \Delta_1, \ldots, \Delta_L|R) = \prod_{i=1}^{L} P(\Delta_i|\mathrm{CI}_i, R) \times \prod_{i=1}^{L} P(\mathrm{CI}_i|R) \tag{8}$$

Similarly, the normalizing part of Equation (2) can be simplified as follows:

$$P(\mathrm{CI}_1, \ldots, \mathrm{CI}_L, \Delta_1, \ldots, \Delta_L) = \prod_{i=1}^{L} P(\Delta_i|\mathrm{CI}_i) \times \prod_{i=1}^{L} P(\mathrm{CI}_i) \tag{9}$$

By applying Equations (8) and (9) to Equation (2), the equation for computing probability of repair for a group of contiguous sections can be expressed as follows:

$$P(R|\mathrm{CI}_1, \ldots, \mathrm{CI}_L, \Delta_1, \ldots, \Delta_L, L) = P(R|L) \times \prod_{i=1}^{L} \frac{P(\mathrm{CI}_i|R)}{P(\mathrm{CI}_i)} \times \prod_{i=1}^{L} \frac{P(\Delta_i|\mathrm{CI}_i, R)}{P(\Delta_i|\mathrm{CI}_i)} \tag{10}$$

where

L: Number of 0.5-mile pavement sections in the group.

$P(R|L)$: Probability that a given group of adjacent pavement sections receives a repair action, given that the total length of that group of sections is L (and without knowing the sections' condition).

$P(\mathrm{CI}_i|R)$: The percentage of repaired sections that have $\mathrm{CI} = \mathrm{CI}_i$.

$P(\mathrm{CI}_i)$: The percentage of all sections that have $\mathrm{CI} = \mathrm{CI}_i$.

$P(\Delta_i|\mathrm{CI}_i,R)$: The percentage of repaired sections with $\Delta = \Delta_i$ that have $\mathrm{CI} = \mathrm{CI}_i$.

3.4 Probability of repair for any given section

Suppose that section X can potentially be part of n M&R projects and the probabilities of these projects (computed using Equation 10) are P_1 to P_n. Then, the probability that section X has been repaired in any given year ($P(X)$) can be calculated using

Equation (11).

$$P(X) = 1 - \prod_{i=1}^{n} (1 - P_i) \tag{11}$$

Finally, all pavement sections in the network are ranked in a decreasing order of probability of repair ($P(X)$) and repaired sections are identified such that the total length of these sections equals the annual portion of the network under repair. This input parameter (i.e., annual portion of network under repair) can be readily obtained from the agency's past experience or historical data (see application discussed in the next section of this article).

4 APPLICATION OF THE DEVELOPED TECHNIQUE

This section discusses the probability density functions (PDFs) required by Equation (10) based on actual data from the Beaumont District of TxDOT. To ensure the applicability of these distributions to pavement networks in other TxDOT districts, these PDFs are used to impute M&R history data in the Bryan District. These districts were used in this study due to their similarities in size (i.e., both districts have similar lane-miles of roadway) and location (both districts are located in east-central Texas). But first, an overview on TxDOT PMS (called PMIS) is provided.

4.1 Pavement management in Texas

TxDOT is divided into 25 Districts and each District includes 6–17 counties. Different highway types exist within each county. For pavement management and data collection purposes, TxDOT divides each highway into a number of sections with an average length of half a mile. TxDOT classifies its pavements into 10 types: continuously-reinforced concrete, jointed concrete reinforced, jointed concrete unreinforced, thick asphalt concrete, intermediate asphalt concrete, thin asphalt concrete, composite, concrete overlaid, flexible overlaid, and thin-surfaced flexible base. Furthermore, each pavement type is classified based on the most recent M&R activity received. These M&R activities include heavy rehabilitation (HR), medium rehabilitation (MR), light rehabilitation (LR), and preventive maintenance (PM).

Each pavement type has its own distress types. For example, asphalt pavement has seven distress types (shallow rutting, deep rutting, patching, failures, block cracking, alligator cracking, and longitudinal cracking).

The percentage of the area or length of the section which is affected by each distress type is recorded as the density of that distress (Q). Q values are converted to utility values using Equation (12) and then aggregated into a DS using Equation (13), which is an overall condition index of the pavement section. Additional information about this condition assessment method can be found in Gharaibeh et al. (2010).

$$U_i = \max\left(0, 1 - \alpha e^{-\left(\frac{\rho}{Q_i}\right)^{\beta}}\right) \tag{12}$$

where U_i is the utility value for distress type i, Q_i is the density of distress type i, and α, β, and ρ are the utility coefficients defined based on pavement and distress types. These factors are always positive.

$$\mathrm{DS} = 100 \prod_{i=1}^{n} U_i \tag{13}$$

where DS is the distress score of a pavement section (0–100 scale), n the number of existing distress types, and U_i the utility value for distress type i.

Since α, β, and ρ are positive, U_i values will always be in the range of [0,1], according to Equation (12). As a result, the multiplication of any number of U_i's will also be in the range of [0,1]. Consequently, DS is guaranteed to be in the range of [0,100] according to Equation (13).

Pavement condition data are stored directly in the PMIS database although maintenance and repair project data are stored in separate databases. In the PMIS database, pavement sections are identified (referenced) by a unique address which is a combination of district name, county name, highway name, and beginning and end reference mile markers. On the other hand, M&R and construction databases identify projects by a code (called control-section-job or CSJ) and an approximate location on the roadway. Also, in many cases, the repair year and specific work type are missing. Additionally, some routine maintenance projects are done by in-house forces and thus are not recorded in the database at all. Thus, it is very difficult to integrate construction and repair history with condition data from these databases.

Pavement condition data are collected annually by a vendor. However, TxDOT collects "audit" condition data on about 5% of the network for validation purposes. To assess the variability of the collected condition data, standard deviation of error is calculated with the assumption that the true DS value is the average of the audit and vendor values. The DS standard deviation of error for all TxDOT roads was calculated as 5.8%, 7.5%, 8.2%, and 6.5% for years 2007–2010, respectively.

A random error with the standard deviation of 5–10 is significant in the sense that it can mask the natural year-to-year deterioration of a pavement section and even show a false improvement in condition. This rationale

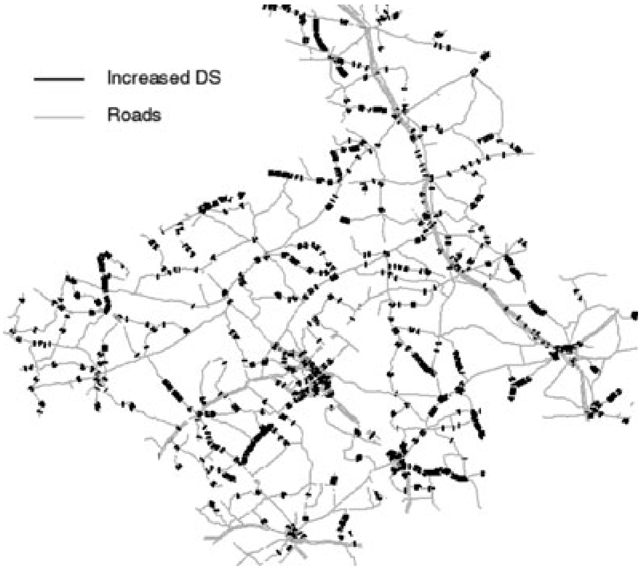

Fig. 5. Sections with improved condition in Bryan District from 2009 to 2010.

was used by the Virginia DOT to assign a maximum acceptable level of variability in distress data: "Data variability for each data element must be smaller than the year-to-year change in that element" (Flintch and McGhee, 2009).

In a perfect data set, DS of all sections in the network should increase upon receiving a repair action. Otherwise, it should stay constant or decrease with time. In such a data set, temporal patterns in condition data are sufficient to impute missing M&R data without considering the spatial patterns of the data. However, such a perfect data set does not exist since, as discussed earlier, pavement condition data are noisy and contain errors. Figure 5 shows changes in DS of pavement sections from 2009 to 2010 in TxDOT's Bryan District.

Dark segments in this map represent pavement sections with improved condition from 2009 to 2010. This improvement can either be due to receiving an M&R action or due to an error in the condition data. As discussed earlier, pavement sections receive M&R in the form of projects. These projects are applied on a number of adjacent pavement sections and it is unlikely to have a project that consists of only one half-mile pavement section. For example, in fiscal year 2009 in the Bryan District, the average project length was 7.2 miles and 95% of projects had a length greater than 2 miles. Thus, one can conclude that the clusters of dark sections represent project boundaries, while scattered single dark sections are likely to be erroneous condition data.

For example, Figure 6 shows the historical DS values for a pavement section located on roadway FM0060 (starting at mile marker 0636 +00.5) in Brazos County. This section has not received any repair in years 2007–2010, and thus the changes in DS values are strictly due to errors in the data. The developed method allows for dealing with this data set so that accurate inference about M&R history can be made.

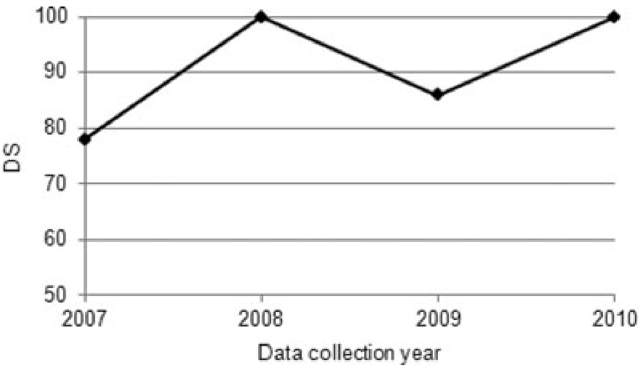

Fig. 6. Temporal variation in DS of an example pavement section (Brazos County, Roadway FM0060 starting at 0636 +00.5).

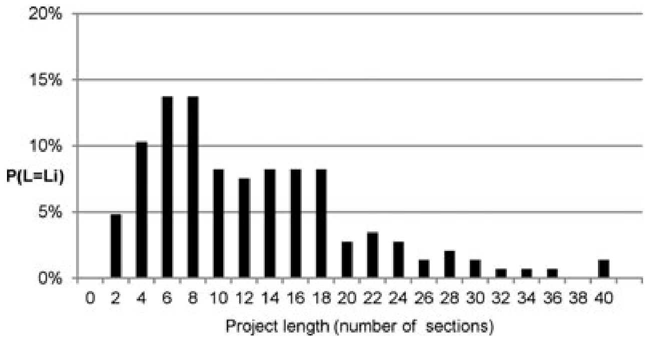

Fig. 7. $P(R|L)$: Percentage of projects with various length values.

4.2 Determining technique parameters

The distribution of project lengths completed by the Beaumont District of TxDOT in 2009 is shown in Figure 7. This distribution is considered as background information and represents $P(R|L)$ in Equation (10).

The other PDFs for the Beaumont District data set are shown in Figures 8–11. It should be noted that in Figures 8 and 9, the percentage of sections with DS = 100 are 65.4% and 44.2%, respectively. These values are much higher than the other bars and thus are noted separately in these graphs. $P(\text{DS} = 100|R) = 44.2\%$ means that 44.2% of the sections that have been repaired in 2010 had a DS of 100 in 2009. The percentage of sections that have a DS value of 100 in 2009 and were repaired in 2010 is 6.8% (313/4,637 = 6.8%). These percentages

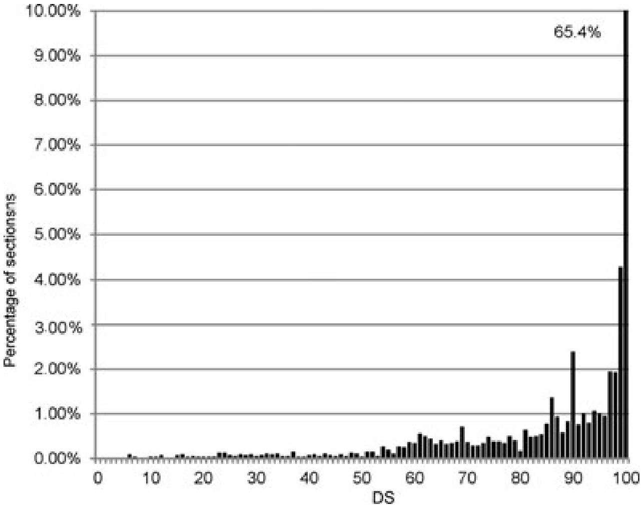

Fig. 8. *P*(DS): Percentage of sections with each DS value.

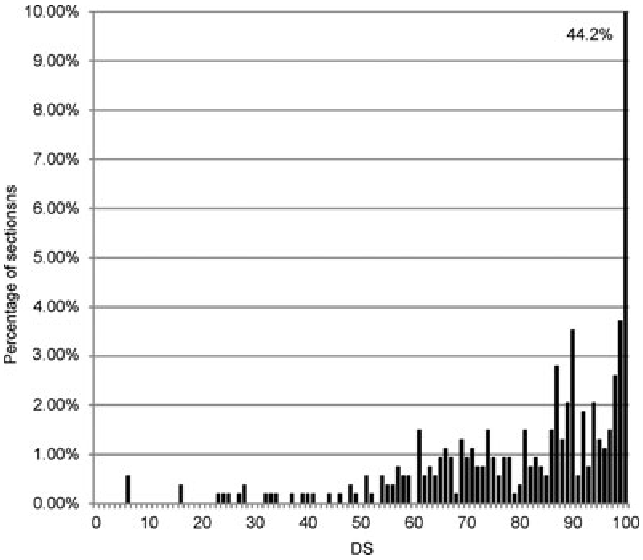

Fig. 9. *P*(DS|*R*): Percentage of repaired sections with each DS value.

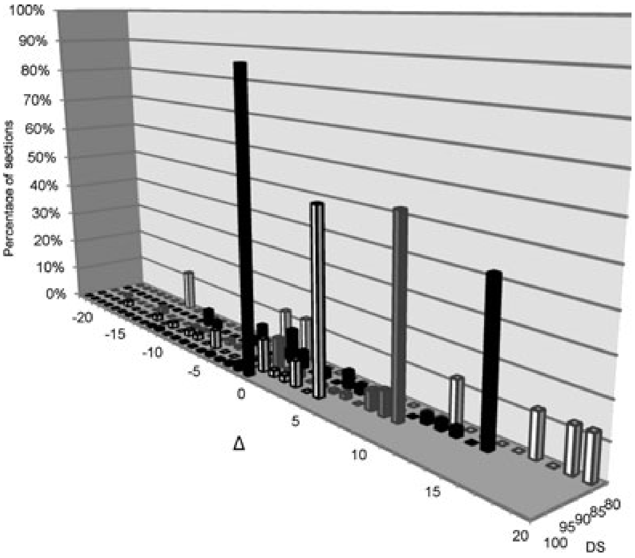

Fig. 10. $P(\Delta|\text{DS})$: Percentage of sections with each Δ value in each DS category.

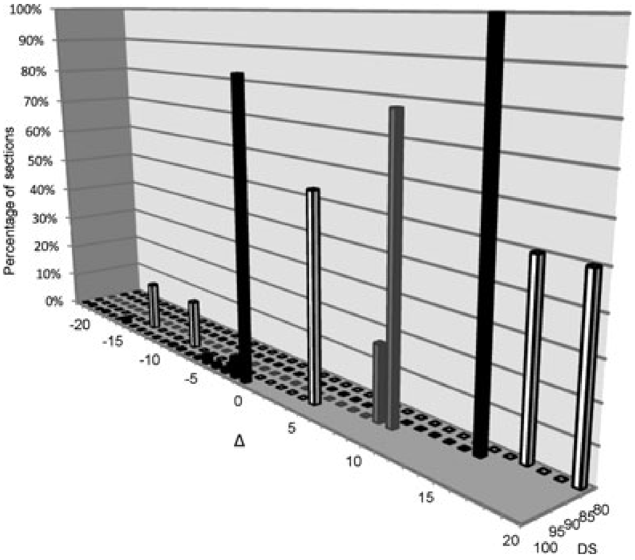

Fig. 11. $P(\Delta|\text{DS},R)$: Percentage of repaired sections with each Δ value in each DS category.

raise the question: Why were so many sections with DS = 100 repaired? Possible explanations include:

1. A pavement section with DS = 100 is not necessarily in perfect condition, as DS does not account for all distress types, severity levels, and densities.
2. Due to economy-of-scale and other practical reasons, M&R actions are normally applied to a group of sections and not to isolated individual sections. Thus, sections with DS = 100 may be repaired only because the neighboring sections needed to be repaired.
3. Condition data contain inherent errors that cannot be eliminated completely.

It should be mentioned that the DS, by definition, is a continuous variable. However, it is normally recorded as a rounded value for practical reasons, as presented in Figures 8–11. For example, in Figure 8, the bar of DS = 90 represents the PDF of a DS range of [89.5, 90.5) and at the same time it represents *P*(DS = 90).

Table 1 shows a 3.5-mile stretch of pavement (composed of seven half-a-mile sections), as an example calculation. In this example, CI is TxDOT's DS. The DS values are obtained from the pavement management database. Δ is computed as Δ = 100−DS (assuming that the repair action increases DS to 100). PDF values for

Table 1
Numerical example of key calculations of the developed technique

Section number (i)	CI	Δ	P(CI)	P(CI|R)	P(Δ|CI)	P(Δ|CI,R)
1	96	4	0.94%	1.11%	65%	100%
2	100	0	65.43%	44.16%	87%	84%
3	90	10	2.38%	3.53%	55%	79%
4	100	0	65.43%	44.16%	87%	84%
5	99	1	4.27%	3.71%	65%	70%
6	94	6	1.05%	2.04%	57%	64%
7	98	2	1.91%	2.60%	64%	86%

each section are obtained from frequency distribution graphs presented in this section (Figures 8–11).

The length of this group is 3.5 miles and according to the distribution shown in Figure 7, $P(R|L) = 14.38\%$. By applying the parameters shown in Table 1 to Equation (10), the probability that this group of pavement sections has received an M&R action between the two conductive data collection years is calculated to be 88%.

5 RESULTS AND DISCUSSION

Using Equations (10) and (11) in conjunction with the PDFs plotted in the previous section, the probability that an M&R action has been applied is calculated for each section in the network. Then, the sections are sorted based on this probability. These calculations are implemented in ArcGIS using the Python coding language. Finally, the sections with the highest probability of being repaired are picked so that the sum of their lengths is equal to a predefined portion of the total length of the highway network. In this analysis, a 10% portion is used since, on average, TxDOT repairs 10% of the network annually. This input parameter (i.e., network repaired portion, hereafter abbreviated as NRP) should be determined based on the agency's past experience or historical data.

Each section classified by the developed method falls in one of four classes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Thus, the precision and accuracy of this binary classification method can be defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{15}$$

where

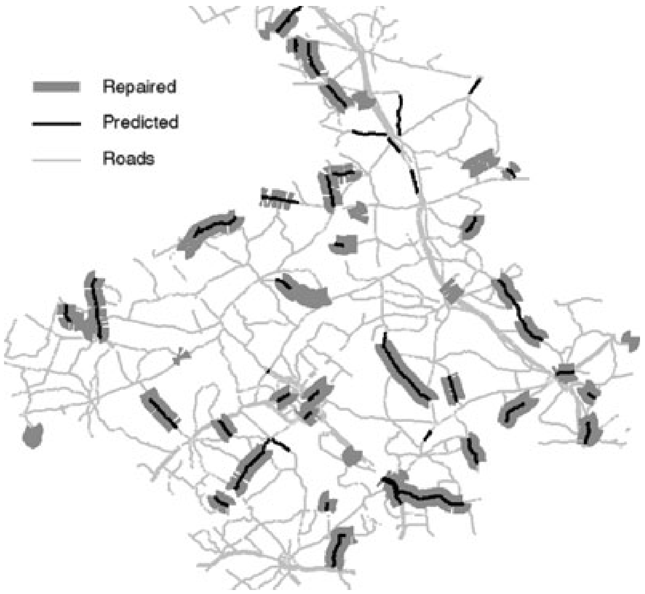

Fig. 12. Estimated M&R projects (dark segments) versus actual M&R projects (thick gray segments) in Bryan District.

TP: The number of sections that have been correctly classified as repaired.

TN: The number of sections that have been correctly classified as not repaired.

FP: The number of sections that have been incorrectly classified as repaired.

FN: The number of sections that have been incorrectly classified as not repaired.

Precision represents the effectiveness of the method in correctly detecting repaired sections. Accuracy, on the other hand, is the proportion of true results (both TP and TN), in the whole population of pavement sections in the network. It represents the veracity of the developed method.

To measure effectiveness of the developed method, the Beaumont District PDFs are used to estimate repair projects in a different district (the Bryan District). The Bryan District highway network consists of 7,090 sections (3,600 miles). A total of 74% of the sections that were classified by the developed spatial-Bayesian technique as repaired were actually repaired in 2010; indicating that precision of this technique has 74% precision. The accuracy in this case was 95%, indicating that 95% of all sections in the network were classified correctly. Predicted versus actual repaired sections are shown in Figure 12 (thick gray segments represent actual M&R projects and dark segments represent M&R projects estimated by the developed technique).

The sensitivity of the developed method to the NRP input parameter is shown in Figure 13. Note that NRP was varied between 1% and 20% since it is unlikely that

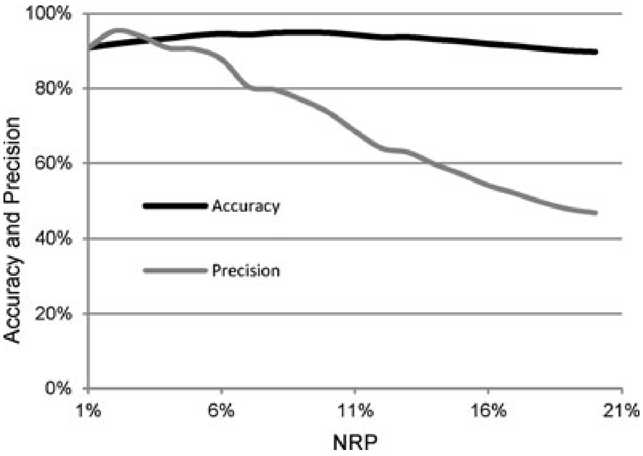

Fig. 13. Sensitivity of the accuracy and precision of the developed method to the NRP input parameter.

a highway agency will repair more than 20% or less than 1% of its network in any given year. This figure shows that accuracy ranges between 90% and 95%, with a maximum accuracy achieved at an optimum NRP of 10%. This optimum value of NRP occurs since as NRP increases, the number of TPs increases, but at the same time, the number of FPs also increases (as more TNs change to FPs). However, the number of FPs increases at a faster rate than the number of TPs; resulting in a decreased precision with increased NRP.

Limitations of the developed method and possible causes of error in the results include:

1. Systematic error in the CI data: More than 60% of falsely classified (FP and FN) sections are those with deteriorated condition (i.e., lower CI value) after they have been repaired, or those with increased CI values without receiving M&R. This indicates that the CI data for these sections are likely to be erroneous.
2. Inability of the CI to reflect the true condition of the pavement: For example, TxDOT's DS does not account for pavement skid resistance and some individual distress types. In such situations, the agency may make repair decisions based on condition indicators that are not accounted for in the CI.
3. Presence of small projects: The calculated probability of repair for short projects can be highly sensitive to error in the CI data.
4. Nonrepresentative or inaccurate prior information: Low-quality prior information can lead to inaccurate prior probability distribution functions and consequently reduce the accuracy and precision of the developed method.

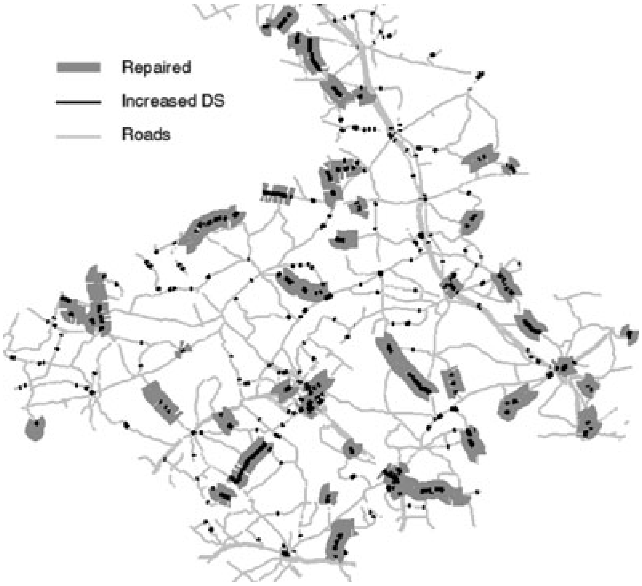

Fig. 14. M&R projects estimated solely based on improvement in DS on a section-by-section basis (dark segments) versus actual M&R projects (thick gray segments) in Beaumont District.

5. The derivation of Equation (11) implicitly assumes that the various project probabilities are independent of each other. This simplifying assumption can be violated since M&R projects can share common pavement sections. To account for this dependency, further research would be needed in modeling the covariance structure of M&R projects.

The accuracy and precision of the developed method can be improved by addressing the above limitations and sources of error. However, reducing systematic error in the CI data has the greatest impact on the effectiveness of the developed method.

Figure 14 shows estimated projects using the section-by-section method. In this method, the estimation of repair history is done solely based on each section's condition history from 2009 to 2010 (i.e., ignoring condition of neighboring sections). In this way, 10% of the sections with the highest DS increase are estimated to have received an M&R action. This results in a prediction precision of 47%; which is much lower than the 74% precision of the developed spatial-Bayesian technique.

To quantify the effect of considering spatial patterns on the robustness of this data imputation method, an artificially error-free pavement condition data set was manufactured. Then, a normally distributed random noise in the DS with average of zero and standard deviation of σ was introduced to the data set. The precision of the developed spatial-Bayesian technique and the section-by-section method (which ignores the

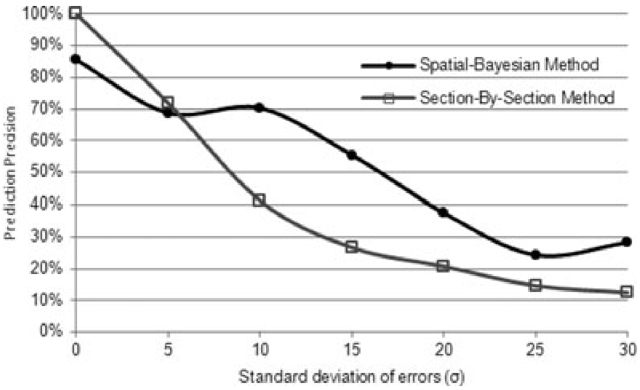

Fig. 15. Precision of spatial-Bayesian and section-by-section methods versus standard deviation of error in pavement condition.

spatial patterns in condition and repair data) for estimating past M&R projects was computed and plotted in Figure 15.

Figure 15 shows how the precision of both methods decreases as the standard deviation of error in the pavement condition data set increases. The nonmonotonic trend of the spatial-Bayesian method in this plot, as opposed to the monotonic trend of the section-by-section method, can be attributed to the PDFs and the interactions among this prior information. Also, the spatial-Bayesian method is not 100% precise even if no error is present in the condition data. This is because of the effect of prior information about project length (i.e., project length PDF). For example, an isolated single 0.5-mile section might have been repaired and its DS increased (and thus should be classified as an M&R project in itself), but the spatial-Bayesian method considers this information unlikely because the project length prior PDF shows that 0.5-mile M&R projects are very unlikely. Similar situations can occur for very long projects (also considered to be unlikely based on project length PDF).

The precision of the spatial-Bayesian method remains stable (ranging between 70% and 85%) up to a DS standard deviation of error of 10 DS points. This behavior suggests that a 10% error in the condition data is a critical point where the method is no longer stable. As discussed earlier (see Section 4.1), σ usually ranges between 5 and 10 DS points. On the other hand, the precision of the section-by-section method decreases steadily as the error level in DS increases.

Predicted versus actual repaired sections for $\sigma = 0$, 5, 10, and 15 using the spatial-Bayesian method and the section-by-section method are shown in Figures 16–19. As shown in these figures, for $\sigma > 10$, the section-by-section method approaches a random estimation of M&R projects (as can be seen in the dotty map). However, for spatial-Bayesian method, the results remain fairly consistent. These comparisons indicate that the developed technique is fairly robust.

6 SUMMARY AND FUTURE RESEARCH

Pavement M&R data and condition data are often stored in disparate heterogeneous databases that are

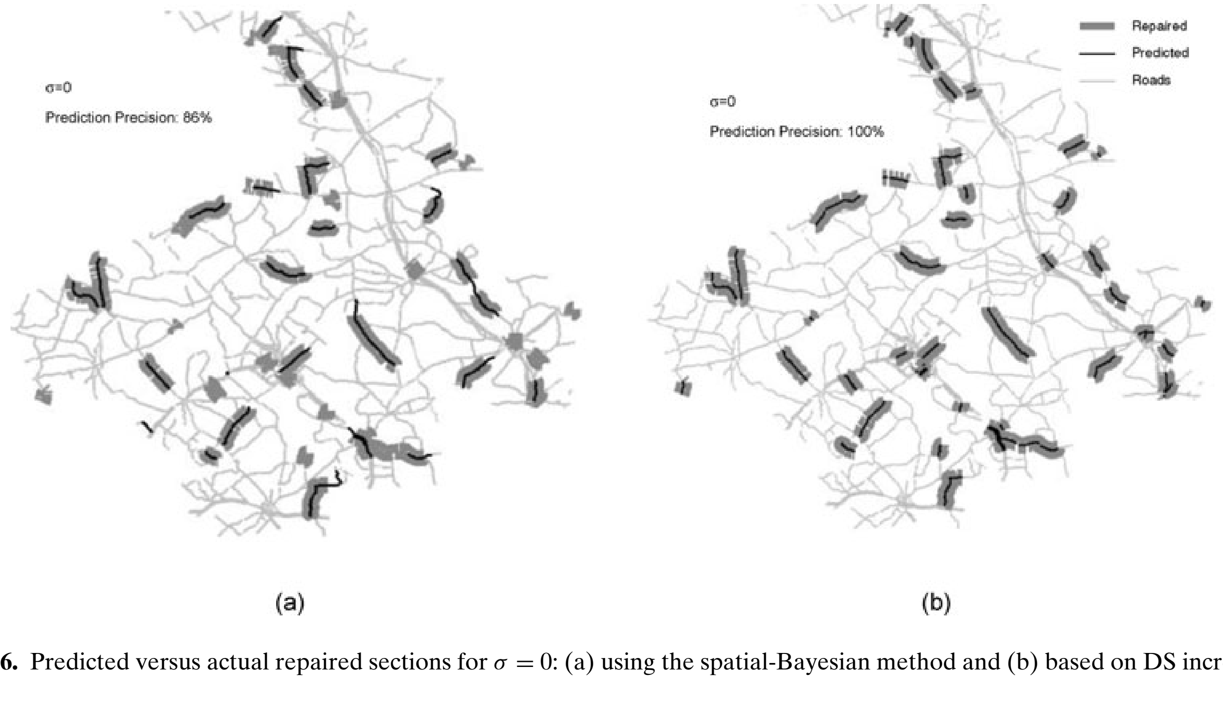

Fig. 16. Predicted versus actual repaired sections for $\sigma = 0$: (a) using the spatial-Bayesian method and (b) based on DS increases.

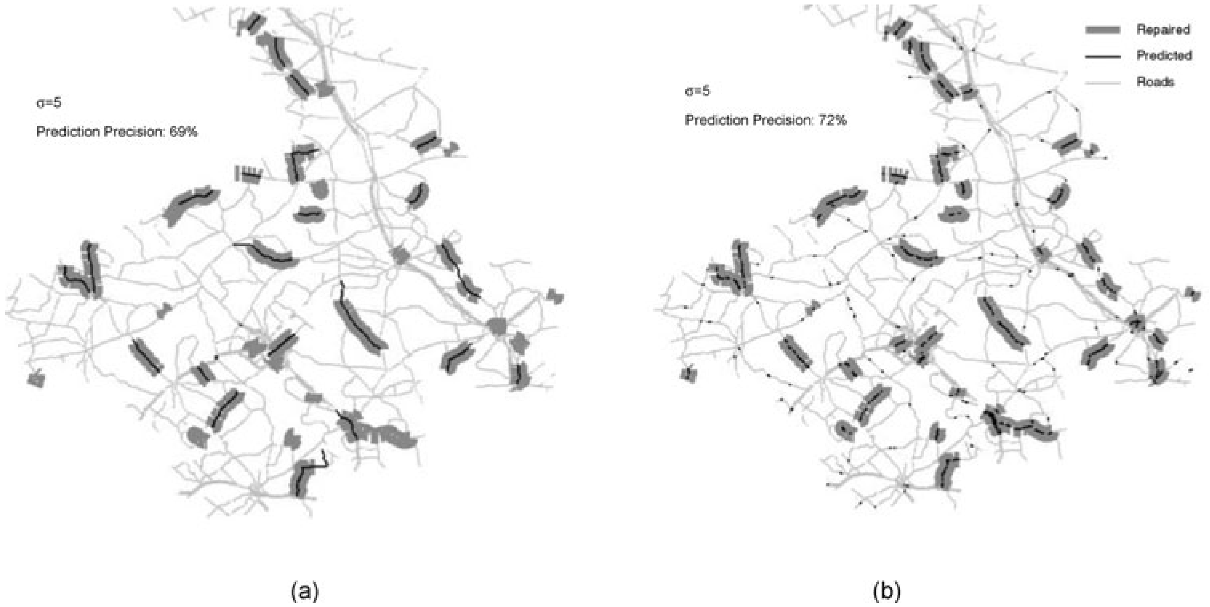

Fig. 17. Predicted versus actual repaired sections for $\sigma = 5$: (a) using the spatial-Bayesian method and (b) based on DS increases.

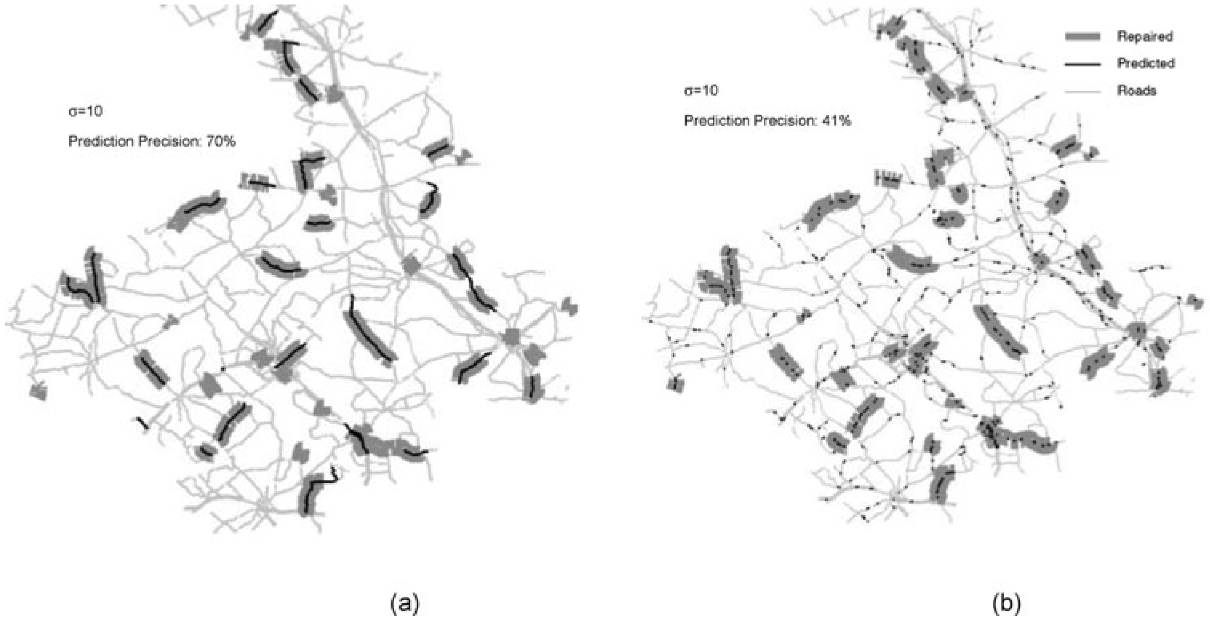

Fig. 18. Predicted versus actual repaired sections for $\sigma = 10$: (a) using the spatial-Bayesian method and (b) based on DS increases.

difficult to integrate (especially legacy databases). A GIS-based Bayesian method has been developed for imputing construction and M&R history of a pavement network by recognizing spatial and temporal patterns in pavement condition data. Traditionally, the focus has been on temporal patterns in condition data (i.e., pavement condition versus age relationships), which tend to be highly noisy. The developed technique searches for evidence of repair in groups of adjacent pavement sections, rather than a single section at a time. This article shows that combined spatial and temporal patterns in condition data can be used for estimating the construction and M&R history of pavement networks more accurately. Analysis of error in pavement condition rating data from Texas showed a standard deviation of error ranging from 5 to 10 points (for a 0–100 condition index). For this range of error in condition data, the developed technique has 74% precision in estimating

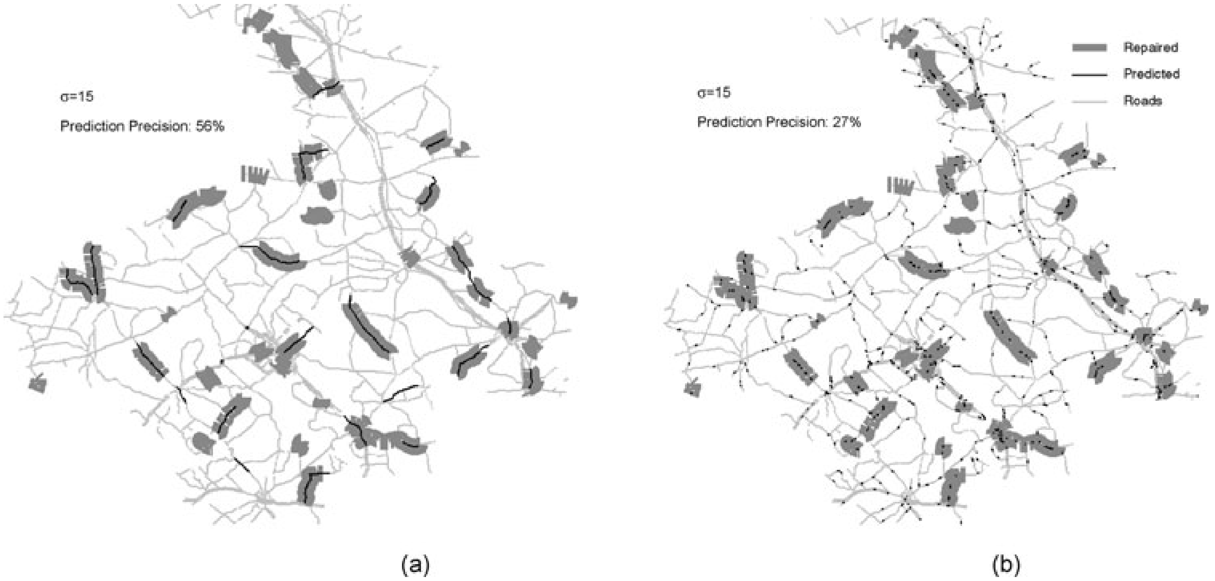

Fig. 19. Predicted versus actual repaired sections for $\sigma = 15$: (a) using the spatial-Bayesian method and (b) based on DS increases.

repair data. The primary sources of classification error in the developed method include systematic error in condition data, inability of the condition index to reflect the true condition of the pavement, presence of small projects, and use of nonrepresentative or inaccurate prior information. The accuracy and precision of the developed method can be improved by addressing these sources of error.

This work has broader implications including the importance of robustness in pavement management models due to the high error level associated with pavement condition data. Determining the required robustness for these models as a function of data set error/noise level and desired model precision can be a subject for future research work. Finally, further work would be needed to explore how the developed method can be implemented in practice to assist transportation agencies in their data integration efforts.

ACKNOWLEDGMENTS

The work presented in this article was funded in part by TxDOT and the FHWA. The contents of this article reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of TxDOT or FHWA. The authors thank Mr. Charles Gurganus of TxDOT for providing insight into the data used in this article. Also, they thank the anonymous reviewers for their helpful and constructive comments.

REFERENCES

Adachi, T. & Ellingwood, B. R. (2009), Serviceability assessment of a municipal water system under spatially correlated seismic intensities, *Computer-Aided Civil and Infrastructure Engineering*, **24**(4), 237–48.

Adams, T. M. (2008), Synthesis of Best Practices for the Development of an Integrated Data and Information Management Approach, Report No. MRUTCD 03-02, Midwest Regional University Transportation Center, Madison.

Bogus, S. M., Song, J., Waggerman, R. & Lenke, L. R. (2010), Rank correlation method for evaluating manual pavement distress data variability, *Journal of Infrastructure Systems*, **16**(1), 66–72.

Bolstad, W. M. (2007), *Introduction to Bayesian Statistics*, John Wiley & Sons, Hoboken, NJ.

Cheng, Z., Amin, H., Zayed, T. M. & Wainer, G. (2005), Representation and analysis of spatial resources in construction simulation, *Proceedings of the 2005 Winter Simulation Conference*, IEEE, Piscataway, NJ, USA, Dec. 4–7, 2005, p. 8.

Cho, H. & Olivera, F. (2009), Effect of the spatial variability of land use, soil type, and precipitation on streamflows in small watersheds, *Journal of the American Water Resources Association*, **45**(3), 673–86.

Daleiden, J. F. & Simpson, A. L. (1998), "Off-the-wall" pavement distress variability study, *Transportation Research Record*, **1643**, 62–70.

de Oliveira, D. P., Neill, D. B., Garrett, J. H., Jr. & Soibelman, L. (2011), Detection of patterns in water distribution pipe breakage using spatial scan statistics for point events in a physical network, *Journal of Computing in Civil Engineering*, **25**(1), 21–30.

Deshpande, V. P., Damnjanovic, I. D. & Gardoni, P. (2010), Reliability-based optimization models for scheduling pavement rehabilitation, *Computer-Aided Civil and Infrastructure Engineering*, **25**(4), 227–37.

Dewan, S. A. & Smith, R. E. (2003), Modifying local agency pavement management system to support governmental accounting standards board 34 requirements, *Eighth International Conference on Low-Volume Roads 2003, June 22, 2003–June 25, 2003*, National Research Council, Reno, NV, pp. 15–23.

FHWA (2010), Data integration primer, Report Number FHWA-IF-10-019, Washington, DC.

Flintch, G. & McGhee, K. K. (2009), *Quality Management of Pavement Condition Data Collection. A Synthesis of Highway Practice*, Synthesis 401, National Cooperative Highway Research Program (NCHRP), Transportation Research Board (TRB), Washington DC.

Gharaibeh, N. G., Yajie, Z. & Saliminejad, S. (2010), Assessing the agreement among pavement condition indexes, *Journal of Transportation Engineering*, **136**(8), 765–72.

Hall, J. P. (2006), Integrating roadway, traffic, and crash data, Transportation Research Circular, No. E-C111, Transportation Research Board, Washington DC.

Jia, J.-H. & Wang, C. (2010), Spatial analysis based on GIS in the application of construction of urban underground rail, *Second International Conference on Information Science and Engineering (ICISE2010), December 4, 2010–December 6, 2010*, IEEE Computer Society, Hangzhou, China, pp. 4030–32.

Jie, G. & Caldas, C. H. (2008), Data processing for real-time construction site spatial modeling, *Automation in Construction*, **17**(5), 526–35.

Kim, S., Damnjanovic, I. & Gunby, M. (2010), Effects of pavement spatial variability on contractor's management strategies, *Journal of Infrastructure Systems*, **16**(4), 231–40.

Koch, K. R. (2007), *Introduction to Bayesian Statistics*, Springer, Bonn.

Lajnef, N., Rhimi, M., Chatti, K., Mhamdi, L. & Faridazar, F. (2011), Toward an integrated smart sensing system and data interpretation techniques for pavement fatigue monitoring, *Computer-Aided Civil and Infrastructure Engineering*, **26**(7), 513–23.

Larson, C. D., Sami, N. & Luhr, D. R. (2000), Structured approach to managing quality of pavement distress data: Virginia Department of Transportation experience, *Transportation Research Record*, **1699**, 72–80.

Lee, S. & Adams, T. M. (2004), Spatial model for path planning of multiple mobile construction robots, *Computer-Aided Civil and Infrastructure Engineering*, **19**(4), 231–45.

Lee, Y.-H., Mohseni, A. & Darter, M. I. (1993), Simplified pavement performance models, *Transportation Research Record*, **1397**, 7–14.

Li, L., Zhu, L. & Sui, D. Z. (2007), A GIS-based Bayesian approach for analyzing spatial-temporal patterns of intracity motor vehicle crashes, *Journal of Transport Geography*, **15**(4), 274–85.

Lytton, R. L. (1987), Concepts of Pavement Performance Prediction and Modeling, Second North American Conference on Managing Pavements, Toronto, Ontario, Canada, November 2–6, 1987, 2.1–2.19.

McNinch, T. J., Dong, J. K. & Colling, T. (2008), Validation of the pavement performance models used in Michigan's roadsoft infrastructure management system & development of parent deterioration curves for local agencies, *Mid-Continent Transportation Research Forum Program*, Midwest Regional University Transportation Center, University of Wisconsin – Madison.

Mishalani, R. G. & Koutsopoulos, H. N. (2002), Modeling the spatial behavior of infrastructure condition, *Transportation Research, Part B (Methodological)*, **36B**(2), 171–94.

Olivera, F. (2001), Extracting hydrologic information from spatial data for Hms modeling, *Journal of Hydrologic Engineering*, **6**(6), 524–31.

Olivera, F., Koka, S. & Nelson, J. (2006), A GIS application for the analysis of hydrologic networks using vector spatial data, *Transactions in GIS* **10**, 355–75.

Olivera, F. & Maidment, D. (1999), Geographic information systems (GIS)-based spatially distributed model for runoff routing, *Water Resources Research*, **35**(4), 1155–64.

Prakash, A., Sharma, B. N. & Kazmierowski, T. J. (1994), Investigation into Observational Variations in Pavement Condition Survey, Third International Conference on Managing Pavements, San Antonio, TX, May 22–26, 1994, 290–301.

Rada, G. R., Wu, C. L., Elkins, G. E., Bhandari, R. K. & Bellinger, W. Y. (1998), Update of long-term pavement performance manual distress data variability: bias and precision, *Transportation Research Record*, **1643**(1), 71–79.

Schabenberger, O. & Gotway, C. (2004), *Statistical Methods for Spatial Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL.

Shahin, M. Y. (2005), *Pavement Management for Airports, Roads, and Parking Lots*, Springer, New York.

Smith, R. E., Freeman, T. J. & Pendleton, O. J. (1998), Evaluation of Automated Pavement Distress Data Collection Procedures for Local Agency Pavement Management, Fourth International Conference on Managing Pavements, Durban, South Africa, May 17–21, 1998, 269–286.

Vandervalk-Ostrander, A., Guerre, J. & Harrison, F. (2003), *Review of Data Integration Practices and Their Applications to Transportation Asset Management*, Report No. FHWA-IF-03-023, Federal Highway Administration (FHWA), Washingotn DC.

Vlahogianni, E. I., Karlaftis, M. G. & Golias, J. C. (2007), Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks, *Computer-Aided Civil and Infrastructure Engineering*, **22**(5), 317–25.

Wang, K. C. P. & Li, Q. (2011), Pavement smoothness prediction based on fuzzy and gray theories, *Computer-Aided Civil and Infrastructure Engineering*, **26**(1), 69–76.

Wang, X. & Kockelman, K. M. (2009), Forecasting network data spatial interpolation of traffic counts from Texas data, *Transportation Research Record*, **2105**(1), 100–8.

Wolters, A. S., McGovern, G. & Hoerner, T. (2006), *Pavement Management; Monitoring, Evaluation and Data Storage; and Accelerated Testing*, National Research Council, Washington DC, pp. 37–46.

Yang, C., Tsai, Y. & Wang, Z. (2009), Algorithm for spatial clustering of pavement segments, *Computer-Aided Civil and Infrastructure Engineering*, **24**(2), 93–108.

Ying, L. & Salari, E. (2010), Beamlet transform-based technique for pavement crack detection and classification, *Computer-Aided Civil and Infrastructure Engineering*, **25**(8), 572–80.

Zhang, Z., Smith, S. G. & Hudson, W. R. (2001), Geographic information system implementation plan for pavement management information system: Texas Department of Transportation, *Transportation Research Record*, **1769**(1), 46–50.