INTEGRATION OF HEURISTICS AND STATISTICS TO IMPROVE THE QUALITY OF NETWORK-LEVEL PAVEMENT CONDITION DATA

A Dissertation

by

SALAR ZABIHI SIABIL

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Nasir Gharaibeh
Committee Members,	Timothy Lomax
	Maryam Sakhaeifar
	Luca Quadrifoglio
Head of Department,	Robin Autenrieth

December 2016

Major Subject: Civil Engineering

ABSTRACT

Transportation agencies use pavement management systems (PMSs) to make efficient decisions about allocating available resources to the maintenance, rehabilitation, and renewal of their roadway networks. One of the most costly parts of the PMS process is collecting pavement condition data. The efficiency and reliability of decisions made based on PMSs depend upon the quality of this data. Thus, transportation agencies need to ensure that dollars invested in this data are well spent, and pavement condition data has the level of quality necessary to meet PMS requirements. Therefore, assessing and improving the quality of pavement management data is a major challenge for both researchers and practitioners.

This study advances the quality assessment of network-level pavement condition data by answering the following questions: (a) How can we identify potential errors in pavement condition data used in PMSs? (b) How do multiple dimensions of error detection affect our ability to detect errors? (c) How does the accuracy of pavement condition data impact predictions of future road network performance? And (d) How do we measure multiple quality dimensions of pavement condition datasets? First, this research devises and implements a computational method to identify potential errors in pavement condition data, integrating conventional statistical methods and heuristics. Second, the effect of considering multiple dimensions of error detection in pavement condition data was investigated. These dimensions are based on data properties, including time series trends in pavement condition data, variability within uniform performance families, and the consistency between several performance indicators.

Third, this research presents a quantitative assessment of the impact of data accuracy on the estimated remaining service life (RSL) of a roadway network as an overall measure of network health. Finally, it provides metrics for measuring data quality dimensions for pavement condition datasets.

The developed technique was validated using pavement condition field data for a road network in Texas. The technique has the advantage of differentiating between extreme yet valid data points and potential errors. In addition, accounting for several properties of pavement condition data to identify potential errors improves the results of this technique. It is hoped that this research will enable pavement engineers to identify potential errors in pavement condition data, and more effectively assure data quality.

DEDICATION

This thesis is dedicated to my family. Thanks for your endless support.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Gharaibeh, for devoting his time to help me complete my dissertation. This document would not be possible without his untiring guidance and support. I am also thankful to my committee members, Dr. Quadrifoglio, Dr. Lomax, and Dr. Sakhaeifar, for their support and valuable comments.

Thanks also go to my friends, colleagues and the department faculty and staff for making my time at Texas A&M University a great experience. I also want to extend my gratitude to Texas A&M Transportation Institute (TTI), which financially supported me during my PhD. Many thanks go to Dr. Schrank, Dr. Turner, Mrs. Geng and other colleagues in the Mobility Division of TTI for helping me develop my background in transportation engineering.

I also recognize that support for this research was provided by a grant from the U.S. Department of Transportation, University Transportation Centers Program to the Southwest Region University Transportation Center.

Finally, thanks to my mother, father and brothers for their encouragement, and to my wife, for her patience, love, and immense support in everything I do.

NOMENCLATURE

AASHTO	American Association of States Highway and Transportation Officials
AADT	Average Annual Daily Traffic
AC	Alligator Cracking
ACP	Asphalt Concrete Pavement
BC	Block Cracking
BCA	Benefit-Cost Analysis
CI	Condition Index
DRUT	Deep Rutting
DS	Distress Score
DOT	Department of Transportation
ESAL	Equivalent Single Axle Load
EUAC	Equivalent Uniform Annual Cost
FLSH	Flushing
FAIL	Surface Failure
FHWA	Federal Highway Administration

GIS	Geographic Information System
HR	Heavy Rehabilitation
IQR	Interquartile Range
IRI	International Roughness Index
LC	Longitudinal Cracking
LR	Light Rehabilitation
M&R	Maintenance and Rehabilitation
MR	Medium Rehabilitation
NCHRP	National Cooperative Highway Research Program
OMB	Office of Management and Budget
PM	Preventive Maintenance
PMIS	Pavement Management Information System
PMS	Pavement Management System
PWV	Present Worth Value
RAV	Raveling
ShRUT	Shallow Rutting

TC	Transvers Cracking
TxDOT	Texas Department of Transportation
USOMB	United States Office of Management and Budget
W-F	Worst-First

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

Motivation and Problem Statement

Transportation agencies use a wide range of data (e.g., safety, mobility, and infrastructure asset management) to make informed decisions about allocating available resources to construct, operate, and maintain a sustainable, safe, and efficient transportation system. These data are rapidly growing due to advanced data collection and storage technologies. New data records are added to store historical data and new data fields are added to store new attributes, leading to increased data management challenges. These challenges are further complicated by changes in data collection methods and equipment over the time. One of these challenges is how to measure and assure the quality of these large and complex datasets. This research addresses this challenge as it applies to pavement condition data at the network level.

Pavement management is defined as a management approach that pavement owners use to make cost-effective decisions about the maintenance, rehabilitation, and renewal of a road network (AASHTO 2001). Pavement management is a data-driven process; with pavement condition data being a key component of any pavement management system (PMS). The quality of this data can affect not only the assessment of current and predicted future condition of the network, but also the quality of decisions regarding maintenance and rehabilitation (M&R) activities (Saliminejad and Gharaibeh 2013). Specific gaps in the pavement management literature that this research seeks to fill are as follows:

- The large size of pavement condition data makes manual error detection methods practically impossible to implement. Thus, automated error detection methods need to be developed and integrated with computerized PMSs.
- Pavement condition is measured using several indicators, such as the International Roughness Index (IRI), rutting, and various types of cracking. Current error detection methods in pavement management usually evaluate each pavement condition indicator individually to identify outliers. Current techniques use the time series trend for each condition indicator to identify any unexpected changes that may denote data quality issues (Pierce et al. 2013) or check for out of range values, missing data, and sample checks of distress ratings (Flintsch and McGhee 2009). By doing so, these methods ignore consistency among multiple pavement condition indicators.
- Current statistical techniques used for outlier detection cannot differentiate between an extreme data point and a potential error. Outliers could be extreme, yet valid, data for pavement sections that are affected by unusual conditions (e.g., premature failure or over-design).

While improving the quality of pavement condition data is desired, it requires additional costs, such as more training for inspectors, more advanced equipment, and more detailed measurements (AASHTO 1993; Livneh 1994). Thus, it is important to measure the impact of data quality improvement on PMS outputs (e.g. predicted network condition) so that the additional cost of quality can be justified. To address this issue, this study provides a quantitative analysis on the effect of erroneous data on PMS output.

Research Questions and General Hypothesis

Based on the aforementioned observations, this research addresses the following specific questions:

- How can we accurately identify potential errors in pavement condition data at the network level?
- How does accuracy of pavement condition data impact the predictions of future performance of the road network (a key capability of PMSs)?
- How can we measure the overall quality of pavement condition datasets?

This research hypothesizes that the integration of heuristic-based consistency checks and statistical outlier detection methods can improve our ability to identify potential errors in pavement condition data. Such method will allow for differentiating between a dissimilar (but valid) outlier and a potential data error, and thus improve the quality of pavement condition data. This general hypothesis is based on the observation that for any given pavement section, different pavement condition indicators change over time in a consistent manner (increase, decrease, or stay the same). Thus, if the change in one of the indicators can be explained by the section deterioration or improvement (e.g., due to a treatment), changes in the other indicators should lead to the same conclusion, and vice versa. To test this hypothesis, an integrated heuristic-statistical method was developed and validated using actual pavement condition data from Texas. This method integrates three properties of pavement condition data to identify data errors and measure accuracy (Figure 1):

- Consistency among multiple condition indicators
- Variability within each uniform performance family
- Time series for each condition indicator

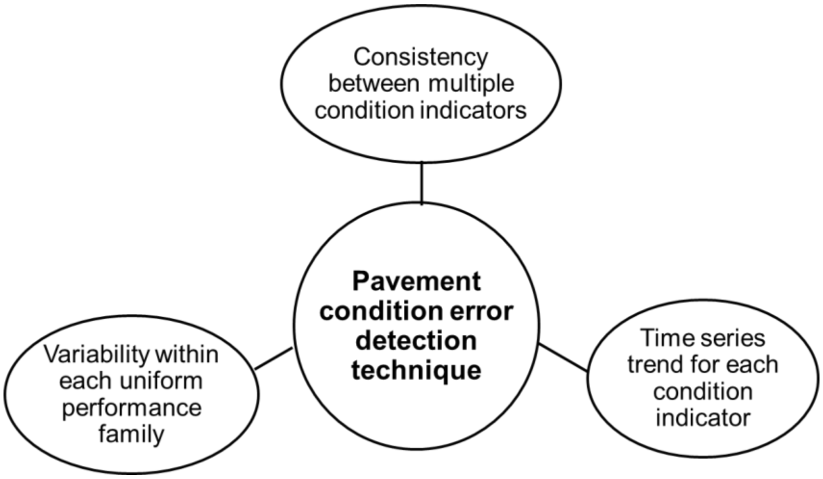

Consistency between multiple condition indicators
Pavement condition error detection technique
Variability within each uniform performance family
Time series trend for each condition indicator

Figure 1 Properties of pavement condition data considered in the developed error detection technique.

Finally, metrics are proposed for measuring several quality dimensions of pavement condition datasets.

Research Objectives

The aim of this research is to develop computational techniques for measuring and improving the quality of pavement condition data at the network level. This entails the following specific objectives:

Develop a Computational Technique to Detect Potential Errors in Pavement Condition Datasets at the Network Level

To accomplish this objective, the following tasks were carried out:

- Establish pavement performance families.
- Detect statistical outliers within each pavement family for each pavement condition indicator.
- Integrate outlier detection method and heuristic-based consistency checks to identify potential errors in pavement condition data.

Validate the Developed Error Detection Technique

This objective was accomplished by testing and validating the developed error detection method using a real world pavement condition data from Texas.

Assess the Impact of Accuracy of Pavement Condition Data on Predictions of Future Performance of the Road Network

This objective was accomplished by comparing the estimated remaining service life (RSL) of the road network (as a measure of network overall health) based on two scenarios: original database (i.e. without any modification) and clean database (i.e. after eliminating potential errors in condition data).

Assess the Effect of Considering Multiple Properties of Pavement Condition Data on Detecting Potential Errors

To accomplish this objective, the following task will be carried out:

- Define three error detection approaches such that each considered a different number of pavement condition data properties to identify likely errors.
- Analyze pavement condition data from the Brownwood District of the Texas Department of Transportation (TxDOT) using these approaches and compare their results (i.e. detected potential error data instances) to each other.

Provide Metrics for Measuring Quality Dimensions of Network-Level Pavement Condition Datasets

To accomplish this objective, the following tasks were carried out:

- Identify metrics for measuring the overall quality of pavement condition datasets
- Apply the identified data quality metrics to a real world pavement condition data from Texas (Bryan district roadway network in east-central Texas).

Organization of the Dissertation

This dissertation is composed of eight sections.

- Section 1 introduced the research hypothesis and objectives.
- Section 2 provides a review of the literature on relevant topics.
- Section 3 describes the development of a new computational technique for detecting potential errors in pavement condition datasets at the

network level. This technique integrates a statistical outlier detection method and heuristic consistency checks to identify likely erroneous data in multiple pavement condition indicators.

- Section 4 validates and tests the developed technique. Also a sensitivity analysis was conducted to show the impact of changes in technique parameters on the results.
- Section 5 compares multi-dimension error detection approaches with a single-dimension approach.
- Section 6 provides a quantitative assessment of the impact of error values on the roadway network RSL. This impact is investigated using both frequency distribution and average RSL of the network.
- Section 7 provides metrics to measuring network-level pavement condition data quality.
- Section 8 presents summary, contributions, conclusions and recommendations.

2. LITERATURE REVIEW

This chapter presents a review of the literature on data quality dimensions, pavement management systems, quality of pavement condition data, and the impact of data quality on PMS outputs.

Data Quality

Data quality is defined as a set of characteristics (e.g., accuracy, completeness, timeliness, and consistency) that a dataset is supposed to possess in order to be trusted to serve its purpose. The stringency of these characteristics can be used to measure the overall level of data quality (Dasu and Johnson 2003). These characteristics, also known as data quality dimensions, make a dataset appropriate for a specific use. Therefore, the importance of specific data quality dimensions varies, depending upon the database and the purpose of the data; a database might be of adequate quality for one purpose, but not for another.

Accuracy is one of the most important dimensions of data quality. The accuracy of a data instance is defined as the closeness of the value in the database to the true value of the phenomenon in the real world. For example, in a pavement condition database, the accuracy of a pavement performance indicator value (e.g., Alligator Cracking [AC]) for a pavement section is defined as the closeness of the AC value in the database to the real amount of AC on the section. Since the true value is often not available, measuring accuracy can be difficult and expensive. Completeness, another dimension of data quality, illustrates what part of the target domain is missing and what part is completely denoted in the database. Completeness can be measured by dividing the portion of recorded data by the total target domain. Timeliness (or currency) is defined as the most recent time the data was updated. In many cases, even a complete and accurate dataset is not fit for use because it does not represent changes that may occur over time. For example, even accurate condition data for a roadway network cannot be used to make cost-effective decisions about future maintenance if it has not been recently updated. Consistency is another frequently-used dimension; it prevents conflicts between different data values. Based on the purpose of the data, additional dimensions of data quality may be necessary, such as level of detail, appropriateness, and interpretability. Greater detail and more definitions can be found in Dasu and Johnson (2003) and Scannapieco and Catarci (2002).

Many studies have defined additional data quality dimensions using a variety of theoretical, empirical, and intuitive approaches. Wand and Wang (1996) used a theoretical method to investigate how an information system (IS) represents the associated real-world system (RW). They defined two conditions to determine proper data quality for an IS. First, every object in the RW should be mapped to at least one data instance in the IS (it could be mapped to more than one data instance). Second, it should be possible, in principle, to map an IS data instance back to the “correct” RW object. Figure 2 shows examples of both proper and incomplete representations.

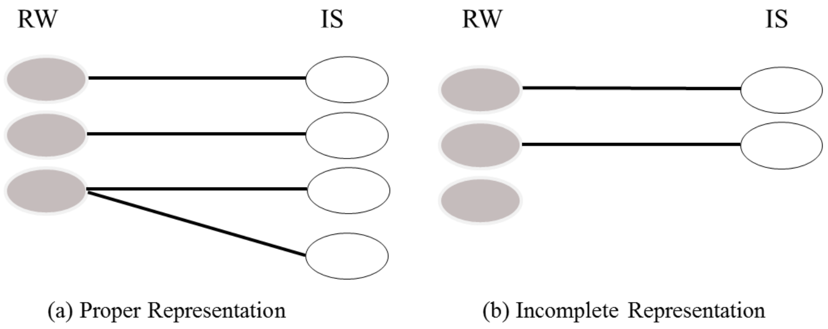


Figure 2 Representations of data quality: (a) Proper representation, and (b) Incomplete representation (adopted from Wand and Wang 1996)

Wang and Strong (1996) developed an empirical approach; they used interviews to determine the data quality dimensions important to certain data consumers. These researchers proposed a two-level classification in which each of four categories was further divided into several sub-dimensions. Table 1 illustrates these categories and quality dimensions.

Table 1 Dimensions proposed in the Wang and Strong (1996) empirical approach (adopted from Batini and Scannapieca 2006).

Category	Dimension	Definition
Intrinsic	Accuracy	data are correct, reliable and certified free of error
	Believability	data are accepted or regarded as true, real and credible
	Objectivity	data are unbiased and impartial
Contextual	Value-added	data are beneficial and provide advantages for their use
	Relevancy	data are applicable and useful for the task at hand
	Timeliness	the age of the data is appropriate for the task at hand
	Completeness	data are of sufficient depth, breadth, and scope for the task at hand
Representational	Interpretability	data are in appropriate language and unit and the data definitions are clear
	Ease of understanding	data are clear without ambiguity and easily comprehended
Accessibility	Accessibility	data are available or easily and quickly retrieved
	Access security	access to data can be restricted Access security and hence kept secure

Redman (1996) used an intuitive approach to determine data quality dimensions. Three categories of dimensions were proposed in his study, including: conceptual schema (referring to the intention of the data), data values, and data format.

Table 2 Dimensions proposed in the Redman (1996) intuitive approach (adopted from Batini and Scannapieca 2006).

Dimension	Type of Dimension	Definition
Accuracy	Data value	Distance between v and v', considered as correct
Completeness	Data value	Degree to which values are present in a data collection
Currency	Data value	Degree to which a datum is up-to-date
Consistency	Data value	Coherence of the same datum, represented in multiple copies, or different data to respect integrity constraints and rules
Appropriateness	Data format	One format is more appropriate than another if it is more suited to user needs
Interpretability	Data format	Ability of the user to interpret correctly values from their format
Portability	Data format	The format can be applied to as a wide set of situations as possible
Format precision	Data format	Ability to distinguish between elements in the domain that must be distinguished by users
Format flexibility	Data format	Changes in user needs and recording medium can be easily accommodated
Efficient use of memory	Data format	Efficiency in the physical representation. An icon is less efficient than a code

No approach will solve all of the data quality issues of a particular database. However, a multi-disciplinary approach is best suited to improving the overall quality of data (Dasu and Johnson 2003). Although data quality can be discussed according to a general framework, the value of each database should be assessed after considering the specific characteristics of that database. Such characteristics play an essential role in detecting potential errors. In most data quality frameworks, conducting a preliminary data review is the first and primary step. Statistical quintiles, graphical representations, and probability distributions are examples of methods for reviewing data before selecting specific data quality tools (EPA 2006).

Data quality can be addressed in three ways: protection, measurement, and improvement. The first approach, protection, focuses on prevention by recommending careful methods for data handling and enforcing quality control procedures (Motro and Rakov 1998). The measurement approach concentrates on means of estimating data quality. It defines levels of data quality, and approximates the value of a database in accordance to those levels. Such estimations provide the opportunity to compare several sources of information in order to select the best. The improvement approach attempts to identify missing values and errors in the data (e.g., outlier values), and either removes these values or replaces them with more likely ones (Motro and Rakov 1998).

This study focuses primarily on the improvement approach. It develops a new technique to detect potential errors in pavement condition databases, and reviews the impact of removing these erroneous data on the output. The first step in this technique is determining outliers in the database. Therefore, a review of existing outlier detection techniques is offered below.

Outlier Detection Techniques

Outliers are defined as data instances that depart from other data, and thus may have been generated by a different mechanism (Hawkins 1980). In the statistics and data mining literature, outliers are also referred to as anomalies, deviants, abnormalities, and surprises (Banerjee et al. 2008, Aggarwal 2013). Outliers do not conform to the behaviors expected of standard data. As such, they are likely to either be erroneous or significant. For example, an outlier in a set of credit card transaction data could be an incorrect value (an error) or a sign of identity theft (Bolton and Hand 2001). Similarly, an outlier in a computer network traffic pattern might represent a cyber-intrusion (Lazarevic et al. 2005). There have been many studies dedicated to developing techniques to find and repair outliers and anomalies. In most cases these techniques are specifically applicable to certain fields; rarely are they general (Chandola et al. 2009). The detection of outliers has a vast arena of applications, such as the detection of network intrusions (Mukherjee et al. 1994), credit card fraud (Ngai et al. 2011), and disease (Wong et al. 2002), as well as image processing (Matteoli et al. 2010).

Differences among the Outlier Detection Techniques

Outlier detection techniques may differ in any number of ways, including the nature of the input data, type of outliers, data labeling, and output. Input data is a collection of data instances usually described as binary, categorical, continuous, etc. (Malik et al. 2014). The nature of these attributes affects the applicability of the outlier detection technique. For example, statistical models are usually applied to continuous or categorical data (Chandola et al. 2009). Another aspect of outlier detection techniques is the desired outlier type, which can be classified as point, contextual, or collective. Point outliers are individual data objects identified as anomalous as compared to the rest of data. They are the most common type of outlier and can occur in any dataset (Gogoi et al. 2010). A contextual outlier, also known as a conditional anomaly, is a data instance occurring within a specific context. Here, each data instance has two attributes: contextual and behavioral. A data instance may be a contextual outlier (based on its behavioral attributes) within a specific context, but another data instance with the same behavioral attributes might be identified as a normal occurrence if it occurs within a different context (Chandola et al. 2009). For example, in a dataset comprised of monthly temperatures recorded in the Houston area over several years, time is considered a contextual attribute and temperature a behavioral attribute. A temperature of 50 degrees Fahrenheit might be considered a normal occurrence in the winter, but an outlier in the summer. A collective outlier is a collection of related data instances found to be anomalous with respect to the entire dataset. Individual data instances occurring in collective anomalies might be normal occurrences by themselves, but their occurrences together, as a collection, may be reason for suspicion (Chandola et al. 2009, Malik et al. 2014).

Data labels denote if the data instance is normal or anomalous. Preparing labeled data requires substantial effort, but it is useful in outlier detection techniques that require a training approach (e.g., supervised and semi-supervised outlier detection techniques). Supervised techniques assume the availability of a training dataset labeled for both normal and anomalous data instances. After building a predictive model using training data, any unseen data instances are identified as either normal or outliers (Gaddam et al. 2007). There are two major obstacles faced by supervised outlier detection techniques. First, the number of data objects labeled as outliers is fewer than those labeled as normal occurrences. Second, since labeling is usually done by experts, it is challenging and expensive to obtain accurately labeled data, especially for outliers (Malik et al. 2014).

Thus, semi-supervised and unsupervised techniques have been developed to solve these issues. Semi-supervised techniques assume that only normal data will be labeled in the training data; resulting models determine normal occurrences and then use them to identify any outliers in the unseen data (Bhuyan et al. 2012). Unsupervised outlier detection techniques do not require labeled training data; they assume that normal data instances are more frequent than any anomalies occurring in the test data (Bhuyan et al. 2012). Consequently, they are more widely applicable than other methods.

Outlier detection methods may differ based on the output reporting the anomaly. Some techniques assign binary labels (i.e., outlier or not outlier) to each data instance. Other techniques provide a ranked list of anomalies and assign an anomaly score to each instance to show the degree to which that instance is considered an outlier (Chandola et al. 2009).

Existing Outlier Detection Techniques

There is extensive and growing literature on developing and applying outlier detection techniques. Much of the work belongs to specific research fields such as information theory and spectral theory. More general outlier detection methods can be categorized into classification, nearest neighbor, clustering, and statistical techniques.

Classification outlier detection techniques develop a classifier model based on a set of labeled data instances (training), and use that model to classify unseen data instances (testing) into a class. This method is divided into two major groups: one-class and multi-class (Chandola et al. 2009). One-class techniques assume that all normal training data have only one class label; thus, any test instance not classified as a member of that class is an outlier. Multi-class techniques assume that training data can belong to more than one normal class, so the classifier model must distinguish between these normal classes and anomalies (Bell 2014, Chandola et al. 2009). Several classification algorithms have been developed to build classifier models, such as neural networks (Augusteijn and Folkert 2002, Vasconcelos et al. 1995), Bayesian networks (Janakiram et al. 2006), and rule-based techniques (Ratsch et al. 2002).

The nearest neighborhood outlier detection technique investigates the distances or similarities between each data point and the next closest point to them. This technique assumes that valid and normal data instances occur close to one another and thus make their "neighborhoods" dense; anomalies and outliers occur farther away from their closest neighbors (Chandola et al. 2009). One approach describes the Euclidean distance of the kth nearest neighbor from a data instance as its degree of outlierness; thus, data instances with higher degree values are more likely to be outliers (Ramaswamy et al. 2000). This method is more robust with noisy data than the standard nearest neighborhood approach (Patcha and Park 2007). Another methodology uses the density of the neighborhood around the observation point to determine its degree of outlierness (Ertoz et al. 2004).

Clustering finds patterns in unlabeled data instances and collects similar data into the same clusters. Clustering-based outlier detection techniques define outliers as data instances that do not belong to any of the established clusters in the dataset (Patcha and Park 2007). Good clustering output has a high level of similarity between data instances within a single cluster, and high levels of difference among the various clusters (Bhattacharyya and Kalita 2013).

Statistical outlier detection techniques assume that the high probability regions of a statistical model contain normal data values, while outliers occur in the low probability regions. Statistical methods fit a certain distribution (e.g., normal distribution) to the data and then apply a statistical inference to see if unseen data objects belong to that distribution. Data objects with a low probability of following the distribution are identified as outliers or anomalies (Chandola et al. 2009). For example, if it is assumed that the data will follow a normal distribution with a 90% confidential interval, data objects occurring in the low probability regions (i.e., 5% each tail) are considered outliers. Statistical outlier detection techniques are divided into two major methods: parametric and non-parametric. Parametric methods assume the distribution of the data and use that data to estimate the distribution parameters; non-parametric data do not assume a knowledge of the distribution.

Another method of detecting errors focuses on the relationships among the different columns (or attributes) of the database. Kinoshita et al. (2003) showed that adding an automated logical checking function capable of restricting the type of data and their range improves the overall accuracy of the database. Rule-based outlier detection techniques provide some rules, mostly among several attributes of the data record; outliers do not obey these rules. This method is often considered a subset of certain classification techniques, if labeled training data is used to develop the rules. Rules with high confidence and support (e.g., confirmed by 90% of the data instances) can define patterns useful in detecting outliers (Maletic and Marcus 2000). Rules between different attributes can be established through logical consistency checks. Consistencies can be spatial, temporal, attribute-based, or any combination of these three (Gong and Mu 2000). Consistency-based outlier detection techniques have mostly been used to detect errors in GIS databases when differently shaped files and their attribute tables are merged together (Phillips and Marks 1996, Griffith et al. 1994). For example, Gong and Mu (2000) used a logical consistency method based on spatial relationships among the neighborhoods to detect errors in GIS maps.

This study integrates both statistical outlier detection methods and consistency reviews to examine the data in a pavement condition dataset for anything likely to be erroneous.

Pavement Management

Pavement management is a mean of making cost-effective decisions about the maintenance, rehabilitation, and renewal of road networks. PMS is a set of tools or methods that help implement such processes (AASHTO 2001). The decisions provide an acceptable level of serviceability for the roadway network by identifying, prioritizing, and planning of treatment activities (Arabali et al. 2016).

Pavement management has two primary management levels: project and network. At the project-level, each pavement section is treated individually for detailed data collection, cause of deterioration assessment, and determination of cost-effective treatment. At the network-level, data collection and treatment/renewal decisions are made for the entire roadway network (Shahin 2005). In PMS, recommended M&R activities for various pavement sections of roadway networks are typically provided in network-level; however, more detailed project-level analysis is needed before the actual application in the field. The project-level analysis considers more factors to drive to the best suited decision for the section. Thus, while there is a good relationship between network-level recommended and actually applied treatment activities for sections, some project-level analysis might lead to a different M&R recommendation. (Hosten et al. 2015)

Additional management levels (such as strategic and project selection levels) can be implemented to support foundational PMS decision-making processes (Saliminejad 2012). The strategic-level is the foremost network-level, used to establish policies and goals that affect the funding allocation process for an agency’s assets. The project selection level is a link between the project-level and the network-level that is used to identify constraints not considered in the higher levels, using more detailed information, in order to refine alternative projects and improve cost estimates (Gurganus and Gharaibeh 2012). Figure 3 shows the relationship between decision making levels of PMSs and the corresponding detail and amount of needed data. Typically in a PMS, higher levels of decision making require less detailed data compared to lower levels which need more specific and detailed data.

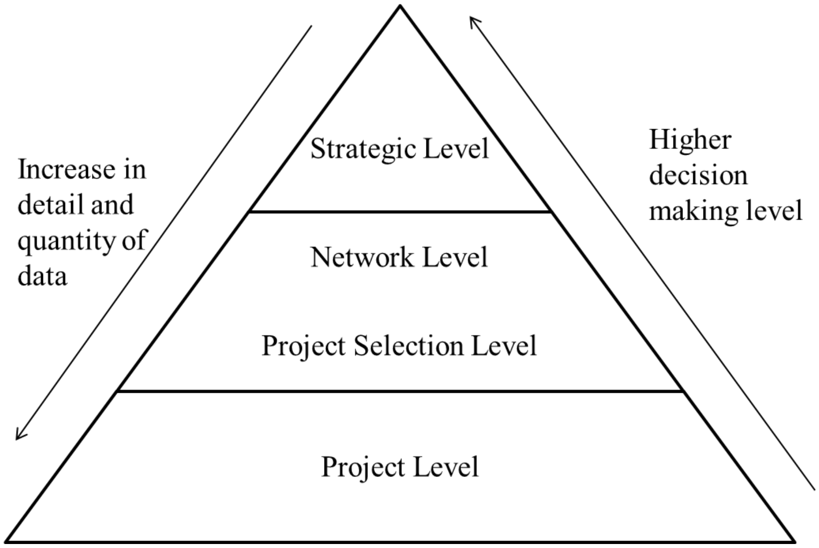

Increase in detail and quantity of data
Strategic Level
Network Level
Project Selection Level
Project Level
Higher decision making level

Figure 3 Relationship between decision making levels and the corresponding detail and amount of required data

The type of pavement condition data and their level of quality are different in several decision making levels. This study focuses on network-level pavement management and aims to assess pavement condition data in order to address data quality issues and their overall impact on pavement management.

Pavement Condition Data Quality Issues

Assessment of the current condition of pavement sections is a substantial and essential task of PMSs. The information such assessments generate allows agencies to determine whether a pavement section is in adequate condition and provides the required level of service, or if M&R actions should be initiated. Moreover, historical condition data can be used to model future performance and identify certain needs, so that accurate and appropriate management plans can be developed using a reliable historical data

(Haas et al. 1994). Thus, the effectiveness of PMS depends greatly on the reliability and accuracy of the pavement condition data (Pierce et al. 2013).

Pavement condition data can be collected either manually or automatically. In manual collection methods, the severity and extent of the pavement condition indicators (e.g., distress types) are visually assessed on site by professional raters. To obtain consistent values, trained raters use the same standard reference for identifying and measuring distress (Shahin and Kohn 1979). Collected data are documented using either pen and paper or a handheld computer, usually equipped with GPS. In manual data collection process, raters might walk from one site to another to inspect and record the pavement condition, or they might perform a windshield survey and use a vehicle to collect data while driving (Flintsch and Bryant 2006). Windshield survey normally is performed in network-level data collection. Manual data collection methods are very rater intensive and their results for a pavement section might vary for different inspectors. In automated data collection, an automated tool or device is used onsite to measure the pavement indicator, or image processing is used offsite to collect the necessary information. This method typically involves a multipurpose vehicle equipped with a distance-measuring device and GPS antennas to capture location data, as well as combinations of video cameras and laser sensors to capture and store pavement condition data. The vehicle also has computer hardware using special software in order to process the collected data (Flintsch and Bryant 2006). According to the literature, automated data collection is generally considered safer and faster; however, manual data collection is more precise (Bogus et al. 2010). Semiautomated methods are other types of pavement data collection methods that use similar equipment as automated methods with lesser degree of automation (Flintsch and Bryant 2006).

Inaccurate or variably assessed pavement condition data makes the PMS yield unreliable results. Errors in the pavement condition data can mislead the evaluation and analysis of the current condition, rate of deterioration, prediction of future condition, M&R needs, and the maintenance cost of pavement sections. It can also impact the M&R treatment selection strategies and the budgeting and needs estimation at the network-level (Tan and Cheng 2015). Consequently, pavement condition data quality is becoming more important; it is also becoming more complicated, due to the increasing size of datasets (e.g., historical data, new fields), and using more advanced data collection and data storage methods.

A pavement condition data quality management plan defines policies and procedures to determine acceptable level of quality and ensure that the data collection procedure provides this level of quality (Pierce et al. 2013). Data quality management plan defines both quality control and quality acceptance activities; the activities can occur before, during, and after data collection.

Quality control includes those activities required to assess and adjust production processes to obtain pavement condition data that meets the desired level of quality (Flintsch and McGhee 2009). The activities are defined in the quality control plan to quantify the variability in the data collection process and maintain it within acceptable level, by identifying and reducing the controllable source of variability in pavement condition data. The most frequent activities used for quality control are: (1) calibration and verification of equipment and methods before the data collection (used by 94 percent of the agencies), (2) testing of known control segments before data collection (94 percent) and during data collection (81 percent), and (3) software routines for checking the reasonableness (57 percent) (Flintsch and McGhee 2009).

Quality acceptance includes those activities are performed by the agency to verify that the collected pavement condition data have met the established quality standards (Flintsch and McGhee 2009). Quality acceptance activities can be divided to three major groups: a) Analysis of the control, verification, and blind site testing, b) Global database checks, c) Sampling checks.

Periodic testing of control sites are used for both quality control and quality acceptance. The site testing data are checked for accuracy, repeatability, and reproducibility; if they cannot meet the acceptance criteria the equipment should be recalibrated and data should be re-collection since the last successful site testing (Pierce et al. 2013). Global checks are conducted after receiving the final condition database and include checking some properties of the entire database such as data format, location accuracy, data completeness, data consistency, and data range. Sampling checks are detailed examination of random samples that provides an approximate estimation of the likelihood of errors in the whole database (Pierce et al. 2013). Primary step of this type of quality acceptance are establishment of acceptance criteria (data accuracy and precision, and reliability) and an appropriate sample size necessary to validate that the data meets these criteria (Flintsch and Bryant 2006). For example TxDOT uses independent auditing to evaluate the accuracy of collected data. The audit data is collected from approximately 5%

of the entire network. The difference between DS values (a composite pavement condition index) of audited and original data is investigated for all samples of each county. The collected condition data of a particular county is rejected and must be recollected if more than 15% of sections within that county has the difference of 10 points or higher (Saliminejad 2012).

The accuracy of the pavement condition data depends on several factors, including the angle and direction of the sunlight and the weather condition during data collection (Smith et al. 1998), the equipment used, the rater or operator training and skills, environmental condition and the shape and condition of pavement sections (Flintch and McGhee 2009), any inability of the collected images and videos to catch thin cracks, and the misclassification of distresses due to ambiguity in the distress definitions (Morian et al. 2002). In automated rut measuring systems errors could not only be related to weather conditions, but also a narrow pavement with no paved shoulder affects the accuracy of collected data(Scullin and Smit 1997). In windshield surveys, data variability can be affected by the rater's ability to see the roadway clearly and the speed of the survey vehicle. The fact that the nature of rating process is subjective and complicated (i.e., the rater must correctly identify both the type of distress and the severity) also leads to data variability (Pierce et al. 2013).

Pavement condition data have been described as variable (Migliaccio et al. 2011) or even erroneous (Larson et al. 2000). In practice, many departments of transportation (DOTs) review the quality of their pavement condition data both during and after data collection to assess and enhance the errors. For example, the Oklahoma DOT checks the final database for out of range or missing data, and performs sampling checks of distress ratings (Flintch and McGhee 2009).

Grabe (2010) categorized measurement errors as either random or systematic. Systematic errors cannot be minimized by repeating the measurement technique for several rounds. Ambiguity in the distress definitions is an example of systematic errors in pavement condition data (Morian et al. 2002). Conversely, the magnitude and direction of random errors are both unpredictable. Thus, random errors can be minimized by repeating data collection. Human errors are one of the primary reasons for random errors in pavement condition data (Flintch and McGhee 2009). Systematic errors are usually fixed and very difficult to detect without secondary field observations. Thus, this study focuses on random errors and tries to provide a method of detection applicable to pavement condition data.

To sum up, there is a general agreement in the literature that errors exist in pavement condition data and transportation agencies need to detect and correct them to decrease their negative effects on the PMS. The following part represents the impact of these errors on PMSs output.

Impact of Data Quality on Pavement Management Outputs

It has generally been accepted that the quality of infrastructure condition data affects the reliability of the output of infrastructure management systems (Livneh 1994, McNeil and Humplick 1991, Shekharan et al. 2007, Saliminejad and Gharaibeh 2013). Therefore, improving the quality of pavement management data is an ongoing goal for transportation agencies.

The quality of pavement condition data not only affects the evaluation of current condition of the roadway network but also it affects the prediction of future pavement condition.

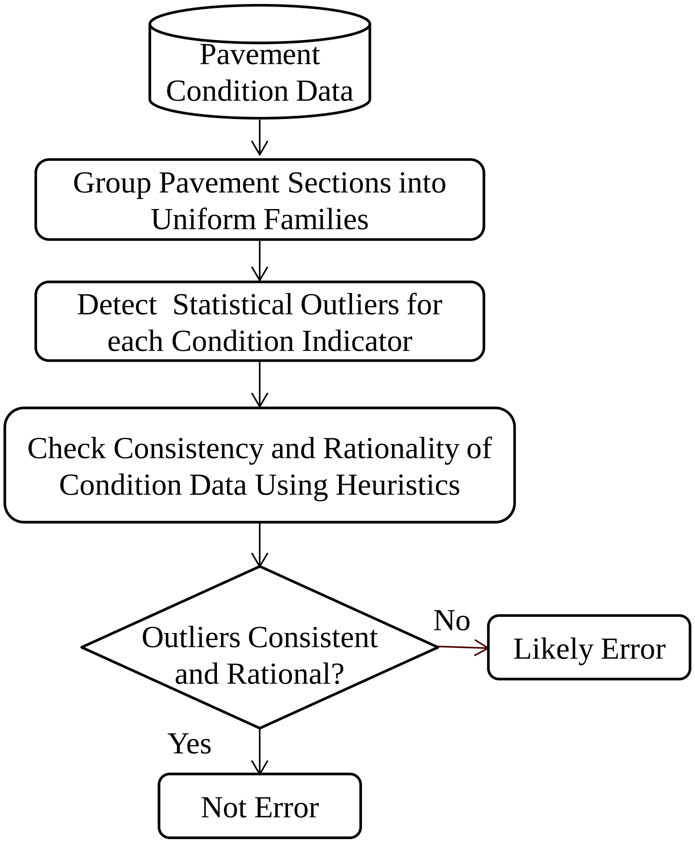
Future condition of pavement sections can be predicted using complex probabilistic models or simpler deterministic techniques (Lytton 1987). Most of the prediction methods assume that the future deterioration trend of a pavement section dependent upon the latest available condition indicator value and parameters of the deterioration curve (Yu et al. 2007). Typically deterioration curves and their parameters are developed by assessing and analyzing historical pavement condition data. Thus future condition of roadway network could not be precisely predicted if database contains errors in current or historical performance indicator values.

One of the most important information that a PMS provides for decision makers is the amount of needed funding for M&R projects to achieve the transportation agency's goal (e.g. have the average condition of pavement sections higher than a predefined threshold). There are several methods to prioritize and select M&R projects such as worst-first (W-F) and benefit–cost analysis (BCA) approaches (Menendez et al. 2013). The quality of pavement condition data affects the need analysis report of the PMS. Although this impact in intuitively accepted, a limited number of studies have been performed that attempt to economically quantify the benefit of high quality pavement condition data. For example, Shekharan et al. (2007) showed that a 21% adjustment in the predicted maintenance costs was the result of implementing an effective quality assurance/quality control plan for the pavement condition data collection system used by the Virginia DOT. Saliminejad and Gharaibeh (2013) showed that both systematic and random errors, even in ranges that in practice may be considered acceptable, can highly distort some PMS outputs. For instance, with 95 percent confidence, a standard error in DS of ±10 (in range of 0-100) can cause overestimation of the annual budget by as much as 85 percent; In this case, needed rehabilitation projects and maintenance projects can have overestimation of 2 percent and 3.8 percent, respectively.

Many PMSs estimate RSL of pavement sections to calculate the LCC and select most efficient M&R projects (Vanier 2001). Little RSL alerts that pavement sections need to receive treatment and adequate RSL represents that pavement sections are in good condition. Baladi et al. (2011) showed that RSL is an effective metric for communicating network health and developing pavement management plans. It demonstrates not only the current condition of roadway network but also the needed service and budget in future, thus it is of interest to decision makers. This research presents a quantitative assessment of the impact of pavement condition data accuracy on the estimated RSL of a roadway network as an overall measure of network health.

3. TECHNIQUE FOR DETECTING ERRORS IN PAVEMENT CONDITION DATA[*]

The overall framework for the proposed technique for detecting potential errors in pavement condition data is depicted in Figure 4.



Pavement Condition Data
Group Pavement Sections into Uniform Families
Detect Statistical Outliers for each Condition Indicator
Check Consistency and Rationality of Condition Data Using Heuristics
Outliers Consistent and Rational?
No
Likely Error
Yes
Not Error

Figure 4 Overall framework for the proposed error detection technique.

[*] Part of this chapter is reprinted with permission from "A Computational Technique for Detecting Errors in Network-Level Pavement Condition Data" by Siabil, S. Z., and Gharaibeh, N. G, 2016. *Transportation Research Record: Journal of the Transportation Research Board,* (2589), 14-19, Copyright [2016] Salar Zabihi Siabil

The components of this methodological framework are discussed next.

Group Pavement Sections into Performance Families

Pavement performance is influenced by several factors (e.g. pavement type, traffic loading, environmental factors, and subgrade properties). Thus, it is reasonable to assume that the performance of a group of pavement sections with common characteristics should change in a similar pattern. Therefore, the first step in the proposed method is to categorize the pavement sections into performance families. In this study, pavement sections of Texas roadway network are grouped based on pavement type, traffic loading, and climate and subgrade zones, as shown in Figure 5.

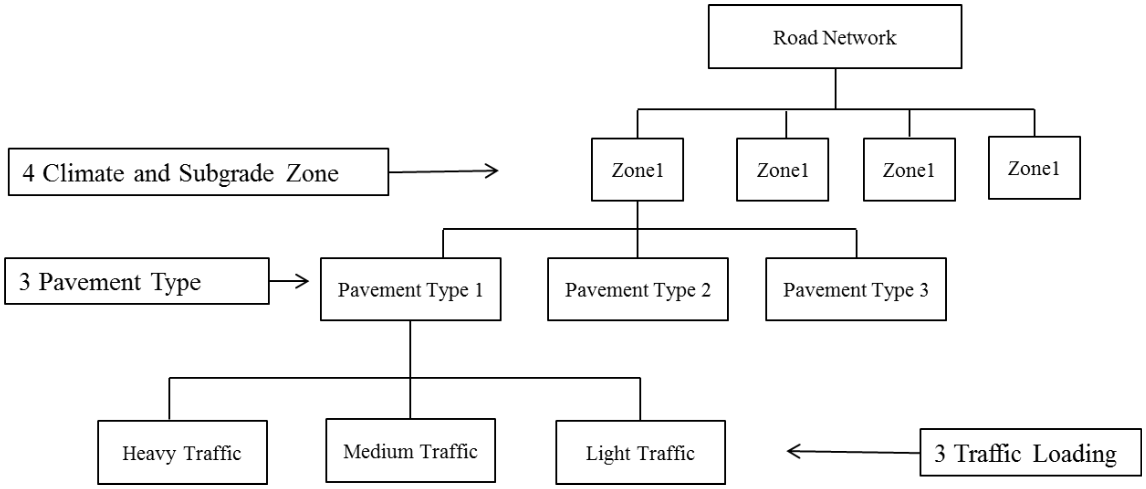


Figure 5 Grouping pavement sections into uniform families (adopted from Gharaibeh et al. 2012).

Based on past studies, Texas is divided into the following four climate and subgrade zones as shown in Figure 6 (Gharaibeh et al. 2012).

Zone 1: Represents wet-cold climate and poor, very poor, or mixed subgrade.

Zone 2: Represents wet-warm climate and poor, very poor, or mixed subgrade.

Zone 3: Represents dry-cold climate and good, very good, or mixed subgrade.

Zone 4: Represents dry-warm climate and good, very good, or mixed subgrade

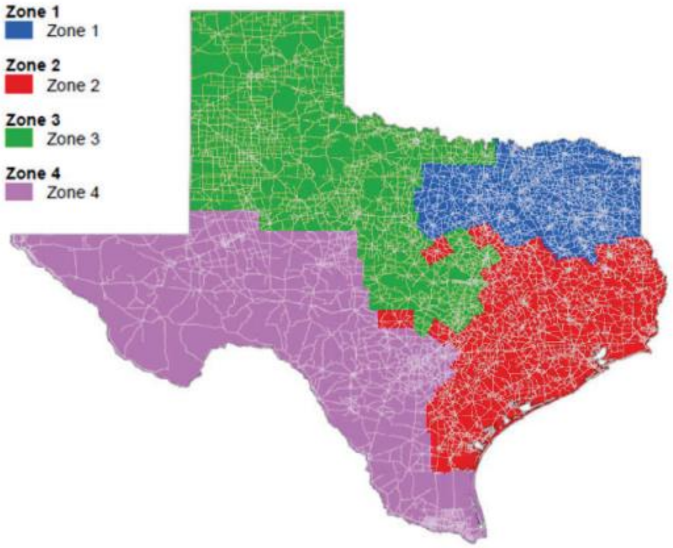


Figure 6 Texas climate and subgrade zones (Gharaibeh et al. 2012).

To consider effects of pavement structure and material type on performance changes, asphalt concrete pavement (ACP) types commonly used in Texas are grouped into three general types as follows (Gharaibeh et al. 2012):

Type A: This pavement type includes thick Asphalt Concrete Pavement (ACP), Intermediate ACP, and overlaid ACP.

Type B: This pavement type includes composite pavement and concrete pavement overlaid with ACP.

Type C: This pavement type includes thin ACP and thin-surfaced ACP.

Since, the majority of pavement sections in Texas road network composed of asphalt concrete this study focuses on several types of ACP and other types of pavement (e.g. continuously reinforced concrete pavement (CRCP), jointed concrete pavement (JCP)) are not considered. However, a similar method can be developed to identify potential errors in CRCP and JCP roads.

Traffic loading has significant effects of pavement performance. Higher number of vehicles passing a pavement section raises the amount of annual deterioration. Pavement sections are classified based on the following three traffic loading levels (Gharaibeh et al. 2012):

Low Traffic Loading: This level includes pavement sections that have a 20-year projected cumulative Equivalent Single Axle Load (ESAL) of less than 1.0 million ESALs.

Medium Traffic Loading: This level includes pavement sections that have a 20-year projected cumulative ESAL greater than or equal to 1.0 million ESALs and less than 10 million ESALs.

Heavy Traffic Loading: This level includes pavement sections that have a 20-year projected cumulative ESAL greater than or equal to 10 million ESALs.

Detect Statistical Outliers of Each Condition Indicator within Each Performance Family

Pavement condition is measured using several indicators such as IRI, rutting, and cracking. Annual changes in each indicator within each performance family are analyzed to identify statistical outliers. A yearly change in performance indicator x is computed as follows:

$$\Delta \ = X_t - X_{t-1} \qquad \text{Eq 1}$$

where Δ_x represents the value of performance indicator *x* in current year (*t*) minus its value in the previous year (*t-1*).

A classic statistical method is used to identify outliers in Δx for each performance family, at any desired confidence interval. These outliers are then investigated (see Step 3 of this method) to delineate them into potential errors and dissimilar (yet valid) data points.

If Δx is normally distributed, the upper limit (UL) and the lower limit (LL) are defined as ($\mu \pm z_\alpha \ \sigma$); where μ and σ are the average and the standard deviation of Δx in the performance family, respectively; and zα is the Z-statistic associated with a desired confidence level (α). Figure 7 shows an example where the extreme 5% of Δx values on each tail of the distribution are considered as statistical outliers (Empirical rule).

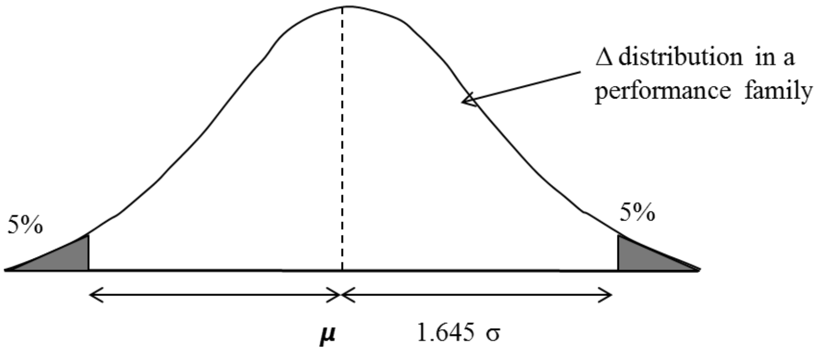


Figure 7 Statistical outliers of normally distributed Δ*x* within a performance family.

The Empirical rule works for data following a normal distribution. If Δx is not normally distributed, Chebyshev's theorem can be used to estimate the minimum confidence interval. According to this theory, at least 1- 1/k$^{\wedge 2}$ of data lie within k standard deviations of the mean (μ ± kσ) regardless of the shape of distribution (Black 2011). The parameter k is termed here as k-value. For example, in an unknown shape of the distribution, at least 75 percent of data are within μ ± 2σ.

The reliability of this outlier detection technique improves as the sensitivity of the performance indicator increases. For example, the dataset used in this study indicates that Δ***x*** for longitudinal cracking has high variability within any given performance family. Thus, this technique is more reliable in detecting outliers in longitudinal cracking compared to less sensitive performance indicators (such as block cracking).

Although, the detected outliers have extreme performance changes (compared to the rest of their family), some of them might be valid and there might be reasonable explanation behind this extreme change (e.g. receiving treatment). Therefore, in the next step, heuristic consistency checks will be defined to investigate whether the outliers are potential errors or they are likely valid values.

Integrate Outlier Detection Method and Heuristic Consistency Checks to Identify Potential Errors in Pavement Condition Data

In this step, the consistency among multiple performance indicators is evaluated for the statistical outliers (described earlier in Step 2) using heuristic checks. If the extreme behavior of an outlier can be explained by heuristic rules, the outlier is considered as a dissimilar (not an erroneous) data; otherwise, the outlier is determined as a “potential error” and needs to be investigated further (e.g. field verification). The development of heuristic checks is described in the following sections of this proposal.

Concept of Consistency

It is logical to assume that for a given pavement section, different pavement condition indicators change from year to year in a consistent manner (i.e., improve, deteriorate, or no change). Thus, if the change in one of the indicator values can be explained by the section deterioration or improvement (e.g., due to treatment), the other indicators’ changes should lead to the same conclusion, and vies versa.

Table 3 shows examples of consistency checks based on yearly changes in performance indicators (Δx). Different M&R treatments affect pavement performance in different ways. Thus, the consistency rules must consider both the performance indicator and how it is affected by different treatment options. Table 3 shows common pavement treatment types and their logical effect (⇓ decrease, ⇑ increase, or ⇔ no change) on different pavement performance indicators one year after the treatment option is implemented.

Table 3 Effect of treatment options on various pavement performance indicators*.

Treatment Option	IRI	RUT	FAIL	BC	LC	TC	AC	RAV	FLSH
Thick overlay or reconstruction	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓
Thin overlay or seal coat	⇑ or ⇔	⇑ or ⇔	⇓	⇓	⇓	⇓	⇓	⇓	⇓
No treatment	⇑ or ⇔	⇑ or ⇔	⇑ or ⇔	⇑ or ⇔	⇑ or ⇔	⇑ or ⇔	⇑ or ⇔	⇑ or ⇔	⇑ or ⇔

*IRI=International Roughness Index; RUT=Rutting, FAIL=Surface failure, BC=Block Cracking, LC=Longitudinal Cracking, TC= Transverse Cracking, AC=Alligator (Fatigue) Cracking, RAV=Raveling, FLSH=Flushing. ⇓= Decrease, ⇑=Increase, ⇔ = No change.

Since surface failure (FAIL), block cracking (BC), longitudinal cracking (LC), transverse cracking (TC), alligator (also called fatigue) cracking (AC), raveling (RAV), and flushing (FLSH) change in a similar manner (increase, decrease, remain unchanged) regardless of the treatment option, they are considered collectively as one group. For this group, the consistency checks are based on their majority. For example, if the number of positive Δs is greater than the number of negative Δs in this group, then the Δx representing this group of performance indicators (called Δsurface) is positive, and vies versa. In case of equal positive and negative Δs for this group, the Δsurface is zero.

The following sections discuss how heuristic checks of the consistency among Δs is used to support or negate statistical outliers. A support of the statistical outlier means that the outlier is an extreme but correct value; whereas, lack of support means that the outlier is a likely erroneous data.

Integrating Heuristic Consistency Checks and Statistical Outlier Detection

Statistical outliers in Δx are either extreme negative or extreme positive compared to the population of their respective families. Extreme negative, yet valid, Δx can be explained by application of a treatment. On the other hand, extreme positive, yet valid, Δx can be explained by rapid deterioration. Therefore, according to the logics of indicators consistent changes, other indicators should change in a similar manner.

There are many possibilities to check consistency between several indicators. Heuristics captures these checks for each indicator. For example, for the seven surface condition indicators (failures, block cracking, longitudinal cracking, transverse cracking, alligator cracking, raveling, and flushing), Δx is checked against $\Delta_{surface}$, Δ_{IRI}, and Δ_{RUT} using heuristics as shown in Figure 8 (which illustrates these checks for longitudinal cracking).

For IRI, Δ_{IRI} is checked against Δsurface and Δ_{RUT} using the rules shown in Figure 9. If one of these indicators does not support the extreme change in IRI, based on the logics of consistency, the IRI value is considered as a potential error. A similar process is conducted for rutting; where Δ_{RUT} is checked against Δsurface and Δ_{IRI}.

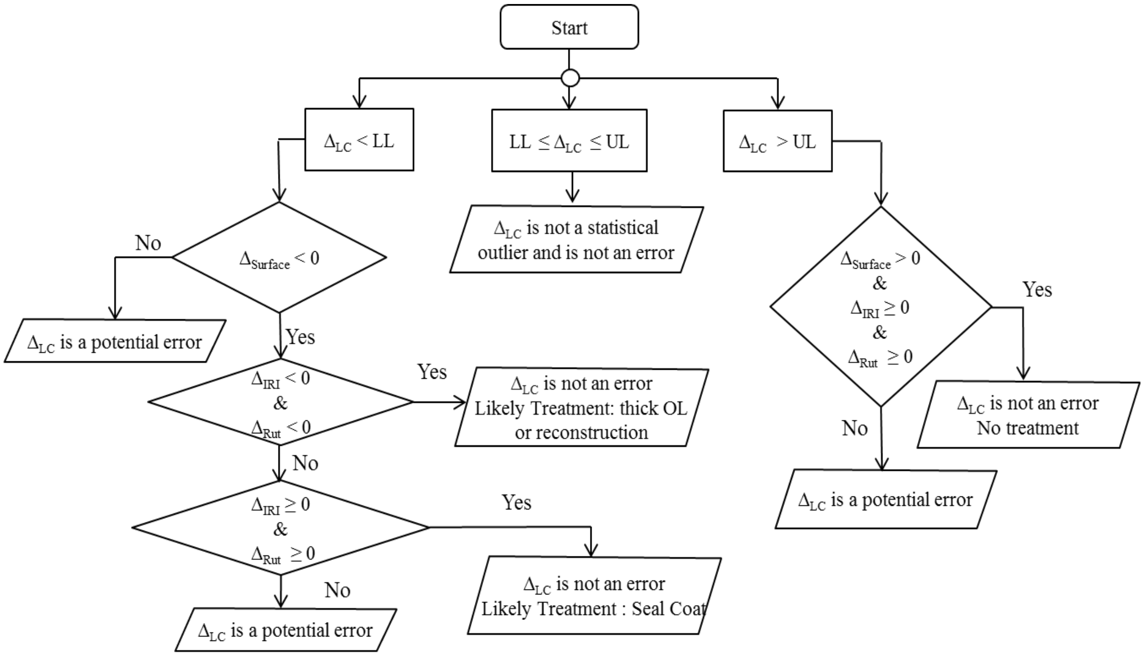

Start
Δ_{LC} < LL
LL ≤ Δ_{LC} ≤ UL
Δ_{LC} > UL
$\Delta_{Surface}$ < 0
No
Δ_{LC} is a potential error
Yes
Δ_{LC} is not a statistical outlier and is not an error
Δ_{IRI} < 0 & Δ_{Rut} < 0
Yes
Δ_{LC} is not an error Likely Treatment: thick OL or reconstruction
No
Δ_{IRI} ≥ 0 & Δ_{Rut} ≥ 0
Yes
Δ_{LC} is not an error Likely Treatment : Seal Coat
No
Δ_{LC} is a potential error
$\Delta_{Surface}$ > 0 & Δ_{IRI} ≥ 0 & Δ_{Rut} ≥ 0
Yes
Δ_{LC} is not an error No treatment
No
Δ_{LC} is a potential error

Figure 8 Heuristic checks for detecting potential errors in LC data

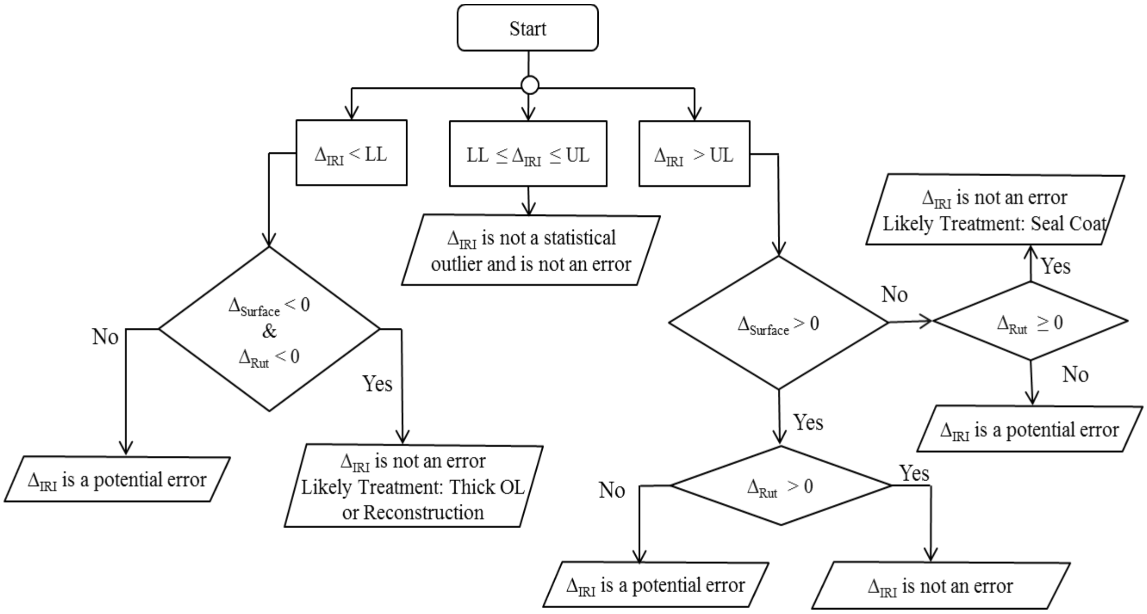

Start
Δ_{IRI} < LL
LL ≤ Δ_{IRI} ≤ UL
Δ_{IRI} > UL
Δ_{IRI} is not a statistical outlier and is not an error
Δ_{IRI} is not an error Likely Treatment: Seal Coat
$\Delta_{Surface}$ < 0 & Δ_{Rut} < 0
No
Δ_{IRI} is a potential error
Yes
Δ_{IRI} is not an error Likely Treatment: Thick OL or Reconstruction
$\Delta_{Surface}$ > 0
No
Δ_{Rut} ≥ 0
Yes
No
Δ_{IRI} is a potential error
Yes
Δ_{Rut} > 0
No
Yes
Δ_{IRI} is a potential error
Δ_{IRI} is not an error

Figure 9 Heuristic checks for detecting potential errors in IRI data

4. VALIDATION OF THE DEVELOPED ERROR DETECTION TECHNIQUE[*]

This section discusses the validation of the developed method for detecting potential errors in pavement condition data (presented in Section 3) using actual pavement condition data obtained from Texas. First, a brief introduction of Texas pavement data is presented. Then, the validation process and results are discussed. Finally, sensitivity of the presented technique to its parameters is investigated.

Validation Data

TxDOT uses private data collection vendors to collect pavement surface distress data every year for approximately its entire roadway network. TxDOT uses this data to measure the network current condition and plan future treatments. This data is stored in the Pavement Management Information System (PMIS) database. The data collection sections are usually 0.5 mile in length and are identified by a unique combination of district name, county name, highway name, and beginning and ending reference mile markers. Distress measurement units vary depending on the distress type, such as length per 100-ft station for longitudinal, number of cracks for transverse cracking, and percent of wheel path area for rutting and alligator cracking. In this study, PMIS distress

[*] Part of this chapter is reprinted with permission from “A Computational Technique for Detecting Errors in Network-Level Pavement Condition Data” by Siabil, S. Z., and Gharaibeh, N. G, 2016. *Transportation Research Record: Journal of the Transportation Research Board,* (2589), 14-19, Copyright [2016] Salar Zabihi Siabil

quantities or densities are investigated using the developed method to identify potential errors.

TxDOT conducts another independent data collection, called Audit data, for approximately 5 percent of its roadway network to control the quality of PMIS data collection. This 5 percent of pavement sections are randomly chosen each year and thus might be completely different from one year to the other. Distress scores (DS) are calculated and compared to check data collection accuracy. In each county, less than 15 percent of sections are allowed to have more than 10 point difference between their DS values in PMIS and Audit data, otherwise the vendor's condition data is rejected.

Figure 10 shows the PMIS road network in 2014 versus the audit sections in 2014. TxDOT collected pavement condition data for more than 193,000 road sections in 2014. The black spots (9,166 sections) in this map represent the roadway sections that are used for validating the proposed error detection method since both PMIS and audit data are available for these sections.

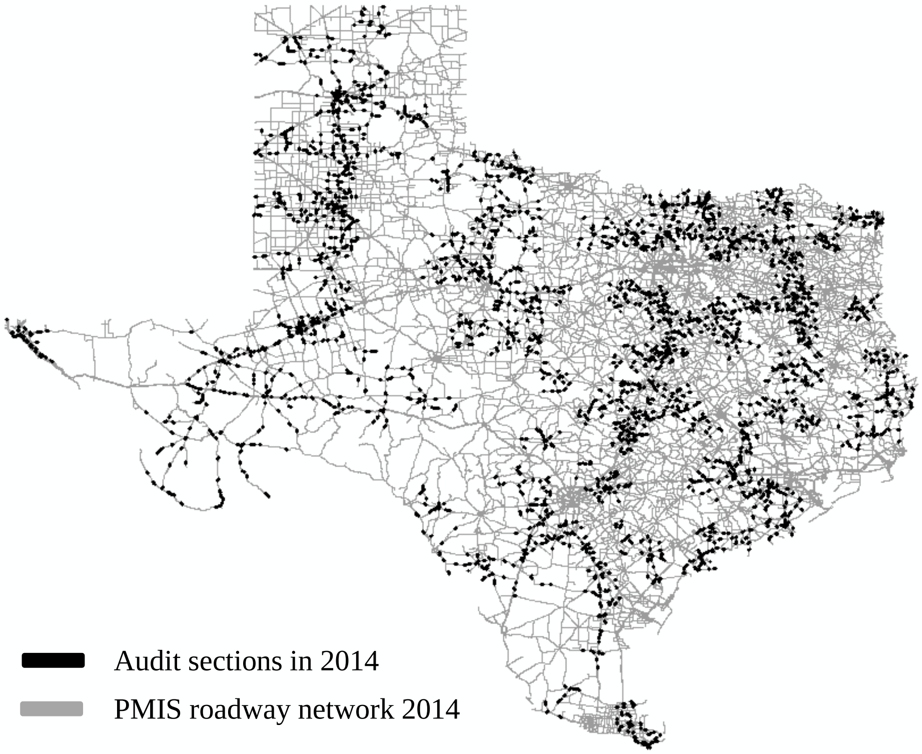


Figure 10 PMIS roadway network in 2014 vs. audit sections in 2014

Validation Process

While the year in which the original data was collected is known, the exact date of collecting this data is unknown. Thus, no true "ground truth" could be established to measure the accuracy of the developed technique on a section-by-section basis. However, as mentioned, audit data is collected on approximately five percent of roadway network to verify the vendor-collected data. The developed technique was validated by comparing Audit data to the original data for the following datasets:

- All sections that have been audited

- Audited sections that have been identified by the developed technique to contain potential errors

The mean absolute error (MAE) was used to measure the difference between the Audit data and the original data for the above datasets (i.e., sections containing potential error and the general population of all audited sections). A divergence between the above datasets in terms of MAE would indicate the ability of the developed technique to identify potential errors. MAE is a well-established statistical metric for measuring model performance (Willmott and Matsuura 2005). In this study, MAE is computed as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \qquad \text{Eq 2}$$

where e_i is the difference between the original distress value and the audit distress in 2014 for each pavement sections; and n is the number of pavement sections in the dataset.

In addition to the MAE test, the Wilcoxon signed-rank hypothesis test was performed. In this statistical test, the null hypothesis is: “the difference between distress values that have been identified as potential error in the original data and their corresponding distress values in the Audit dataset comes from a distribution with zero median.” Analysis of 2014 Audit data and original data for several distress types showed that the difference between these two populations is not normally distributed; and thus the Wilcoxon signed-rank test (Corder and Foreman 2014) is used here. This test is non-parametric (i.e., does not assume that the data is normally distributed).

Validation Results

The above validation process was implemented on several distress types for which 2014 audit data is available, including longitudinal cracking (LC), alligator cracking (AC), and transverse cracking (TC). This process was not implemented on IRI and rutting data because these condition indicators are measured using automated data collection methods and thus are not audited.

Potential Errors in LC Data

LC runs approximately parallel to the pavement centerline. LC values in PMIS represent the total length of longitudinal cracks in feet per 100-ft station of the rated lane. The developed error detection method was applied to the LC values of PMIS data in 2014 and 2013 and identified 336 sections with potential errors in LC values (out of 9,166 sections for which both audit and PMIS data is available). The locations of these sections of potential LC errors are shown in Figure 11.

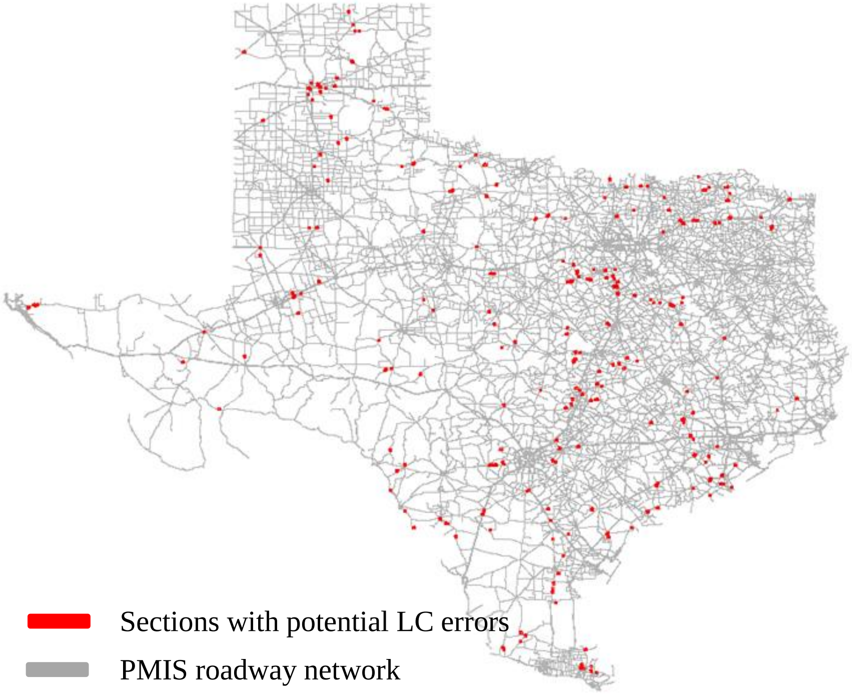


Figure 11 Sections with potential LC errors based on PMIS$_{2014}$ and PMIS$_{2013}$

To check the normality of the difference between Audit2014 and PMIS2014 LC data, a chi-square test is applied and its q-q plot was developed, as shown in Figure 12. The chi-square test was rejected with a 95 percent confidence level and thus it can be concluded that these values are not normally distributed. Therefore, the Wilcoxon signed-rank test is used to determine whether Audit2014 and PMIS2014 of potential errors are significantly different.

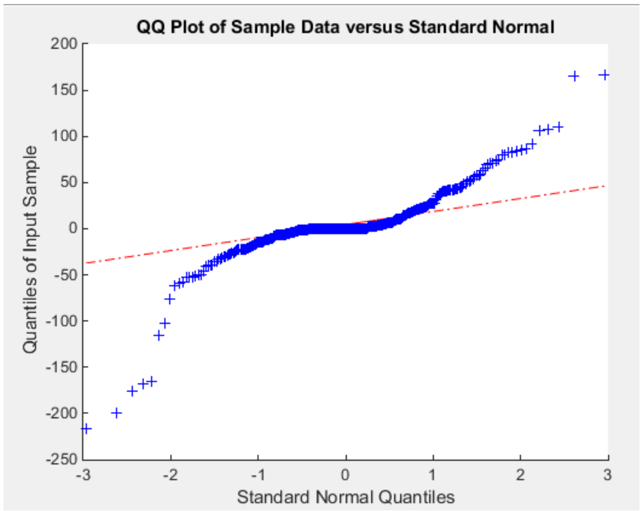


Figure 12 Q-Q plot of the difference between LC values in Audit$_{2014}$ and LC values in PMIS$_{2014}$ for sections with potential LC error

In the Wilcoxon signed-rank test, the null hypothesis is: "the difference between LC values that are potential error in Audit2014 and PMIS2014 comes from a distribution with zero median." This hypothesis is rejected with 95 percent confidence (α =0.05). The p-value associated with the hypothesis is calculated as 0.016. Since the p-value is less than α, it can be inferred that the difference between LC values that have been identified as potential error in the PMIS2014 dataset and their corresponding LC values in the Audit2014 dataset comes from a distribution with nonzero median. In other words, for potential errors, Audit2014 data and PMIS2014 data come from populations with different medians.

Table 4 compares the MAE of LC values for all audited sections in 2014 (Audit2014) and PMIS sections with potential LC errors (PMIS2014). It can be seen that sections with potential LC errors have higher MAE compared to all audited sections.

Table 4 MAE of LC values for sections with potential LC error vs. MAE of LC values for all audited sections.

Population	MAE of LC, ft /100ft	Number of Sections
All Audited Sections	6.48	9166
Sections with Potential LC Errors	21.671	336

Potential Errors in AC Data

AC consists of interconnecting cracks which form small shaped blocks like alligator's skin. AC values in PMIS represent the percentage of the wheel path area that is covered by AC in the rated lane of the data collection section. The developed method was applied to AC values in PMIS 2014 and 2013 and detected 242 sections with potential AC errors. The locations of these sections are shown in Figure 13.

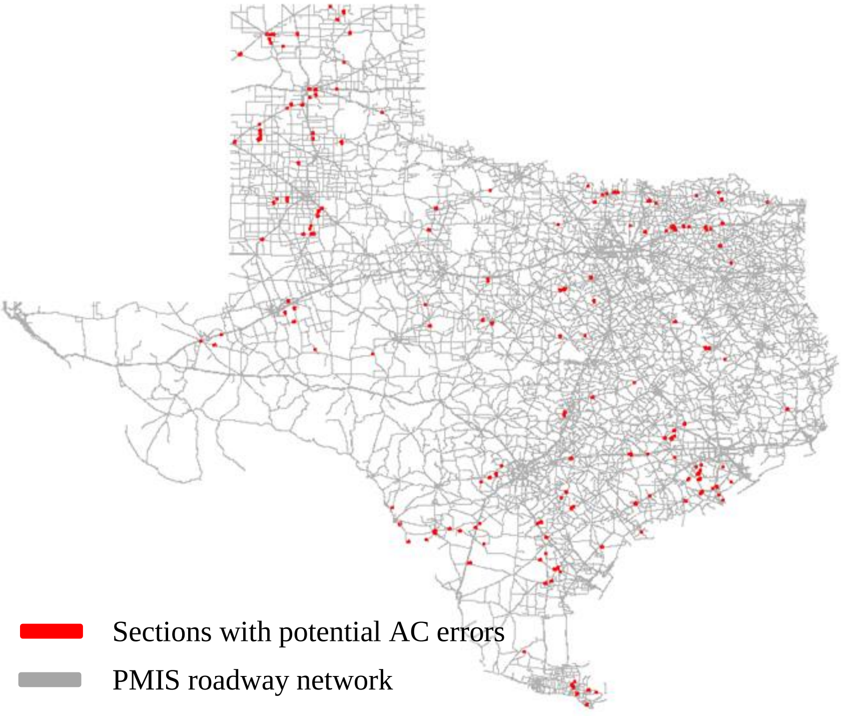


Figure 13 Sections with potential ac errors based on PMIS$_{2014}$ and PMIS$_{2013}$

The normality of the difference between Audit2014 and PMIS2014 AC data was checked using the chi-square test and q-q plot (see Figure 14). The chi-square test was rejected with a 95 percent confidence level and thus it can be concluded that these values are not normally distributed. Therefore, the Wilcoxon signed-rank test is used to determine whether Audit2014 and PMIS2014 of potential errors are significantly different. In the Wilcoxon signed-rank test, the null hypothesis is: "the difference between AC values that are potential error in Audit2014 and PMIS2014 comes from a distribution with zero median." This hypothesis was rejected with 95 percent confidence (α =0.05) and it can be inferred that the difference between AC values that have been identified as potential error in the PMIS2014 dataset and their corresponding AC values in the Audit2014 dataset comes from a distribution with nonzero median..

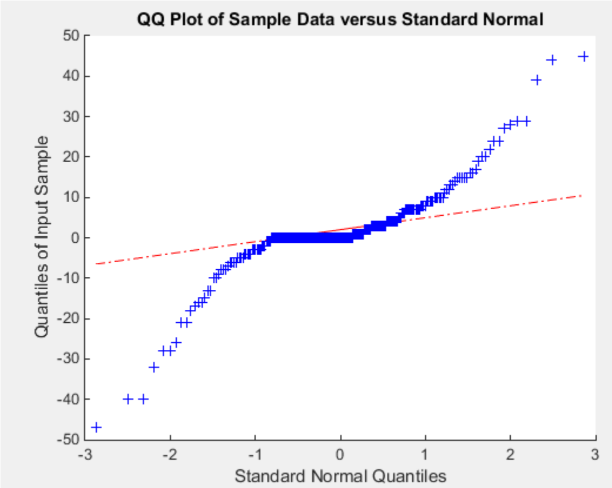


Figure 14 Q-Q Plot of the difference between AC values in Audit$_{2014}$ and AC values in PMIS$_{2014}$ for sections with potential AC error

Table 5 compares the MAE of AC values for all audited sections in 2014 (Audit2014) and PMIS sections with potential AC errors (PMIS2014). It can be seen that sections with potential AC errors have higher MAE compared to all audited sections.

Table 5 MAE of AC values for sections with potential AC error vs. MAE of AC values for all audited sections.

Population	MAE for AC, %	Number of Sections
All Audited Sections	0.9564	9166
Sections with Potential AC Errors	6.152	242

Potential Errors in TC Data

TC values in PMIS represent the number of visually observed transverse cracks per 100-ft long station. The developed method was applied to TC values in PMIS 2014 and 2013 and detected 230 sections with potential TC errors. The locations of these sections are shown in Figure 15.

Similar to LC and AC data, the chi-square test (95 percent confidence level) and q-q plot (see Figure 16) suggest that the difference between Audit2014 and PMIS2014 TC data is not normally distributed. However, the null hypothesis of the Wilcoxon signed-rank test for TC data could not be rejected with α =0.05 because p-value was 0.121. The MAE of TC values for PMIS sections with potential AC errors (PMIS2014) remains noticeably higher than that for all audited sections in 2014 (Audit2014) (see Table 6).

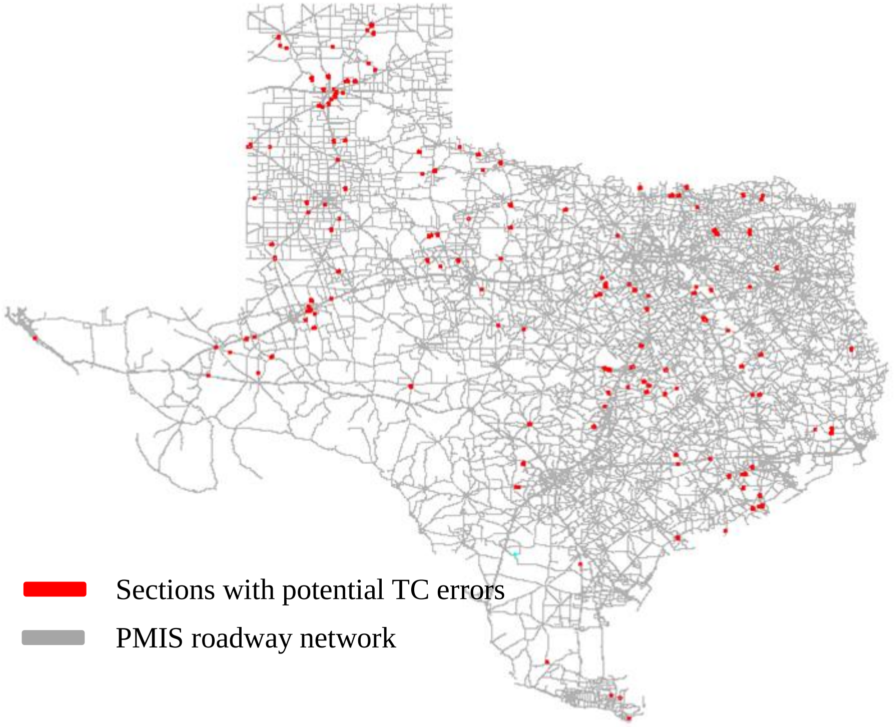


Figure 15 Sections with potential TC errors based on PMIS_{2014} and PMIS_{2013}

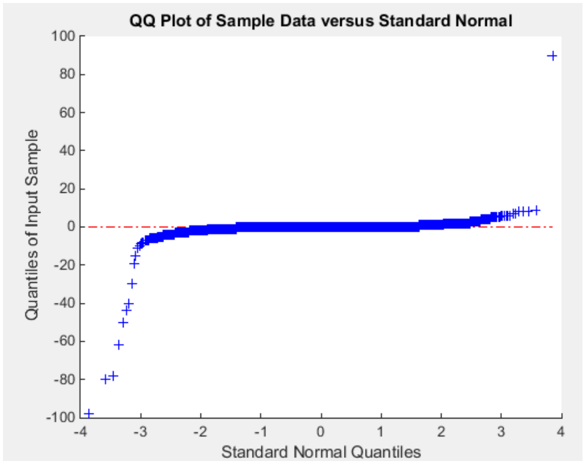


Figure 16 Q-Q Plot of the difference between TC values in Audit_{2014} and TC values in PMIS_{2014} for sections with potential TC error

Table 6 MAE of TC Values for sections with potential AC error vs. MAE of TC values for all audited sections.

Population	MAE of TC, number/100ft	Number of Sections
All Audited Sections	0.2877	9166
Sections with Potential TC Errors	1.5217	230

Summary of Validation Results

Potential errors in pavement condition data from 2013 and 2014 (i.e., PMIS2014 and PMIS2013) were detected using the proposed error detection method. Validation process was implemented on several distress types for which 2014 audit data is available, including longitudinal cracking, alligator cracking, and transverse cracking. The Wilcoxon signed-rank test with α =0.05 validated the developed error detection method for both longitudinal cracking and alligator cracking. Table 7 compares the MAE for all audited sections in 2014 (Audit2014 dataset) to the MAE for sections that have been identified to have potential errors. For all cracking types, sections with potential errors have higher MAE compared to the general population (i.e., all audited sections).

Figure 17 shows the distribution of Texas roadway sections with potential errors in 2014 condition data. Out of 9,166 audited pavement sections, 681 sections were detected to have potentially erroneous cracking data. This map shows that these potential errors are randomly distributed across the Texas roadway network and are not related to a specific geographic region, district, climatic zone, etc.

Table 7 MAE for cracking data identified as potential error vs. MAE for all audited sections.

Distress	Potential Errors		All Audit Data	
	MAE	Number of Sections (n)	MAE	Number of Sections (n)
Longitudinal Cracking	21.67 (ft /100ft)	336	6.48 (ft /100ft)	9,166
Alligator Cracking	6.15 (%)	242	0.96 (%)	9,166
Transverse Cracking	1.52 (No./100ft)	230	0.29 (No./100ft)	9,166

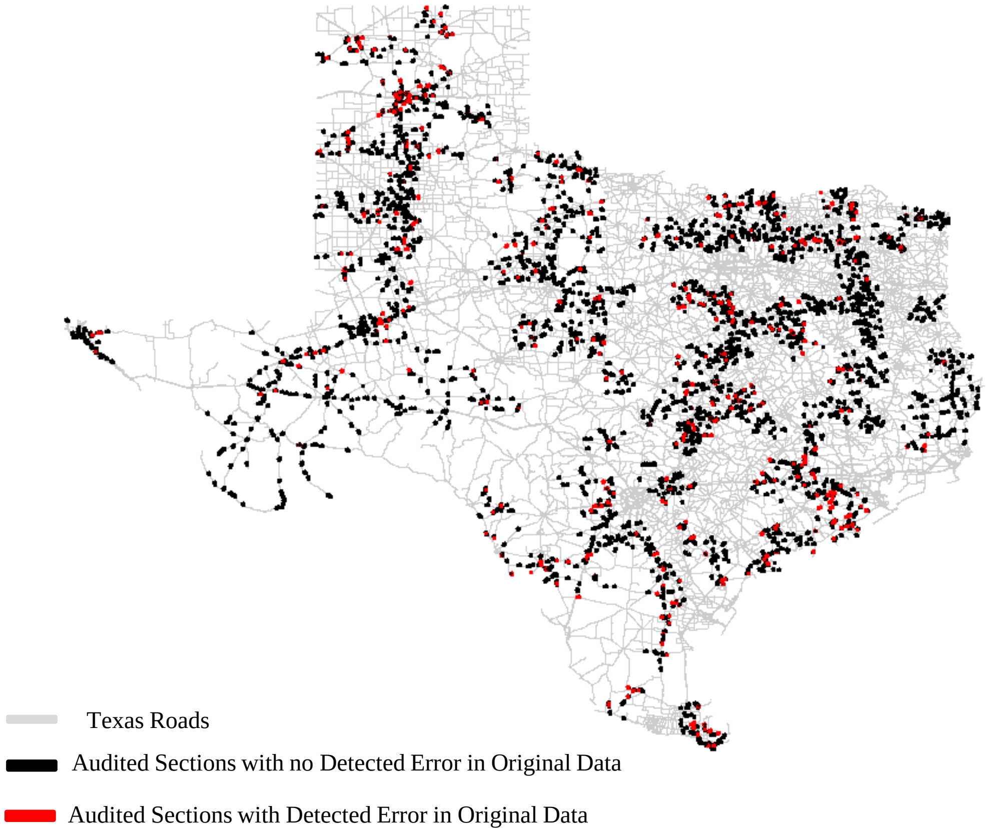

Texas Roads
Audited Sections with no Detected Error in Original Data
Audited Sections with Detected Error in Original Data

Figure 17 Sections identified by the developed technique to have potential error in at least one cracking type out of all audited sections in 2014.

Sensitivity of Error Detection Technique

This section evaluates the sensitivity of the developed error detection technique to changes in k-value in outlier and potential error identification process.

As discussed earlier, in each performance family, the data instances (Δx) lie within the k standard deviations of the mean are considered as regular data, and the ones greater than UL ($\mu + k\ \sigma$) or less than LL ($\mu - k\ \sigma$) are identified as outliers. Therefore, as k-value increases, the number of Δx outliers decreases, and thus the consistency checks are applied to fewer pavement sections. On the other hand, selecting small k-value results in detecting Δx data points with even a slight distance from the median as outliers. Because the developed technique considers outliers as irregular subjects that need to be investigated, an extremely small k-value leads to higher number of suspicious data points that need to be checked for consistency. In this case, some of these pavement sections that have been flagged as outliers might be incorrectly investigated, and eventually identified, as potential errors. The question becomes what k-value should be used? To answer this question, a sensitivity analysis on the k-value was carried out and discussed next.

In this analysis, two variables are considered: number of sections that have outliers and number of sections that have potential errors. Figure 18 shows that the number of sections that have outliers decreases as the k-value increases. For example, k equals to 4 yields 7,199outliers (i.e. 4.5 percent of the population is identified as outliers) and a k of one yields 63,899 outliers (35.6 percent of the population is identified as outliers).

Figure 19 shows that the relationship between k-value and the number of potential errors is similar to that between k-value and number of outliers. For example, a k of 4 yields 3802 potential error (2.3 percent of the population is identified as potential errors) and a k of one yields 38600 potential errors (23.3 percent of the population is identified as potential errors).

Ideally, sections with potential errors should be re-inspected. Thus, the selection of appropriate value for k-value is dependent on availability of resources to audit and re-inspect these sections. For example, a k of 2 yields 15606 potential errors (i.e., 9.5 percent of the population). Typically, highway agencies audit about 5-10 percent of their networks. Thus, a k-value of 2 is used in the next Section of this study to investigate the impact of erroneous data on the PMS output. Different highway agencies might select different values of k-value based on these sensitivity results.

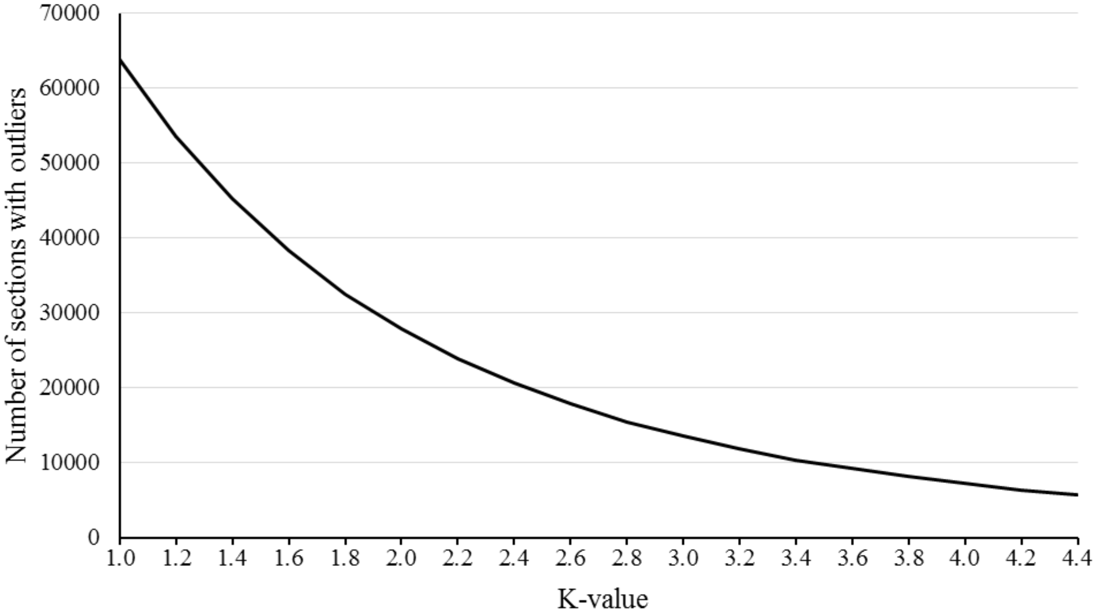


Figure 18 Relationship between k-value and number of sections with outliers

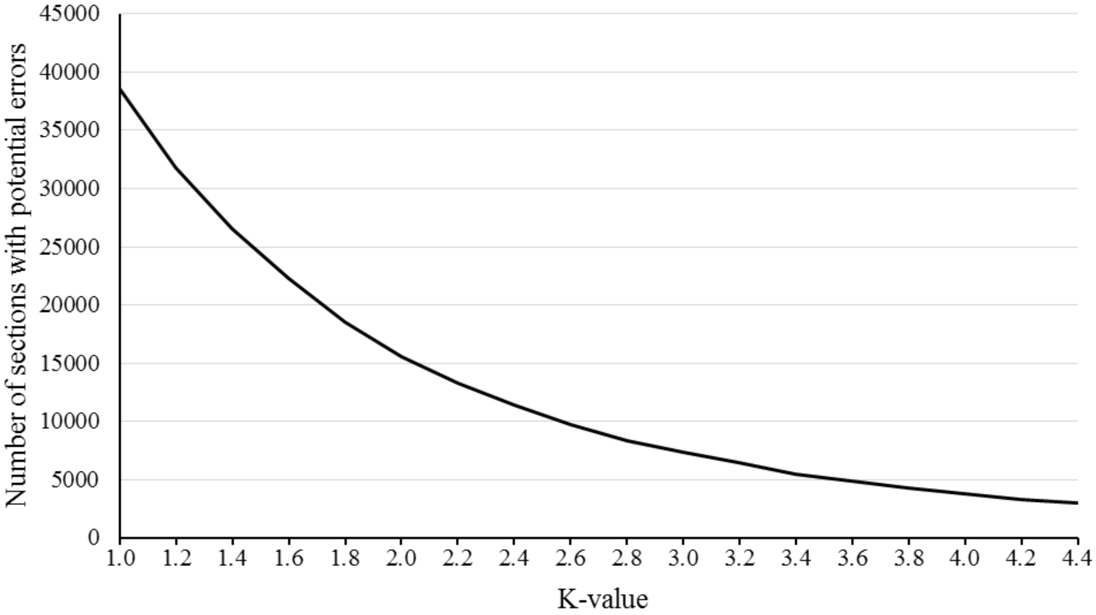


Figure 19 Relationship between k-value and number of sections with potential errors

5. EFFECT OF CONSIDERING MULTIPLE DIMENSIONS OF ERROR DETECTIONS IN PAVEMENT CONDITION DATA

As discussed earlier, the technique developed in this study considers three properties of pavement condition data as a means for detecting potential errors. These properties are:

- Time series trend for each condition indicator
- Variability within each uniform performance family
- Consistency between multiple condition indicators.

In this section, the influence of these dimensions of pavement condition data on error detection is investigated, as follows:

- Case 1 - Time series for the entire network (one-dimensional technique): this case is similar to some of the current error detection techniques used in PMSs. It compares annual distress changes (time series) in pavement condition indicators, and recognizes outliers as likely errors. It does not consider variability in uniform performance families or consistency checks. For example, in this investigation, this technique assumed that points with normal changes would lie within the distance of twice standard deviation from the average ($\mu \pm 2\sigma$). The reminders (which had annual changes outside of this range) were likely errors.
- Case 2 - Time series for individual performance families (two-dimensional technique): this case defines uniform performance families for pavement sections, based on their performance characteristics (e.g., climate zone, traffic

loading, pavement type, etc.). Pavement sections in any given family generally have the similar attributes and are supposed to deteriorate in a similar fashion. Thus, the outlierness of an indicator's annual change is determined respective to the changes of other members of the family. An annual change of one indicator might be recognized as an anomaly in one performance family, but the same amount of change could be seen as a normal data instance in another family.

- Case 3 - Potential errors (three-dimensional technique): this is the new technique that was developed as part of this research. It not only considers annual changes in uniform families in order to detect outliers, but also implements a heuristic consistency check as an additional level of investigation (as discussed in Section 3). Therefore, this technique has the advantage of differentiating irregular, yet valid outliers, from potential errors.

In this analysis, false positive and false negative outliers were investigated. False positive outliers are data points incorrectly identified as errors; while in fact they are correct (e.g., dissimilar but valid data). False negative outliers are data points that are indeed erroneous but were not detected by the technique.

Each of the above analysis cases was conducted for each condition indicator for the entire TxDOT network. The results for LC are shown in Figures 20, 21, and 22 (the results for the other condition indicators are presented in Appendix B). From 165,400 pavement sections considered in this analysis, Case 1 analysis identified 8,892 pavement sections as having outlier LC values. After considering the various pavement families

(Case 2 analysis), the number of sections with outlier LC values decreased to 8,521. Although both cases detected approximately similar number of pavement sections as having irregular LC values, they did not identify the same exact sections. The two cases agreed on 6,623 pavement outlier sections in terms of LC values. There were 2,269 LC values identified as statistical outliers in Case 1, but were considered non-outlier in case 2 analysis. This is because these sections were non-outlier with respect to the other members of their pavement families. On the other hand, the case 1 analysis missed 1,898 sections that, based on their performance families, should have been classified as outliers.

The Case 3 analysis, which considers both consistency checks among multiple condition indicators and performance families, identified only 4,656 pavement sections as potential errors. Of those potential errors, 3,577 sections were correctly detected by Case 1 analysis, missing 1,079 sections (i.e. false negatives). Table 8 shows a summary of the results in a pairwise comparison matrix. Each cell lists the number of pavement sections detected by the pair of analyses as potential errors in LC values. For example, 6,623 pavement sections were detected by both Case 1 analysis and Case 2 analysis as likely to have errors in their LC values (representing 74 percent agreement between Case 1 and Case 3). Similarly, 3,577 pavement sections were detected by both Case 1 analysis and Case 3 analysis as likely to have errors in their LC values (representing 40 percent agreement between Case 1 and Case 3). Since the results of the three-dimensional technique were a subset of the two-dimensional technique result, their intersection was the number of pavement sections detected by the three-dimensional technique.

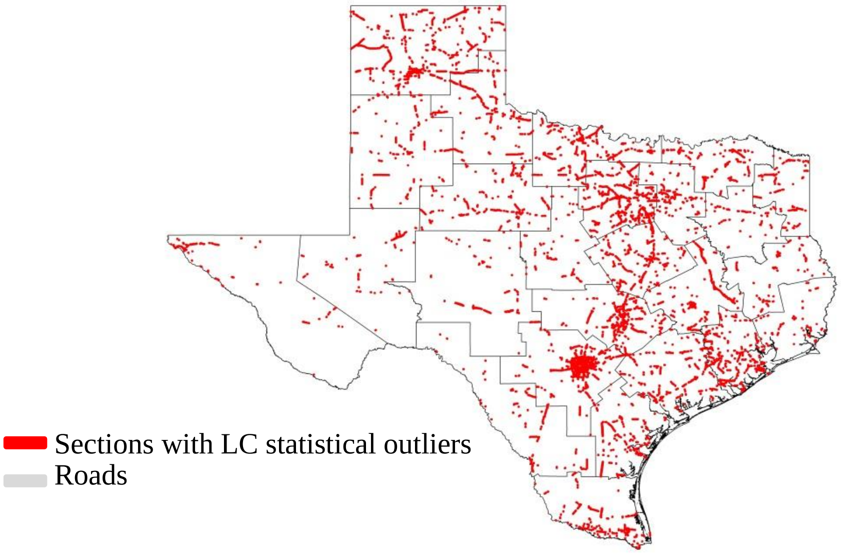


Figure 20 Results of Case 1 analysis: sections with erroneous LC values

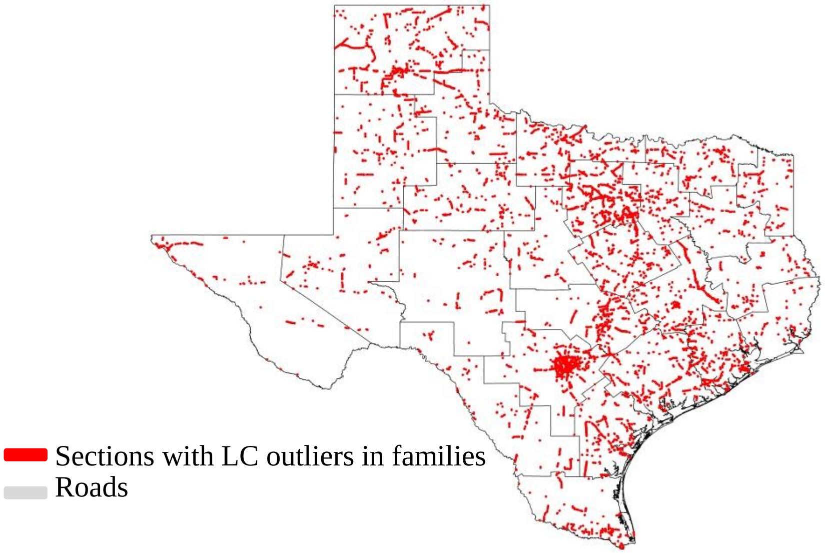


Figure 21 Results of Case 2 analysis: sections with erroneous LC values

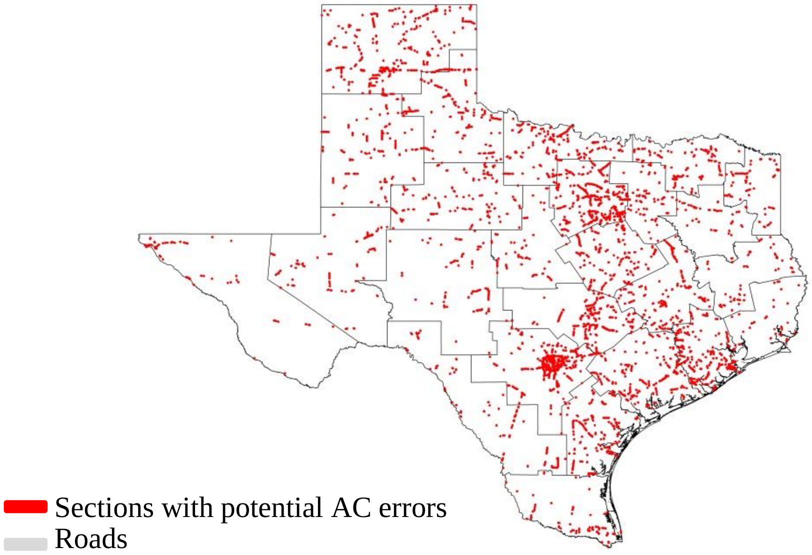


Figure 22 Results of Case 3 analysis: sections with erroneous LC values

Table 8 Agreement between the three case analyses expressed in terms of number of pavement sections detected as potential errors in LC values (entire TxDOT network).

	Case 1 Analysis	Case 2 Analysis	Case 3 Analysis
Case 1 Analysis	8,892	-	-
Case 2 Analysis	6,623	8,521	-
Case 3 Analysis	3,577	4,659	4,659

Texas is a vast state with an extensive roadway network that encompasses many different types of performance families (e.g., climate zone, pavement type, traffic loading, etc.). Thus, it is not surprising that considering all pavement sections to be of one performance family leads to inaccuracy and provides many false positive and negative outliers. Therefore, similar analyses were conducted on a single TxDOT district (Brownwood District) to demonstrate the effect of considering multiple error detection dimensions for a smaller network with fewer performance families. The results for LC are shown in Figures 23, 24, and 25. Table 9 shows the agreement between the three cases in terms of the number of pavement sections detected as potential errors in LC values. Based on the changes in LC values for 5,505 pavement sections, 245 sections were identified as statistical outliers by Case 1 analysis (i.e., without considering performance families or consistency checks). For Case 2 analysis, 267 sections were detected as outliers, including and 219 sections in common with Case 1 analysis (representing 89 percent agreement between Case 1 and Case 2). For Case 3 analysis, only 135 sections were detected as outliers that are potential LC errors, including 113 sections in common with Case 1 analysis (representing 46 percent agreement between Case 1 and Case 3). While this analysis shows that for a smaller network there is more agreement between Case 1 and Cases 2 and 3, differences remain noticeable.

Table 9 Agreement between the three case analyses expressed in terms number of pavement sections detected as potential errors in LC values (Brownwood District).

	Case 1 Analysis	Case 2 Analysis	Case 3 Analysis
Case 1 Analysis	245	-	-
Case 2 Analysis	219	267	-
Case 3 Analysis	113	135	135

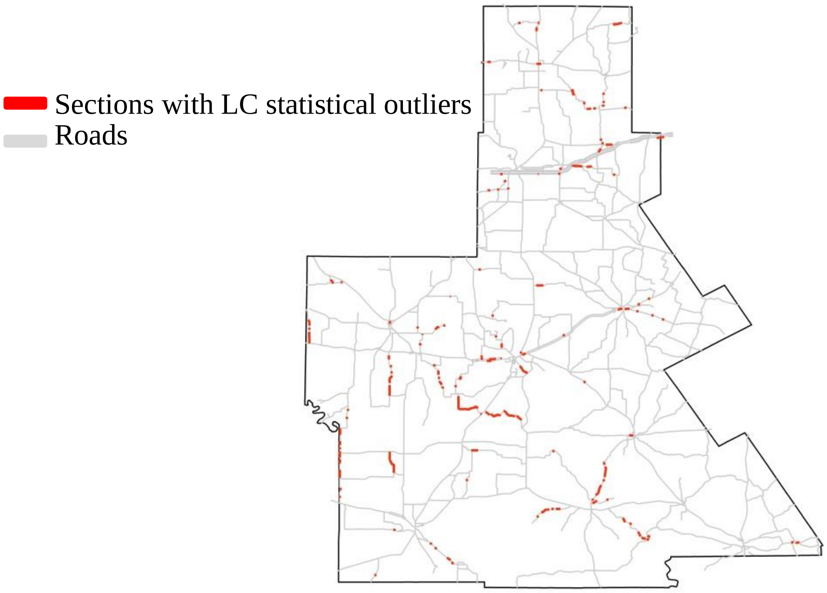


Figure 23 Results of Case 1 analysis: sections with erroneous LC values (Brownwood District)

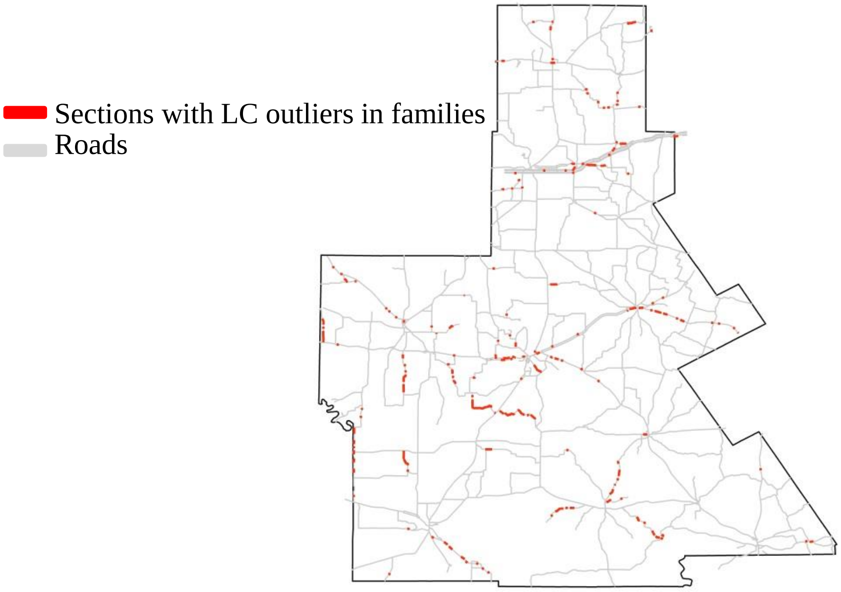


Figure 24 Results of Case 2 analysis: sections with erroneous LC values (Brownwood District)

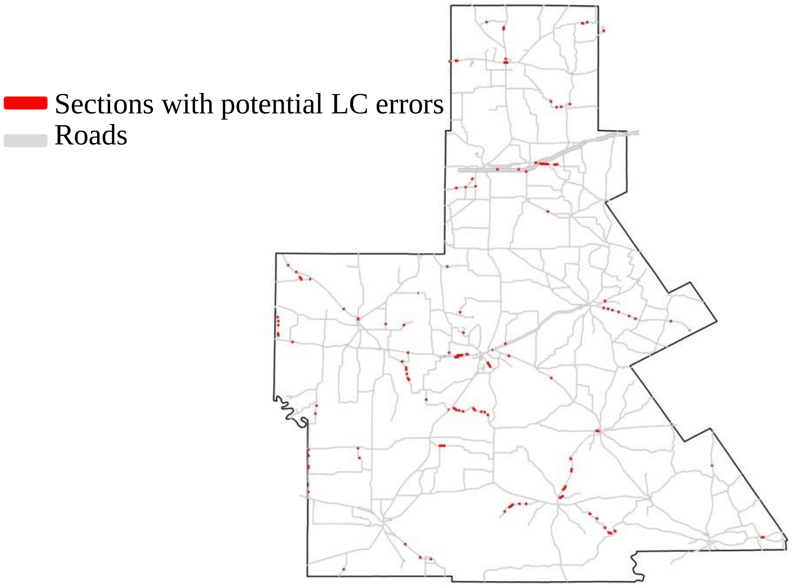


Figure 25 Results of Case 3 analysis: sections with erroneous LC values (Brownwood District)

Discussion of Results (Case 1 Analysis vs. Case 2 Analysis)

The causes of difference in results between Case 1 and Case 2 analyses are discussed through examples. Table 10 shows examples of detected potential errors by Case 1 analysis, Case 2 analysis, or both. Section US0084R05321.0 belongs to climate zone 3, pavement family A, and medium traffic loading. The change in LC value between 2013 and 2014 is 41 (i.e., Δx = 41). This Δx is outside the UL and LL for both cases (i.e., entire network and the section's performance family). Thus, 2014 LC value is detected as potential error by both Case 1 and Case 2 analyses.

Section SH0071K04461.0 belongs to climate zone 3, pavement family A, and medium traffic loading. The change in LC value between 2013 and 2014 is 41 (i.e., Δx

= 41). This Δx is outside the UL and LL for the performance family; however, it is within the LL and UL range for both cases entire network. Thus, 2014 LC value is detected as potential error by Case 2 analysis only.

Section FM1176K03440.5 belongs to climate zone 3, pavement family C, and low traffic loading. The change in LC value between 2013 and 2014 is -32 (i.e., Δx = -32). This Δx is outside the UL and LL for the entire network; however, it is within the LL and UL range for the 3-C-Low performance family. Thus, 2014 LC value is detected as potential error by Case 1 analysis only.

Table 10 Examples of sections detected as potential errors by Case 1 analysis and Case 2 analysis.

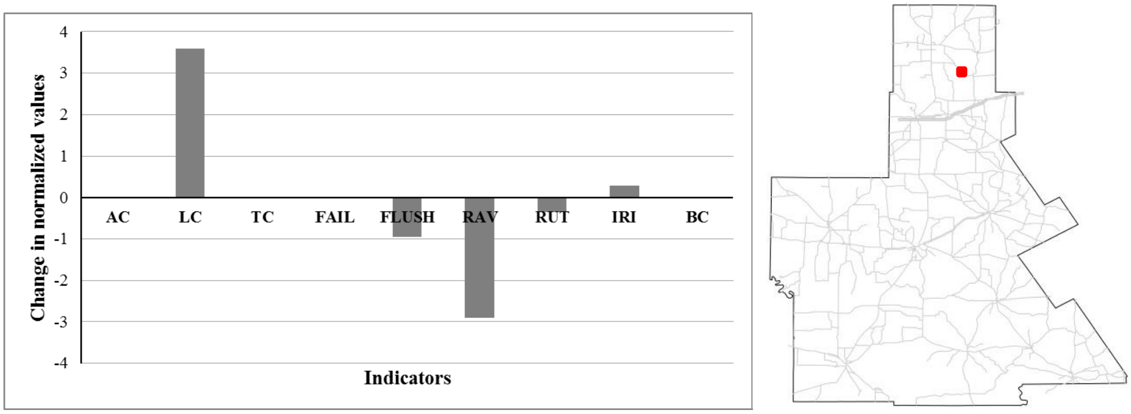
Section number	Change in LC value (Δx)	Case 1 Analysis		Case 2 Analysis			Potential Error?
		LL	UL	Performance family*	LL	UL	
US0084R05321.0	41	-31.3	30.7	3-A-Medium	-25.8	26.6	Both
SH0071K04461.0	30	-31.3	30.7	3-A-Medium	-25.8	26.6	Case 2
FM1176K03440.5	-32	-31.3	30.7	3-C-Low	-34.2	32.6	Case 1
FM0502K04501.0	-50	-31.3	30.7	3-C-Low	-34.2	32.6	Both
US0067K05440.5	-30	-31.3	30.7	3-A-Medium	-25.8	26.6	Case 2

* Climate Zone – Pavement Family – Traffic Loading

Discussion of Results (Case 1 Analysis vs. Case 3 Analysis)

Similar to the previous discussion of Case 1 analysis vs. Case 2 analysis, the causes of difference in results between Case 1 and Case 3 analyses are discussed through examples.

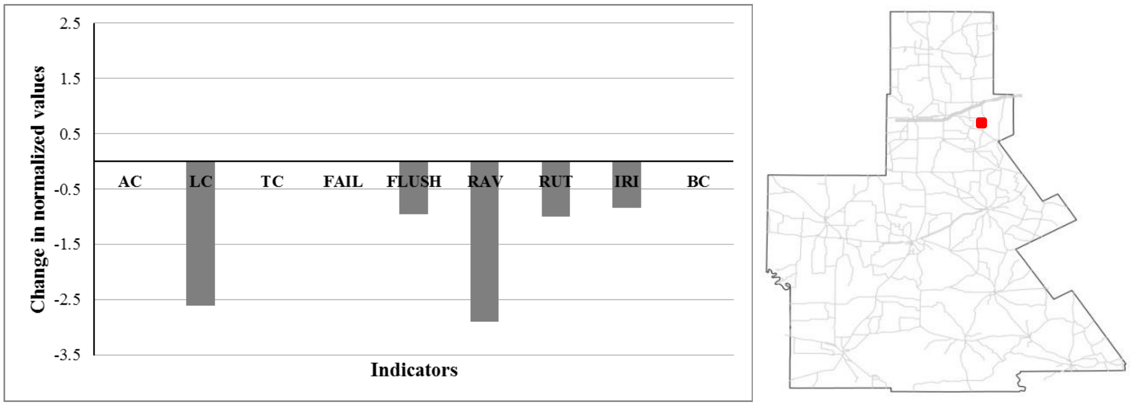
Figure 26 shows an actual pavement section along with changes in LC and other condition indicators after being normalized (condition changes from 2013 to 2014). The LC values for the section were detected as potential error by both Case 1 and Case 3 analyses. This is because Δx was outside the LL and UL range for the entire network (Case 1 analysis) and at the same time, this Δx is inconsistent with the Δx for the majority of the other condition indicators (i.e., violating the consistency checks). Thus, LC data was detected as potential error by both Case 1 and Case 2 analyses.



(a) (b)

Figure 26 An actual example pavement section with correctly detected outlier: (a) normalized Δx for multiple condition indicators in 2013 and 2014, and (b) Location of the pavement section

Figure 27 shows an actual pavement section that serves as an example of a false positive outlier by Case 1 analysis. This section was detected by Case 1 analysis as a statistical outlier based on its LC change. However, this Δx is consistent with the Δx for the majority of the other condition indicators for this particular pavement section (i.e., passing the consistency checks). All Δx values are negative; suggesting that the condition of this pavement section has improved likely due to applying a thick overlay (which results in reduction to all surface cracking as well as rutting and IRI).



(a) (b)

Figure 27 An actual example pavement section with false positive LC error identification by Case 1 analysis: (a) normalized Δx for multiple condition indicators in 2013 and 2014, and (b) Location of the pavement section

Figure 28 shows an actual pavement section that serves as an example of a false negative outlier by Case 1 analysis. This section was not detected by Case 1 analysis as a statistical outlier based on its LC change. However, this Δx is outside the LL and UL for its performance family and thus it warrants a consistency check. Figure 37 shows that this Δx is inconsistent with the Δx for the majority of the other condition indicators (i.e., violating the consistency checks). Thus, this is a likely error.

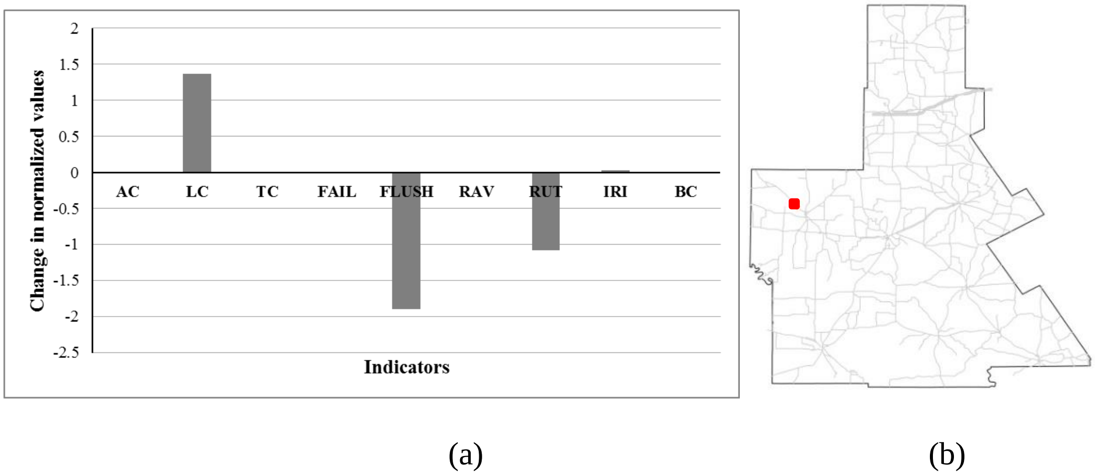


Figure 28 An actual example pavement section with false negative LC error identification by Case 1 analysis: (a) normalized Δx for multiple condition indicators in 2013 and 2014, and (b) Location of the pavement section

Figure 29 shows false positive and false negative in LC error identification for Brownwood district. Case 3 analysis detects more false positives compare with false negatives.

Table 11 shows the differences between pavement sections detected by Case 1 analysis and Case 3 analysis for all performance indicators in Brownwood District. This table shows that the developed technique has the advantage of recognizing false positives in traditional statistical methods because of its ability to consider consistency among multiple condition indicators. Also, the developed technique has the advantage in reducing false negatives in traditional statistical techniques because of its ability to consider performance families. It can be seen that the developed technique has greater power in reducing false positives compared to reducing false negatives.

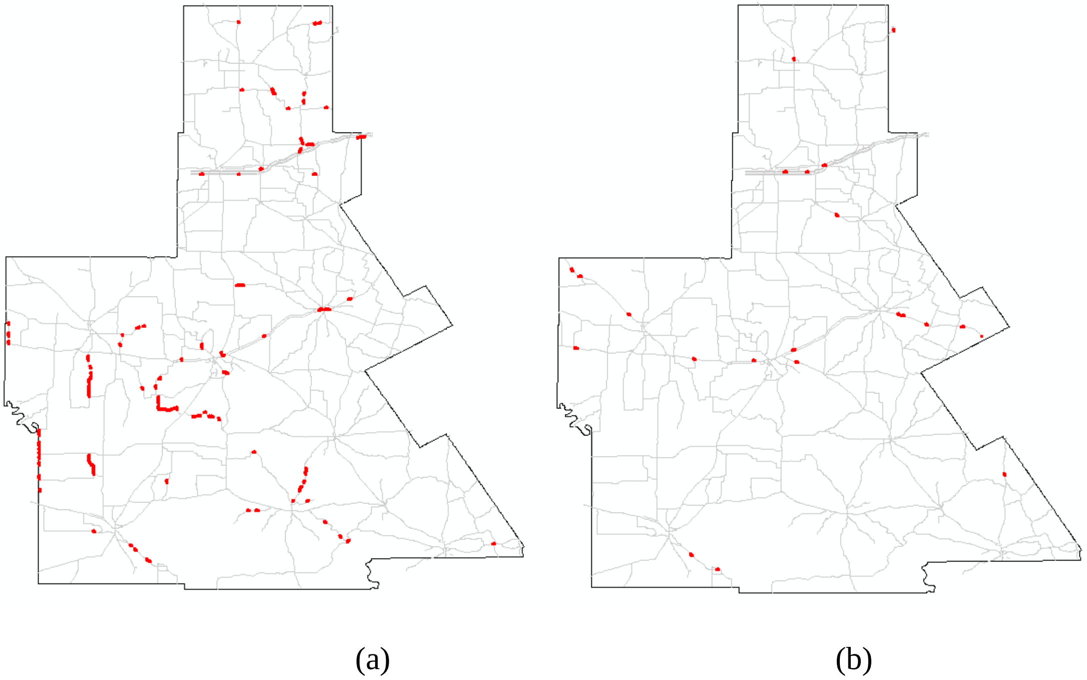


Figure 29 False positive and false negative in LC error identification (a) false positive (b) false negative

Table 11 Number of pavement sections from the Brownwood District detected by Case 1 and Case 3 analyses.

Condition Indicator	No. of Section with Potential Errors (Case 3 Analysis)	No. of Section with Potential Errors (Case 1 Analysis)	No. of Section detected by both cases	False negative	False positive
LC	135	245	113	22	132
AC	87	176	76	11	100
TC	99	291	99	0	192
RUT	184	287	165	19	122
IRI	131	256	115	16	141

False negative: Sections were detected by Case 3 to have potential error values but Case 1 failed to detect them

False Positive: Sections were detected by Case 3 to have extreme yet valid data but Case 1 falsely detected them as errors

6. IMPACT OF ACCURACY IN PAVEMENT CONDITION DATA ON THE ASSESSMENT OF NETWORK HEALTH

This part of the research addresses the second research question; which is: How does accuracy of pavement condition data impact the predictions of future performance of the road network? The impact was investigated in both project and network level. In project level investigation, the effect of potential erroneous indicators was investigated on the health measurement of sections detected to have these potential errors. To answer the question in network level, the road network health is measured based on two scenarios: original database and clean database. The original database includes all original data (without any modifications). The clean database does not include data points that are identified by the developed method as potential errors. Remaining Service Life (RSL) is used as an overall measure of network health to compare these two scenarios. RSL represents the timeframe within which the pavement is expected to require an M&R treatment. RSL indicates not only the current condition of roadway network but also the needed service and budget in future, thus it is of interest to decision makers.

In this chapter, first the RSL estimation process is described. Then, the presented technique is used to identify potential errors. Next, impact of accuracy on RSL of the sections detected to have potential errors was investigated. Finally original database and clean database are compared based on the estimated RSL of the network (i.e. the average RSL of the network and the distribution of RSL for the network).

RSL Estimation Process

RSL is defined as the estimated number of years from the last data collection year to the time when the pavement condition indicator (e.g. distress index) reaches a threshold value of minimum acceptable service level (Baladi and Novak 1992, Baladi et al. 2011). Figure 30 shows the estimation of RSL using measured and predicted distress index over time. Elkins et al. (2013) suggested that predicting the remaining life of pavement is an essential capability for PMSs.

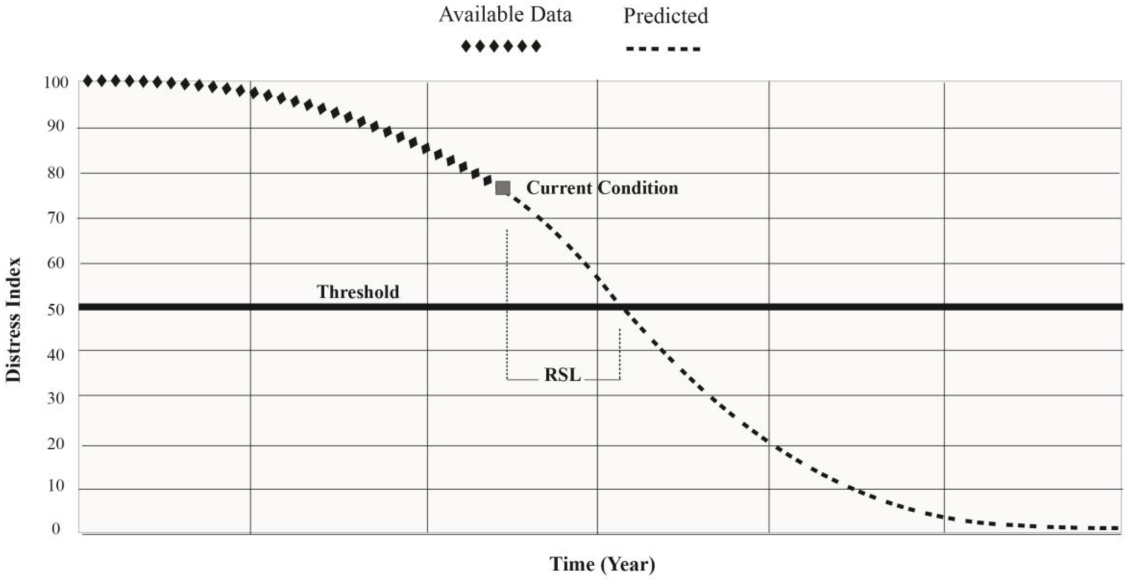


Figure 30 Estimation of RSL using measured and predicted distress index over time (adopted from Baladi et al. 2011)

Since, pavement condition is measured using several indicators, RSL is calculated for each indicator individually. The minimum RSL among the multiple condition indicators is the pavement section’s overall RSL.

RSL of a given pavement network is computed as the weighted average of all network pavement sections, as below:

$$RSL_{(network)} = \frac{\sum_{i=1}^{n} RSL_i * SL_i}{\sum_{i=1}^{n} SL_i} \qquad \text{Eq 3}$$

where RSL$_i$ is the RSL of pavement section i, and SL$_i$ is the length of section.

Figure 31 shows a step by step method implemented in this study to compute RSL for any given pavement section. Pavement performance indicators are used to measure pavement condition and specify whether the pavement section needs to receive an M&R treatment. In the RSL estimation process, considered indicators are required to have current values and reliable deterioration models to predict the pavement future condition. Naturally, more accurate current condition data and deterioration model results in more accurate estimation of RSL.

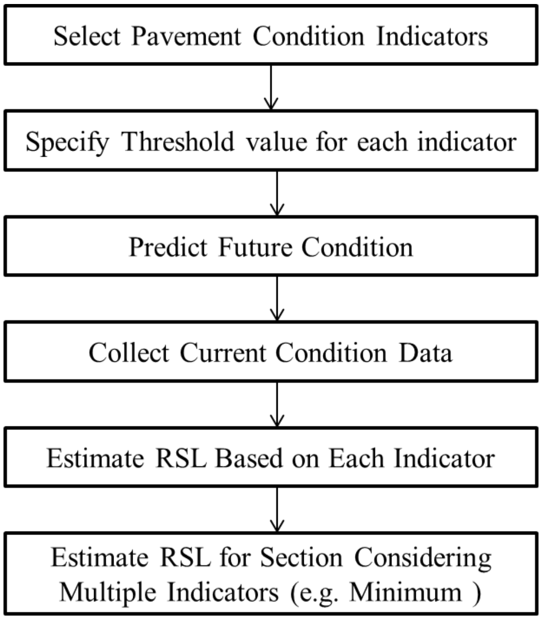

Select Pavement Condition Indicators
Specify Threshold value for each indicator
Predict Future Condition
Collect Current Condition Data
Estimate RSL Based on Each Indicator
Estimate RSL for Section Considering Multiple Indicators (e.g. Minimum)

Figure 31 RSL estimation process

Threshold values are defined as the amount of distress that shows the pavement section needs corrective treatment because preventive maintenance can no longer effectively improve pavement condition. In this study, these values are determined based on needs estimate trigger criteria defined by TxDOT to receive at least Light Rehabilitation (RL) treatment (Gharaibeh et al. 2012). Threshold values are different based on Average Daily Traffic (ADT) levels of pavement sections. Table 12 shows selected indicators and their threshold values for different ADT levels in the RSL estimation process. There indicators are defined as follows (TxDOT 2015):

- Alligator cracking (AC): Percentage of the rated lane's total wheelpath area that is covered by alligator cracking
- Longitudinal Cracking (LC): Linear feet per station (i.e. average feet of cracking in each 100 feet of surface)
- Transverse Cracking (TC): number of transverse cracks per 100-ft station.
- Shallow Rutting (ShRUT): Percent of the section's total wheel path area that is rutted between 0.25 in and 0.49 in.
- Deep Rutting (DRUT): Percent of the section's total wheel path area that is rutted between 0.5 in and 0.99 in.
- FAIL: Total number of failures observed along the entire rated section.
- Block Cracking (BC): percentage of the rated lane's total surface area with block cracking.
- Patching: Percentage of the rated lane's total surface area with patching.

Table 12 Performance indicators and threshold values (adopted from Gharaibeh et al. 2012).

Condition Indicator	Threshold Value to Receive LR			
	ADT ≤ 99	99 < ADT ≤ 999	999 < ADT ≤ 4,999	4,999 < ADT
AC (%)	25	20	15	10
LC (ft/100ft)	126	101	101	101
TC (No./100ft)	7	7	7	5
ShRUT (%)	12	12	10	10
DRUT (%)	9	9	9	9
FAIL (No.)	3	2	2	2
BC (%)	16	16	16	12
Patch (%)	42	32	22	12
DS	70	70	70	70

* A Distress Score (DS_ below 70 indicates that the pavement section is in Poor or Very Poor Condition

TxDOT performance prediction models for different distress types are used for estimating RSL. These models are s-shaped (Figure 32) and have the following general equation:

$$L_i = \alpha e^{-(\frac{A}{Age})^{\beta}} \qquad \text{Eq 4}$$

where:

L_i: represents the density of distress i (i.e., distress quantity normalized for section length or percent ride quality lost).

Age: number of years since last construction on the pavement section

α: maximum loss factor which controls the maximum Li

β: slope factor which controls how steeply Li increases in the middle of the curve

A: prolongation factor controls the location of the Li curve’s inflection point.

For Equation 4, separate values of α, β, and A were developed in previous research for all combinations of pavement type, distress type, subgrade type, climate and traffic level (Gharaibeh et al. 2012). These coefficients are presented in Appendix A.

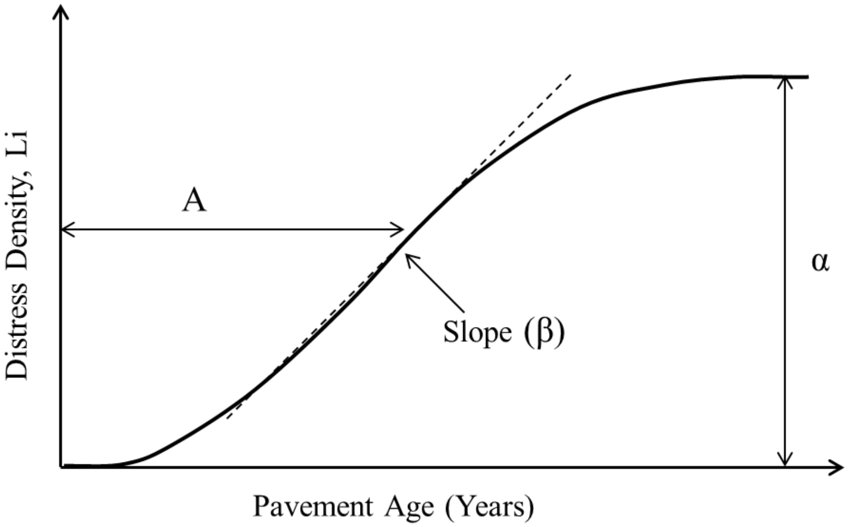


Figure 32 Typical S-shaped pavement performance prediction models used by TxDOT (Gharaibeh et al. 2012)

The change of DS versus age follows a sigmoidal curve too, and estimated using the following equation (Gharaibeh et al. 2014):

$$DS = DS_0\left[1 - e^{-(\frac{\rho}{Age})^{\beta}}\right] \qquad \text{Eq 5}$$

where DS_0 is the DS immediately after M&R treatment or construction; Age is the number of years after the last treatment on the pavement section; β and ρ are the slope factor and prolongation factor, respectively. These coefficients change for different combination of climate zone, pavement family, loading traffic, and the type of last treatment. For example, Figure 33 shows the pavement DS prediction model for a pavement section in climate zone one, pavement family B, and medium loading traffic which has received a LR as its last treatment.

DS ranges from 0 to 100; where DS value equal to 100 represents a recently constructed pavement section in a great condition. Pavement sections with DS values less than threshold value (e.g. 70) are in poor condition and need to receive M&R treatment.

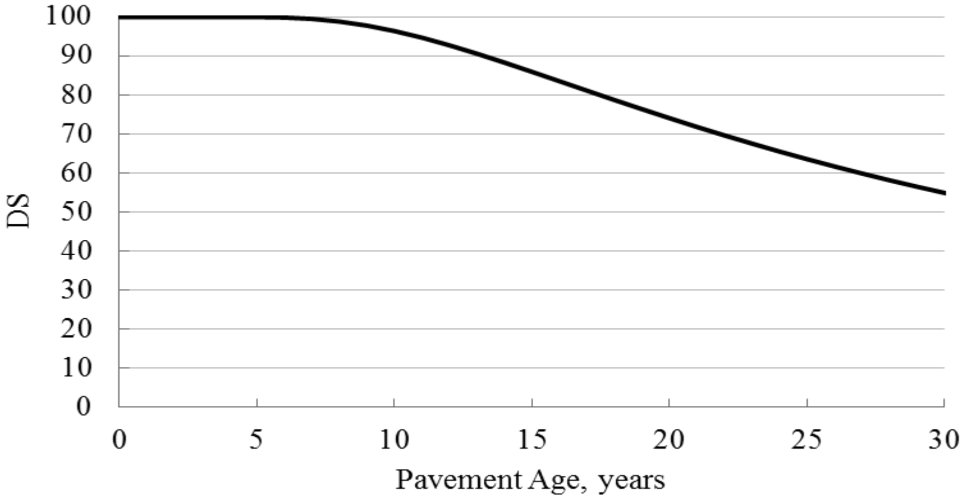


Figure 33 Example DS prediction model (pavement B, Zone 1, medium traffic loading, and LR treatment).

RSL for each performance indicator is the difference between the pavement age based on the condition indicator value in current year and the pavement age when the indicator reaches its threshold value. Thus, any given pavement section has several estimated RSL (i.e. one for each condition indicator). The minimum RSL among the multiple indicators is considered as the pavement section overall RSL. Figure 34 shows how several condition indicators of a pavement section might reach their thresholds in different periods of time (age). This example pavement section is located in climate zone one, pavement family B, and medium loading traffic road and has recently received MR (thus all condition indicators have zero value and DS is equal to 100). Therefore, estimated RSL of each indicator is equal to the period of time for that indicator to reach its threshold limit. The estimated RSL for this example pavement section is controlled by alligator cracking (AC) and is estimated to be 16 years. If the current value of anyone of the condition indicators changes, the section's RSL would be different. For example if the shallow rutting in current year was equal to 5 percent, then the current age of shallow rutting would be equal to 14 and the pavement section RSL would change to 6 years (i.e. 20 year threshold age minus 14 year current age of shallow rutting).

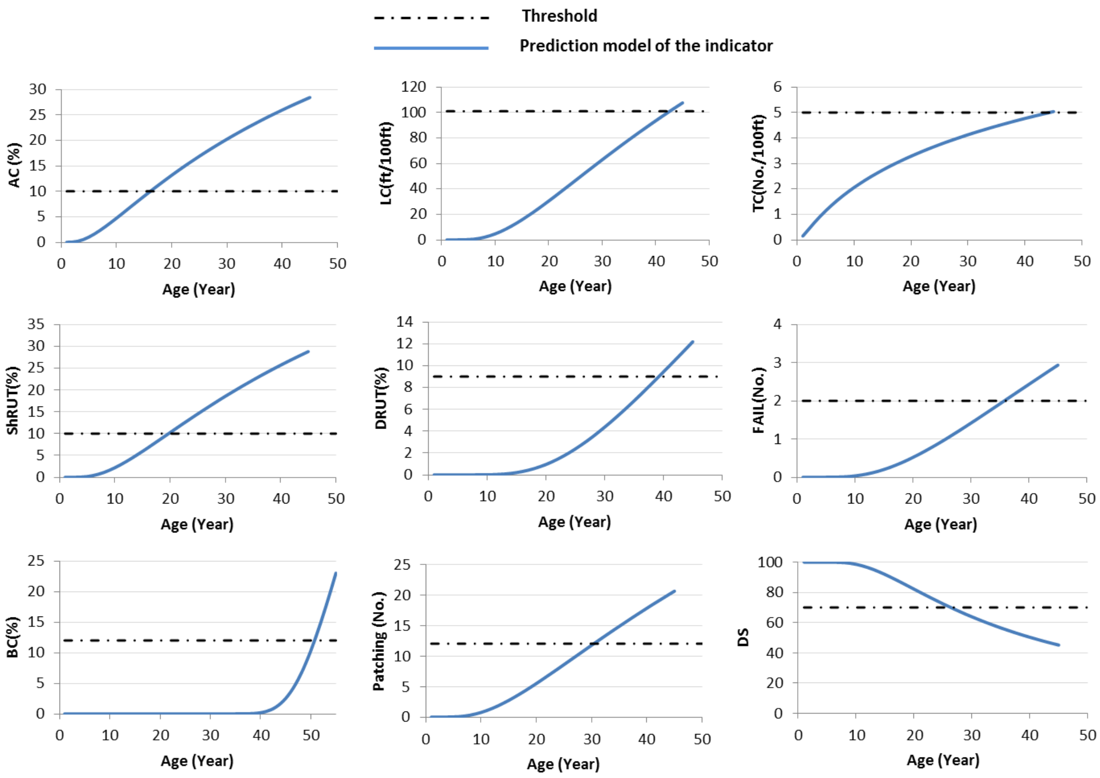


Figure 34 Pavement age at which several indicators reach their threshold values for an example pavement section (pavement B, Zone 1, medium traffic loading, MR treatment, AADT >4999)

Erroneous data might wrongly increase or decrease current age by denoting an invalid current condition values. In next section, the impact of erroneous data instances on the estimated RSL of pavement sections is investigated.

Impact of Condition Data Accuracy on RSL for Sections Affected by Bad Condition Data

Detecting a potentially invalid performance indicator value of a pavement section does not necessarily mean that other indicator values of the section are erroneous too.

However, even only one invalid data value might change the estimated RSL of the section, and thus deceive the PMS decision making process. In order to assess the impact of erroneous condition data on the estimated RSL of pavement sections with potential errors, the following scenarios were investigated:

- First, RSL of the section was estimated considering all data, including potentially erroneous values (Original RSL).
- Second, a condition indicator with erroneous values was removed from the RSL process, and the RSL of the section was estimated again (Permuted RSL).

To prevent effects of other indicators on the result of removing suspicious data, the process was conducted on the sections that were detected to have only one potential error value (e.g., AC or LC). For example, if a given pavement section was identified to have likely error data for both AC and TC values it was not considered in individual checking, however it has been considered in checking all indicator effects.

Since all distress types studied here affect the computed DS, errors in any condition indicator affects the DS value. Thus, in this analysis, RSL is computed without considering DS.

Table 13 shows the results of removing suspicious indicators (one at the time) from the RSL estimation process. Figure 35 compares the distribution of permuted RSL and original RSL for the sections affected by bad condition data.

Table 13 Effect of data errors in individual condition indicators on estimated RSL.

Condition Indicator	Sections with Likely Error Data (No.)	Average RSL (Years)	
		Original Data	Permuted Data
AC	2,094	3.5	7.7
TC	1,287	7.1	9.2
LC	3,615	5.8	10.1
RUT	3,761	1.7	16.7
All	12,127	4.2	11.9

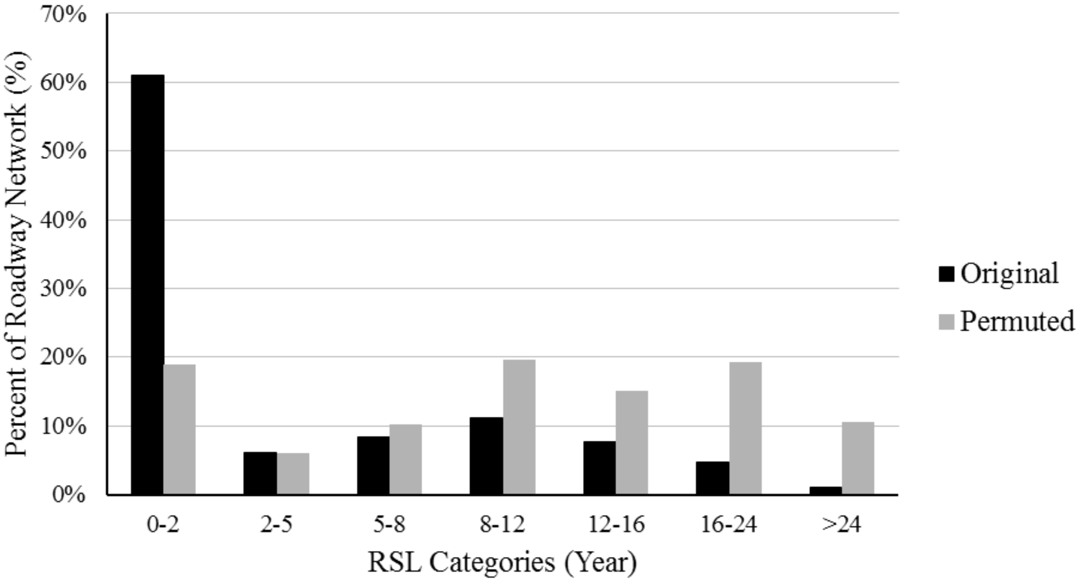


Figure 35 Distribution of original RSL vs. permuted RSL for sections affected by bad condition data

Impact of Condition Data Accuracy on RSL for Entire Network

In this section, the influence of pavement condition data accuracy on the predictions of future performance is investigated using all 165,469 PMIS pavement sections. The following procedure was implemented:

- First: RSL of all pavement sections were estimated using 2014 pavement condition data, as original data RSL.
- Second: Potential errors in condition data (i.e., PMIS2014 and PMIS2013) were detected using the proposed method in Section 3
- Third: Sections with potential errors were removed from the original data and the reminder was considered as clean data. RSL of clean data was compared to the RSL of original data to evaluate the impact of condition data accuracy.

The RSL comparison of these scenarios (i.e. origin data and clean data) is expressed in terms of the weighted average RSL (weighted by pavement section length) and the distribution of RSL for network. Weighted average RSL represents the overall health of the network. Assessing the distribution of RSL enables decision makers to focus on different parts of the network (e.g., parts needing M&R work in the near future vs. parts needing M&R work in the long-term).

Table 14 shows the percent of roadway network in different RSL categories for both original and clean data. Figure 36 compares the distribution of RSL of Texas roadway network using these categories (i.e. clean condition data vs. original condition data).

Table 14 Number of sections and percent of Texas roadway network in each RSL categories based on clean data and original condition data

RSL Category (Year)	Number of Sections		Percent of Network	
	Original Data	Clean Data	Original Data	Clean Data
0_2	46,594	37,939	28.2%	25.3%
2_5	21,723	19,985	13.1%	13.3%
5_8	19,812	18,416	12.0%	12.3%
8_12	32,457	30,644	19.6%	20.4%
12_16	21,368	20,291	12.9%	13.5%
16_24	17,680	16,939	10.7%	11.3%
>24	5,835	5,649	3.5%	3.8%

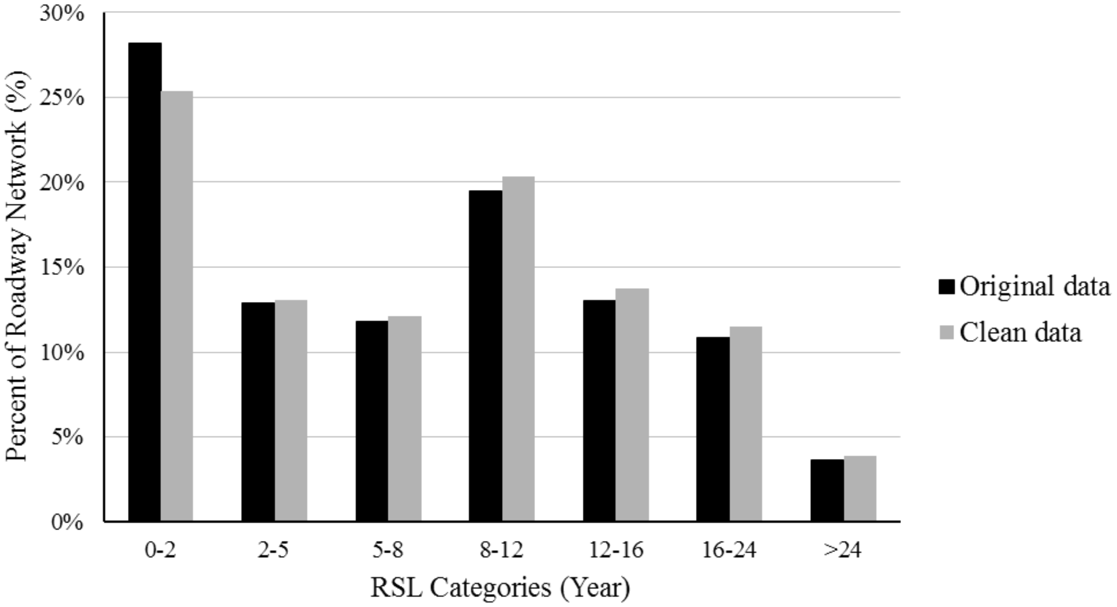


Figure 36 Distribution of RSL of the Texas roadway network using original data vs. clean data

The largest difference between the two scenarios occurs in the first category, where RSL is less than 2 years. Based on the original data, 28.2 percent of the network needs rehabilitation in the next two years; however, based on the clean data 25.4 percent of the roadway network is expected to need rehabilitation in the next two years. Thus, according to the clean data, the M&R budget in the next two year is 3,276 million dollar (i.e. assuming all pavement sections receive LR); which is 361 million dollar (i.e., 11% difference) less than the required budget estimated using the original data for the same period of time. Both clean and original data determine almost the same percent of the roadway network in the range of 2 to 8 year RSL. For the categories with RSL greater than eight years, the percent of network in these categories is greater for the clean data than for the original data. This outcome may be attributed to the fact that the developed error detection method identifies fewer potential errors among sections in good condition.

This analysis reveals that errors in pavement condition data results in the overestimation of required budget for short term plans (i.e. 1-2 year M&R planning). There is an 11 percent difference in estimated budget for the next two years between the two scenarios (i.e. original data and clean data). This difference decreases when the two scenarios are compared based on longer planning period, since clean data shows that a greater portion of the network needs M&R after 2 years (Figure 26).

The weighted average RSL (weighted by section length) of all network is calculated using Equation 3 for both original and clean data. The average RSL of the network increased after removing sections with potential errors in condition data. Table

15 shows statistical parameters of estimated RSL for both scenarios. The original PMIS data underestimates the overall condition of the roadway network compared to the clean data (i.e. after removing about 10 percent of pavement sections with potentially erroneous data). These results suggest that, in this particular database, errors mostly exaggerate the amount of deterioration. The condition of sections with potentially erroneous data is worse than the average condition of total roadway network. This can be explained by the fact that inspecting and rating pavement sections with few or no distress is easier than inspecting sections that have several types of distress with different levels of severity.

Table 15 Statistical parameters of estimated RSL for original data vs. clean data

Population	Mean*	Std	Min	1st Quartile	Median	3rd Quartile	Max	Number of Sections
Original Data	8.00	7.3	0	1.2	7.4	12.2	30.3	165,469
Clean Data	8.39	7.3	0	1.9	7.7	12.9	30.3	149,863

* Weighted average RSL of all network

Based on the original data, on average, the roadway network needs at least a LR treatment in 8 years, however without considering the erroneous data roadway network needs to receive the same treatment in 8.39 years. In order to show the economic effect of these results, the Present Worth Value (PWV) and the Equivalent Uniform Annual Cost (EUAC) of both scenarios are calculated by discounting future treatment costs to the current year as shown in Equation 6 and Equation 7.

$$PWV = Cost_{init} + \sum_{k=1}^{N} Cost_{Future} \left[\frac{1}{(1+i)^{n_k}}\right] \qquad \text{Eq 6}$$

$$EUAC = PWV * i * \left[\frac{(1+i)^m}{(1+i)^m - 1}\right] \qquad \text{Eq 7}$$

where:

$Cost_{init}$: represents the cost of initial treatment.

i: represents discount rate.

k: is the number of future treatment applied during the analysis period (k= 1 to N).

n_k: shows the year at which the kth treatment is applied.

m: is the number of years in analysis period.

The magnitude of difference between the PWVs for these scenarios is related to three factors: the treatment cost, the year of treatment application, and the discount rate. The average unit cost of LR, treatment is estimated to $76,086 per lane mile (Gharaibeh et al. 2014). This unit cost is used in estimating the PWV for both scenarios. The difference between the PWV for the two scenarios comes from the difference in the timeframe (i.e., number of years into the future) for applying the treatment. PWV has a negative relationship with the timeframe; the longer the timeframe is, the lower the PWV would be. Thus, PWV is lower for the clean data scenario than for the original data scenario. The discount rate represents the rate of change in the value of money over time. Several studies have indicated the significant influence of the discount rate on Life Cycle Cost Analysis (LCCA) results. Thus, the discount rate should be as realistic as possible (Hall et al. 2003; Ferreira and Santos 2012). The U.S. Office of Management and Budget (OMB) annually update nominal and real discount rates. Real discount rates represent the true time value of money with removing the inflation premium; however, the nominal discount rate includes an inflation component. The real discount rate is commonly used in pavement management LCCA (Hall et al. 2003; Wall and Smith 1998). In 2015, the OMB has recommended the real interest rate value equal to 0.9 percent for a 10-year analysis period (USOMB 2015). Beside the OMB report, historical trends and previous studies should be considered in determining discount rates. Figure 37 shows real discount rates for various analysis periods published by OMB in last 20 years.

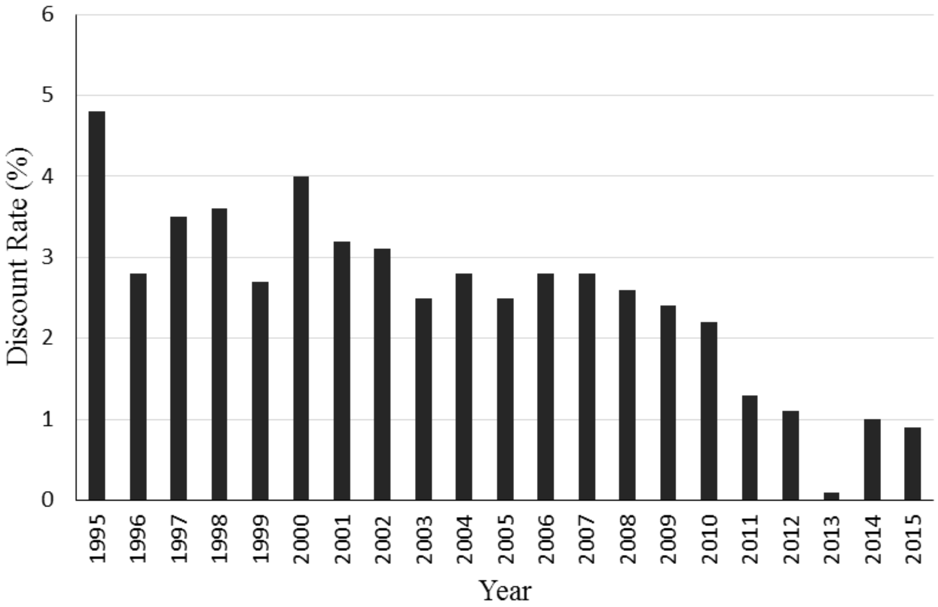


Figure 37 Trends of the real discount rates for 10-year analysis period in last 20 years

To account for the uncertainty in discount rate value in this study, the difference between the two data scenarios is calculated using several discount rates in the range of

0 percent to 5 percent. Figure 38 shows the difference between PWV of required treatment for clean data and original data versus discount rate. The amount of difference varies between zero and 170 million dollar based on the discount rate. The value most often used in federal LCCA is 4 percent (Kim et al. 2014), making the difference between the two PWVs equal to 145 million dollar ($9.43 Billion for original dataset vs. $9.29 billion for clean dataset, or about 1.5 percent difference).

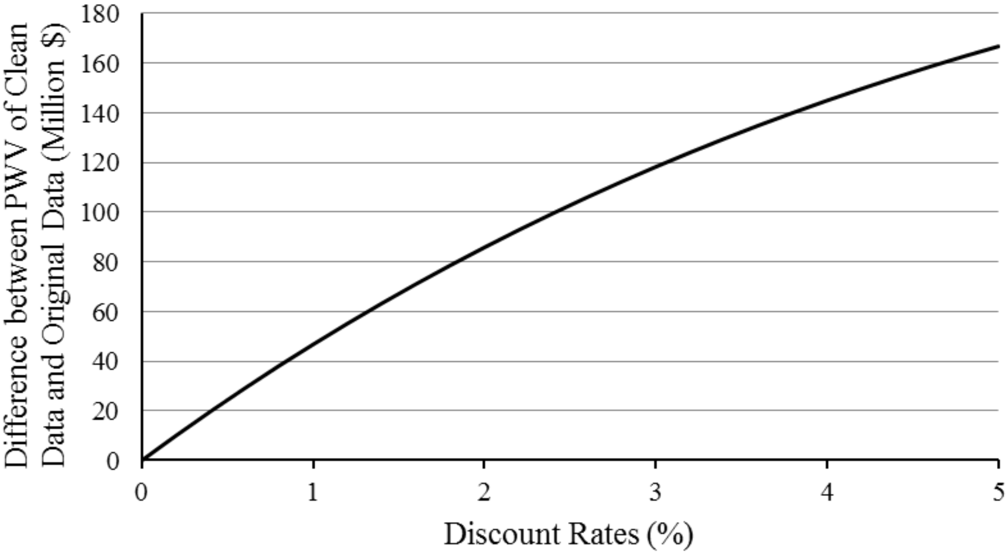


Figure 38 Difference between PWV of required treatment for clean data and original data vs. discount rate

Figure 39 shows the difference between the EUAC of required treatment for the clean data and original data, at varying discount rate. For a 4 percent discount rate, the difference between EUAC for clean data and original data is equal to 21 million dollar.

In other words, in next 8 years the yearly needed budget would be overestimated by 21 million dollars due to errors in the pavement condition data. Although this is only 1.5 percent of total estimated yearly budget, 21 million dollars remains a significant amount of money. These results may vary for other pavement networks.

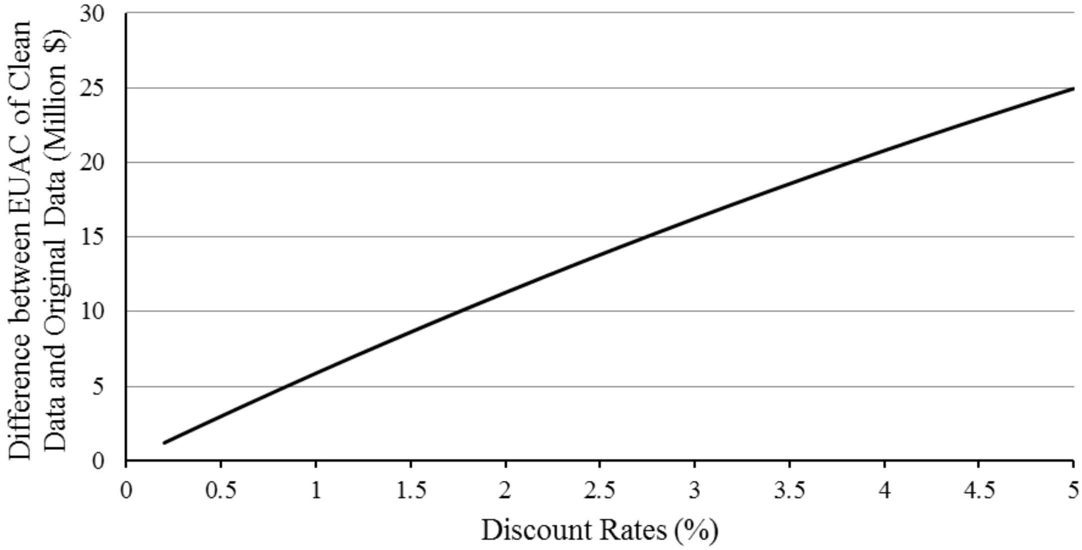


Figure 39 Difference between EUAC of required treatment for clean data and original data vs. discount rates

7. METRICS FOR MEASURING PAVEMENT CONDITION DATA QUALITY

Transportation agencies own large amounts of pavement condition data that feed into PMSs. Assuring the overall quality of these datasets is critical to the reliability of the outputs of PMSs. The quality of a pavement condition dataset has several dimensions, such as accuracy (closeness between a data value and the real-world value that it represents), completeness (absence of missing values in the dataset), consistency (captures the violation of semantic rules defined over a set of data items), and timeliness (how up-to-date the data is with respect to the task at hand). This chapter provides metrics for measuring these quality dimensions for pavement condition datasets. Ultimately, these metrics can be used by transportation agencies to determine if pavement condition datasets are of acceptable quality. As a case study, the metrics are applied to a pavement condition dataset for TxDOT's Bryan district roadway network in east-central Texas. This dataset consists of approximately 7,000 records and 70 columns. Each record represents a pavement section (approximately 0.5-mile in length). The columns include pavement inventory, individual distress types, and pavement condition indexes. It is hoped that these metrics will enable pavement engineers to better assess the overall quality of pavement condition datasets.

Methodology

In this research, data quality is defined as a set of characteristics (e.g., accuracy, completeness, timeliness, and consistency) that a dataset is supposed to possess in order for it to be trusted to serve its purpose. The stringency of these characteristics can be used to measure the overall level of data quality (Dasu and Johnson 2003). These features, also known as data quality dimensions, make a dataset appropriate for a specific use. Therefore, the importance of specific data quality dimensions varies, depending upon the database and the purpose of the data; a database might be of adequate quality for one purpose, but not for another. These dimensions can be defined using theoretical, empirical, or intuitive approaches, as is discussed in Chapter 2. This work defines six major quality dimensions for network-level pavement condition data, as shown in Figure 40. Each dimension is explained below, as well as the metrics that were determined to quantify the quality of the pavement condition data with respect to the particular dimensions required for this research. Other dimensions could be added if transportation agencies or other users found them to be important.

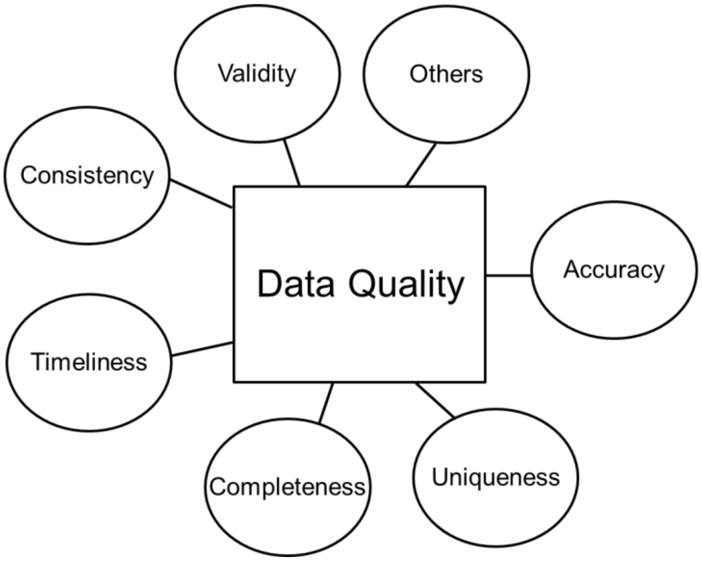


Figure 40 Pavement condition data quality dimensions

Timeliness

Timeliness represents the most recent time the pavement condition data was updated. It can also demonstrate the frequency of change in the particular database. Timeliness is one of the easiest data quality dimensions to measure. The unit of timeliness can be either months or years. Pavement continuously deteriorates because of factors such as traffic loading, aging, and environmental effects. Thus, it is important that decision makers use data representing the most recent condition of the particular roadway network. Old data might not be useful for making pavement management decisions; even if the data were both accurate and complete at the time the decision was made. Typically, the frequency of the pavement condition data collection depends upon the state's DOT policy; most DOTs collect data either annually or biennially.

Uniqueness

Uniqueness means that each data record should be distinctive and there should be no duplicates. In pavement condition data, there may be two records providing information about the same pavement section. This can affect PMS decisions and result in either an over- or under-estimation of needs. Thus, one of the first steps in reviewing the quality of pavement condition data is detecting and removing duplicated sections. Following equation calculates uniqueness of network-level pavement condition data.

$$Uniqueness\ (\%) = (1 - \frac{\sum_{i \in D}(L_i * ESAL_i)}{\sum_{i \in I}(L_i * ESAL_i)}) * 100 \qquad \text{Eq 8}$$

where, L_i is the length of pavement section i; $ESAL_i$ is the current 18 kip equivalent single axel load value of the section; D is the set of pavement sections that are duplicated in the dataset; and I is the entire set of pavement sections in the dataset.

Completeness

Completeness represents the extent to which the desired attribute data has been provided. It shows the percentage of the real world that is covered in the database. In most databases, completeness can be very difficult to measure because the actual inventory is not specifically known. In pavement condition data, completeness demonstrates the number of pavement sections for which surface conditions have been collected and recorded in the database, as compared to the total roadway network inventory. It should reflect both pavement sections missing from the database and pavement sections in the database that have blank (or missing) values for desired attributes. Completeness of pavement condition data can be calculated using following equation:

$$Completeness\ (\%) = \frac{\sum_{i \in A}(L_i * ESAL_i)}{\sum_{i \in I}(L_i * ESAL_i)} * 100 \qquad \text{Eq 9}$$

where L_i is the length of pavement section i; $ESAL_i$ is the current 18 kip equivalent single axel load value of the section; A is the set of pavement sections in the dataset with available condition data; and I is the entire set of pavement sections in the dataset.

Validity

Validity is defined as the degree to which the data values pass the necessary validity criteria (within a set of predefined accepted values). For example, Alligator

Cracking (AC) ratings represent the percentage of the rated lane's total wheel path area that is covered by alligator cracking. Thus, AC values must range from 0 to 100 percent; any number outside of this range is invalid. Validity criteria (or validation checks) are usually based upon expert opinions or logical facts and can be defined as a simple rule (e.g., not exceeding a specific threshold) or several complex rules.

$$Validity\ (\%) = \frac{\sum_{i \in V}(L_i * ESAL_i)}{\sum_{i \in A}(L_i * ESAL_i)} * 100 \qquad \text{Eq 10}$$

where L_i is the length of pavement section i; $ESAL_i$ is the current 18 kip equivalent single axel load value of the section; V is the set of pavement sections in the inventory with valid data; and A is the set of pavement sections in the dataset with available condition data.

Figure 41 compares completeness with validity. This chart represents the length of the entire roadway network that is expected to be collected (Li). The completeness demonstrates the lengths of the sections with valid and invalid values, divided by the length of the entire roadway network. The validity represents the length of the sections with valid values divided by the length of the sections with available data (i.e., both valid and invalid data).

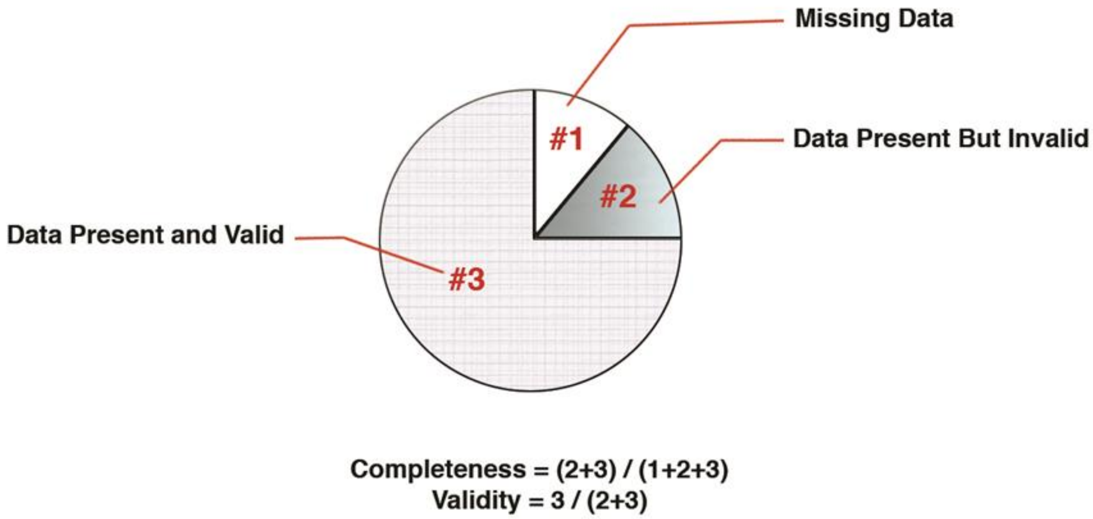


Figure 41 Illustration of completeness and validity measure (Wang et al. 1995)

Consistency

Consistency can consider several aspects. It can refer to the provision of the same data for the same object, even if these data are collected at another location or time. In pavement condition data it refers to the reliability of performance indicator definitions and data collection methods. For example, the equipment calibration method should produce the same results, even when data collection occurs at different times and locations. In order to have consistent data, some agencies inspect the control site each time data collection is initiated in a new district, or each time the data collection vehicle leaves the state. Inspection results are more consistent if the collectors are well-trained raters using identical distress rating guides (Pierce et al. 2013).

Another aspect of consistency that this study focuses on is the degree to which a dataset agrees with itself. In this sense, the values in one dataset must align with the values in another dataset. The necessary alignment can be in a variety of directions, depending on the context (Loshin 2006), including between one set of attribute values and another attribute set within the same record (record-level consistency), between one set of attribute values and another attribute set in different records (cross-record consistency), and between one set of attribute values and the same attribute set within the same record at different points in time (temporal consistency). Consistency may also take into account the concept of reasonableness, in which some range of acceptability is imposed upon the values of a set of attributes.

In a consistent database, information comes from several attributes of a data instance that must not be in conflict with one another. Based on the attributes' dependencies, certain rules and constraints can be defined to capture violations of consistency. For example, pavement conditions cannot improve if the section in question receives no treatment. In other words, data consistency checks are performed to look for data that does not make sense. A type of logic test is usually implemented to check for unexpected values. Although the conflict might be explained in some data instances, the more conflicts that are found, the higher the chance of errors. Values for different attributes must satisfy both logical and statistical constraints. Increasing the number of constraints also increases the number of data instances (i.e., pavement sections) and thus escalates the likelihood of suspicious data values. The consistency of an entire database is calculated using the following equation:

$$Consistency\ (\%) = \frac{\sum_{i \in S}(L_i * ESAL_i)}{\sum_{i \in A}(L_i * ESAL_i)} * 100 \qquad \text{Eq 11}$$

where L_i is the length of pavement section i; $ESAL_i$ is the current 18 kip equivalent single axel load value of the section; S is the set of pavement sections in the inventory with available data that satisfy the required consistency rules; and A is the set of pavement sections in the dataset with available condition data.

For demonstration purposes, consistency of the Bryan dataset was measured for IRI data only by comparing the IRI for the right wheel path and left wheel path.

Consistency of IRI Data

TxDOT collects IRI data for every 0.1 mile at both wheel paths. An average IRI value is used to characterize surface roughness. The lower the calculated IRI, the smoother the pavement will ride (and vice versa); higher IRI values represent rougher pavement. The calculated IRI values at the two wheel paths should be close to one another; otherwise, the average will overestimate the pavement condition of one wheel path and underestimate it in the other (Jai et al. 2016). Thus, the difference between the calculated IRI values along the two wheel paths demonstrate the variability of the data collected. Although the difference in the IRI of the two wheel paths might be true (e.g., due to true differences in the actual pavement condition of the two wheel paths), significant differences usually represent inconsistencies in the IRI data that should raise suspicion. Figure 42 shows the correlation between the IRI values of the two wheel paths in a set of Texas pavement condition data collected in 2014. In most sections, the calculated IRI values from the left wheel path (IRI_LT) are similar to the calculated IRI values collected from the right wheel path (IRI_RT). This fact can be used to determine a general consistency rule for IRI data that should be satisfied if the overall data consistency is to be verified.

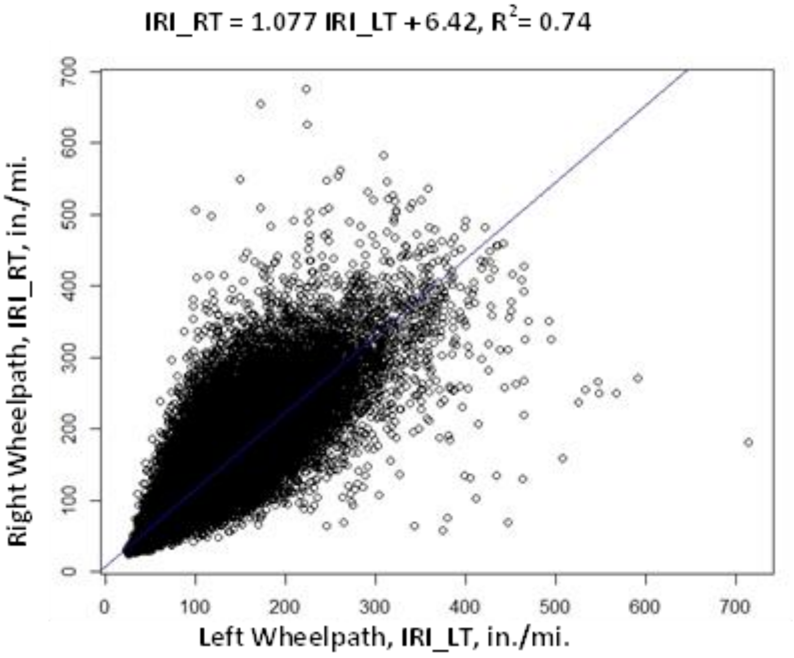


Figure 42 Comparison of the IRI values of both wheel paths for Texas roadways in 2014

The difference in IRI values between the two wheel paths represents the variability of the roughness data; Equation 12 is used to define the variability of such IRI values (Jai et al. 2016).

$$Er = \frac{|IRI_{LT} - IRI_{RT}|}{|IRI_{LT} + IRI_{RT}|} * 100\% \qquad \text{Eq 12}$$

where Er is the relative error of the roughness data.

A high relative number of errors represents inconsistencies between the IRI values of the two wheel paths. In order to define a consistency rule for the IRI data, a threshold for acceptable Er values should be determined. The threshold may differ based on the desired confidence level. Figure 43 shows the cumulative distribution of relative errors in the IRI data collected for Texas roadways in 2014. More than 90 percent of the roadway network had Er values of less than 18.1 percent. Thus, in this study it was assumed that relative errors of more than 18.1 percent represented inconsistencies. In other words, the IRI data for a given section was considered consistent if the Er values were less than 18.1 percent.

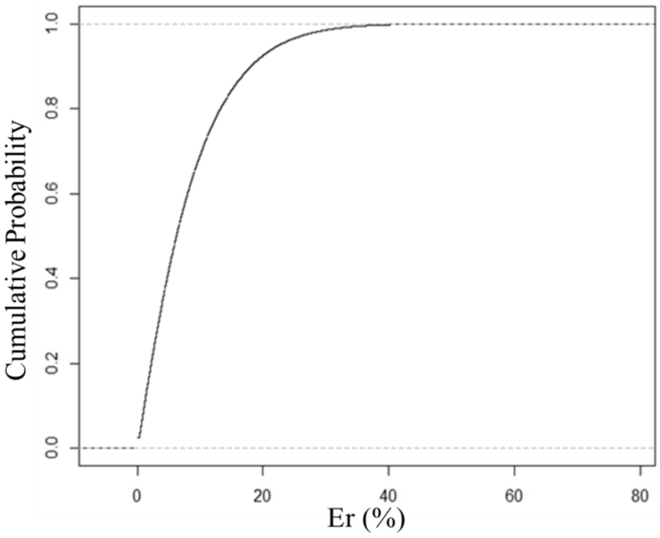


Figure 43 Cumulative distribution of relative IRI value errors for Texas roadways in 2014

Accuracy

Accuracy is the degree to which a value in the database is close to the true value of the phenomenon in the real world. This dimension of data quality is also referred to as data correctness. In pavement condition data, accuracy represents the closeness of several types of distress ratings to their real values in the field.

Although accuracy is one of the most important dimensions of data quality, it is very difficult to quantify because the true value is not known for every instance in the dataset. In pavement management, to verify and control the accuracy of collected pavement condition data, independent auditors resurvey and reanalyze random samples from the data collection (Flintsch and McGee 2009). For instance, TxDOT uses private vendors to collect pavement condition data every year for nearly its entire roadway network. These data, called PMIS, are used to measure the network's current condition and plan M&R treatments. TxDOT also conducts an independent data collection survey, called audit data, on approximately five percent of its network, in order to verify the vendor-collected data (Siabil and Gharaibeh 2016). It is assumed that audited data represent the true values and can be used to measure the accuracy of the vendor-collected data. For each data instance (i.e., pavement section), the distance function in Equation 13 (Heinrich et al. 2007) can be used to quantify the distance between the distress rating in the PMIS database (Xp) and the corresponding attribute value in the audit data (Xa):

$$d\left(X_p, X_a\right) = \frac{|X_p - X_a|}{max(X_p, X_a)} \times 100 \qquad \text{Eq 13}$$

where d (Xp, Xa) represents the distance between any distress rating (e.g., TC, AC, and LC) in the database and its corresponding true value in the field. d (Xp, Xa) is computed using a sample of the dataset. In this case, the audited sections represent the sample used for computing d (Xp, Xa).

Large distance values represent a lack of accuracy in the data for the pavement section; conversely, when d = 0 the distress value should be considered accurate. If both Xp and Xa = 0, then d = 0 because the data instance correctly represents the real world. Thus, accuracy has a negative correlation with distance, d (Xp, Xa), and is defined as within a range of 0 to 100. The accuracy of a pavement section data for any indicator can be calculated using Equation 14:

$$Accuracy_i = 100 - \ d\left(X_p, X_a\right) \qquad \text{Eq 14}$$

where accuracy$_i$ is the accuracy of a data instance (i.e. pavement section i) for an indicator (X); and d(Xp, Xa) represents the distance between any distress rating (e.g., TC, AC, and LC) in the database and its corresponding true value in the field.

The accuracy of the entire data collection for any indicator can be represented by the weighted average accuracy of the sections with audit data, as shown in Equation 15.

$$Accuracy_{net}\ (\%) = \frac{\sum_{i \in T} Accuracy_i (L_i * ESAL_i)}{\sum_{i \in T} (L_i * ESAL_i)} * 100 \qquad \text{Eq 15}$$

where L$_i$ is the length of pavement section i; ESAL$_i$ is the current 18 kip equivalent single axel load value of the section; accuracy$_i$ is the accuracy of a data instance (i.e. pavement section i) for the indicator; and T is the set of pavement sections with audit data (i.e., sample size).

Results

The above metrics for measuring data quality were implemented for network-level pavement condition data collected for the Bryan district roadway network in 2014. This dataset consisted of approximately 7,000 records and 70 columns. Each record represented a pavement section (approximately 0.5-miles in length). The columns included pavement inventory, individual distress types, and pavement condition indexes. Metrics were applied individually to several pavement performance indicators, including IRI and surface cracking data (i.e., LC, AC, TC, and BC). The results are summarized in Table 16. TxDOT collects pavement condition data annually, so all data points had the same timeliness value. All of the records were unique and there were no duplicates in the database. It was assumed that the only missing data was contained in the blank records, which were rarely found in this database. The IRI values were more complete but less valid than the surface cracking data. Audit data were not available for the IRI dataset, so the accuracy could not be measured. Overall, the Bryan roadway network dataset had a very high level of quality based on the proposed metrics. However, the consistency and accuracy dimensions had less strength than the other quality dimensions.

Generally, there is a greater chance of collecting inaccurate data if the pavement is deteriorated and distress value is not zero. Therefore, distress types that are less sensitive and rarely change (like BC), tend to have more zero values and less erroneous data. This might be why there were more accurate data for the BC values.

Table 16 Quantified data quality dimension for several performance indicators in Bryan roadway network in 2014.

Performance Indicator	Timeliness (year)	Completeness (%)	Validity (%)	Uniqueness (100)	Consistency (%)	Accuracy (%)
IRI	1	100	99.1	100	95.2	NA
AC	1	99.5	100	100	NA	77.2
LC	1	99.5	100	100	NA	77.4
TC	1	99.5	100	100	NA	79.7
BC	1	99.5	100	100	NA	99.7

8. SUMMARY, CONTRIBUTION, CONCLUSIONS, AND RECOMMENDATIONS

Summary

Transportation agencies allocate a portion of their annual budget to collecting pavement condition data as a part of their PMSs. These PMSs help agencies make efficient decisions about allocating available resources to the maintenance, rehabilitation, and renewal of roadway networks. Due to their use of more advanced technologies in collecting, storing, and manipulating data, the size of these pavement condition databases are growing rapidly, which makes controlling the quality of data even more complicated and expensive. On the other hand, high quality data is essential to efficient and reliable decision making. Thus, transportation agencies need to ensure that the dollars they invest in data are well spent, and their data is of the level of quality necessary to meet the requirements of their PMSs (e.g., data is complete, accurate, consistent, and up-to-dated). Despite the importance of this area, the majority of pavement management literature has been dedicated to data collection processes, data analysis, and decision making modeling. Only a few studies have investigated the quality of pavement condition data and its effect on PMS results. In order to fill this gap, this research assesses and enhances the quality of network-level pavement condition data.

First, this research devised and implemented a new computational technique to identify potential errors in pavement condition data. This technique integrates conventional statistical methods and heuristics. The statistical methods are used to identify outliers in uniform performance families, and the heuristics are used to delineate potential errors from extreme yet valid behaviors within these outliers. Compared to conventional statistical methods, the developed technique has the advantage of differentiating between extreme, yet valid, data points and potential errors. Second, the new technique was validated using actual pavement condition data from Texas. Audit data was compared with original data for two datasets: all sections that were audited, and the audited sections that were identified by the developed technique as containing potential errors. Third, the effect of considering multiple dimensions of error detection in pavement condition data was investigated. These dimensions are based on data properties, including time series trends in pavement condition data, variability within uniform performance families, and the consistency between several performance indicators. The results of the error detection using single or multi-dimensional techniques were compared using data from TxDOT's Brownwood district. Next, an assessment was made of the impact of the data's accuracy on predictions of the road network's future performance. RSL was used as an overall measure of network health in order to quantify this impact at both the project and network levels. Finally, several metrics were proposed for measuring quality dimensions of pavement condition data. The metrics were applied to the Bryan district roadway network dataset

Contribution

The key contributions of this research include the following:

a) Developing a new computational technique to detect potential errors in network-level pavement condition data

b) Providing a quantitative assessment of the impact of data accuracy on the output of PMSs (i.e., assessing the estimated RSL).

c) Providing metrics for measuring quality dimensions of network-level pavement condition datasets

The main merit of the developed technique is the ability to distinguish between extreme data points and potential errors, as compared to conventional statistical methods. The impact of data accuracy was quantitatively investigated and the overestimation (i.e., because of erroneous data) of the required budgets for future M&R plans was calculated. Provided metrics enables engineers to assess the quality of pavement condition datasets, and ultimately can be used to determine if pavement condition datasets (as a whole) are of acceptable quality.

Conclusions

Key conclusions of this study are defined in four categories and summarized as follows:

- A new technique was developed to detect potential errors in pavement condition data. In order to validate the results, the technique was tested on Texas pavement sections that had been audited in 2014. The following key conclusions can be made based on the new technique and validated sections:
    - Potential errors in pavement condition data can be identified and delineated by integrating statistical methods and heuristic consistency

checks. This new technique has the advantage of differentiating between extreme yet valid data instances (pavement sections that had unusual but explainable performances such as receiving treatments or rapid deterioration) and potential errors.

  - A Wilcoxon signed-rank test with α =0.05 (i.e., 95 percent confidence) validated the developed error detection method for both LC and AC. For all cracking types (i.e., LC, AC, and TC), the data points identified as potential errors had higher MAEs, as compared to the general population of audited pavement sections.
  - Potential errors were found to be randomly distributed across the Texas roadway network and were not related to a specific geographic regions, districts, climatic zones, etc.

- Accounting for several properties of pavement condition data (i.e. multiple dimensions of error detection) to identify potential errors improves the results of pavement condition error detection techniques:
  - The impact of considering performance families increases as the networks get larger and more diverse
  - Considering consistency among several indicators significantly improves the results of the error detection techniques by identifying false positives (i.e., the valid data instances that falsely detected as outliers).

    - Considering the variability in pavement performance families in conjunction with consistency checks improve error detection by identifying false negatives.

- The impact of pavement condition data accuracy on the estimated RSL of a road network was assessed using data obtained from Texas. The following conclusions are related to the impact investigation segment of this research:
    - Pavement sections detected to have potential errors had lower RSLs than the average roadway network. This meant that the errors mostly occurred in deteriorated sections with many distresses; the data for these sections cannot be easily and precisely collected, as compared with pavement sections in good condition and with little or no distress.
    - Considering the entire roadway network, errors in pavement condition data result in the underestimation of RSL and overestimation of required budget for short term plans (i.e. 1-2 year M&R planning).
    - The estimated RSLs of pavement sections detected to have potentially erroneous indicator values were highly sensitive to suspicious values of condition data. Ignoring these suspicious data (i.e., potential errors) in the RSL estimation process dramatically increased the amount of the RSL. From 165,469 pavement sections in the Texas roadway network, 12,127 sections were detected to have potential errors in at

least one of their condition indicators (i.e., rutting, AC, LC, TC). The average estimated RSL of these sections increased from 4.2 years to 11.9 years when the suspicious data was removed.

- Several metrics are provided to measure quality dimensions of pavement condition datasets. The metrics were applied to a pavement condition dataset for TxDOT's Bryan district roadway network:
    - Metrics for measuring several dimensions of pavement condition data quality were defined.
    - The developed metrics were demonstrated on a case study; however they should be validated in future research.

Recommendations

The following are recommendations based on this research for both practitioners and researchers.

- This research provided a new computational technique for detecting errors in network-level pavement condition datasets. The technique can be used by transportation agencies to determine candidate sections for field audits.
- The developed technique can be applied to check the accuracy of data collected by vendors.

- Similar techniques could be developed to detect errors in condition data for other infrastructure systems such as bridges, traffic signs, pipelines, etc.
- This study provided metrics for measuring quality dimensions of pavement condition datasets. Further testing and validation of these metrics are needed.
- Develop a composite metric for measuring the overall quality of pavement condition datasets.
- Errors and outliers can affect performance prediction models. Thus, it is recommended that this technique be implemented on datasets to remove potential errors before the development of pavement deterioration models.
- Field data collection can be used to provide ground truth of likely errors and consequently improve the effectiveness and reliability of the proposed technique.
- Future research could continue the process of data quality assessment and improvement by introducing the best method to correct and impute uncovered errors.

REFERENCES

AASHTO (1993). *Guide for Design of Pavement Structures,* American Association of State Highway and Transportation Officials, Washington, D. C.

AASHTO (2001). *Pavement Management Guide,* American Association of State Highway and Transportation Officials, Washington, D.C.

Aggarwal, C. C. (2013). *Outlier Analysis*, Springer, New York.

Arabali, P., Sakhaeifar, M., Freeman, T., Wilson, B., and Borowiec, J.(2016). "Decision-Making Tool for the Selection of Pavement Preservation Treatments in General Aviation Airport Pavements." *International Conference on Transportation and Development*, 30-41.

Augusteijn, M. and Folkert, B. (2002). "Neural Network Classification and Novelty Detection." *International Journal on Remote Sensing*, 23 (14), 2891–2902.

Baladi, G. Y., and Novak, E. C. (1992). "Pavement Condition Index-Remaining Service Life." *Pavement Management Implementation Symposium,* Atlantic City, New Jersey.

Baladi, G. Y., Dawson, T. A., Dean, C. M., Haider, S. W., and Chatti, K. (2011). "The Calculation of the Remaining Service Life Based on the Pavement Distress Data." *In Proceedings of the ASCE Conference on Integrated Transportation and Development for a Better Tomorrow,* Chicago, Illinois.

Banerjee, A., Chandola, V., Kumar, V., Srivastava, J., and Lazarevic, A. (2008). "Anomaly detection: A tutorial." *SIAM Conference on Data Mining,* Atlanta, Georgia.

Batini, C. and Scannapieca, M. (2006). *Data Quality Concepts, Methodologies and Techniques*. Springer-Verlag Berlin Heidelberg.

Bell, J. (2014). *Machine Learning: Hands-On for Developers and Technical Professionals*. John Wiley & Sons. New Jersey

Beaumont, L. R. (1996). "Metrics: a Practical Example." *The PDMA Handbook of New Product Development,* Toimittaneet Rosenau, Milton D., Griffin, Abbie, Castellion, George ja Anschuetz, Ned. John Wiley & Sons, New Jersey.

Bhuyan, M. H., Bhattacharyya, D. K., and Kalita, J. K. (2012). " Survey on Incremental Approaches for Network Anomaly Detection." *International Journal of Communication Networks and Information Security,* vol. 3, no. 3, p. 14

Black, K. (2011). *Business Statistics: for Contemporary Decision Making*. John Wiley & Sons, New Jersey

Bogus, S. M., Migliaccio, G. C., and Cordova, A. A. (2010). “Assessment of Data Quality for Evaluations of Manual Pavement Distress.” *Transportation Research Record* 21(70), 1-8.

Bolton, R. J., and Hand, D. J. (2001). “Unsupervised Profiling Methods for Fraud Detection.” *Credit Scoring and Credit Control VII*, 235-255.

Chandola, V., Banerjee, A., and Kumar, V. (2009). “Anomaly Detection: A Survey.” *ACM computing surveys (CSUR)*, 41(3), 15.

Dasu, T., and Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning (Vol. 479)*. John Wiley & Sons. New Jersey

Elkins, G. E., Thompson, T. M., Groeger, J. L., Visintine, B., and Rada, G. R. (2013). *Reformulated Pavement Remaining Service Life Framework.* Publication No. FHWA-HRT-13- 038, Federal Highway Administration, McLean, Virginia, USA, August 2013.

Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P. N., Kumar, V., Srivastava, J., and Dokas, P. (2004). “Minds-Minnesota Intrusion Detection System.” *Next Generation Data Mining*, 199-218.

Ferreira, A., and Santos, J. (2012). “LCCA System for Pavement Management: Sensitivity Analysis to the Discount Rate.” *Procedia-Social and Behavioral Sciences*, 53, 1172-1181.

Flintsch, G., and McGhee, K. K. (2009). *Quality Management of Pavement condition data Collection*, National Cooperative Highway Research Program, Washington, D. C.

Flintsch, G. W., and Bryant, J. W. (2006*). Asset Management Data Collection for Supporting Decision Processes.* US Department of Transportation, Federal Highway Administration, Washington, DC.

Gaddam, S. R., Phoha, V. V., and Balagani, K. S. (2007). “K-Means+ ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods.” *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 345-354.

Gharaibeh, N., Freeman, T., Saliminejad, S., Wimsatt, A., Chang-Albitres, C., Nazarian, S., Abdallah, I., Weissmann, J., Weissmann, A., T. Papagiannakis, A., and Gurganus, C. (2012). *Evaluation and Development of Pavement Scores, Performance Models* *and Needs Estimates for the TxDOT Pavement Management Information System.* Final Report (No. FHWA/TX-12/0-6386-3). Texas A&M Transportation Institute, Texas A&M University System.

Gharaibeh, N. G., Narciso, P., Cha, Y., Oh, J., Menendez, J. R., Dessouky, S., and Wimsatt, A. (2014)."*A Methodology to Support the Development of 4-year Pavement Management Plan"* (No. FHWA/TX-14/0-6683-1)

Gogoi, P., Borah, B., and Bhattacharyya, D. K. (2010*).* "Anomaly Detection Analysis of Intrusion Data Using Supervised & Unsupervised Approach." *Journal of Convergence Information Technology,* 5(1), 95-110.

Gong, P., and Mu, L. (2000). "Error Detection through Consistency Checking." *Geographic Information Sciences*, 6(2), 188-193.

Grabe, M. (2010). *Generalized Gaussian Error Calculus,* Springer, New York, NY.

Griffith, D. A., Haining, R., and Arbia, G. (1994). "Heterogeneity of Attribute Sampling Error in Spatial Datasets." *Geographical Analysis*, 26(4), 300-320.

Groeger, J., Stephanos, P., Dorsey, P., and Chapman, M. (2003). "Implementation of Automated Network-Level Crack Detection Processes in Maryland." *Transportation Research Record: Journal of the Transportation Research Board,* (1860), 109-116.

Gurganus, C., and Gharaibeh, N. G. (2012). "Pavement Preservation Project Selection and Prioritization: A Competitive Approach." *91st Annual Meeting of the Transportation Research Board, Transportation Research Board,* Washington, D. C.

Haas, R., W. R. Hudson, and J. Zaniewski. (1994) *Modern Pavement Management.* Krieger Publishing, Malabar, Fla.

Hall, K. T., Correa, C. E., Carpenter, S. H., and Elliott, R. P. (2003). "Guidelines for Life-Cycle Cost Analysis of Pavement Rehabilitation Strategies." *Urbana*, 51, 61801.

Hawkins, D. M. (1980). *Identification of Outliers,* Springer, New York.

Hosten, A., Chowdhury, T., Shekharan, R. A., Ayotte, M., and Coggins, E. (2015). "Use of VDOT's Pavement Management System to Proactively Plan and Monitor Pavement Maintenance and Rehabilitation Activities to Meet the Agency's Performance Target". *In 9th International Conference on Managing Pavement Assets*. Washington D.C. Metropolitan Area.

Heinrich, B., Kaiser, M., and Klier, M., (2007). “How to Measure Data Quality? A Metric Based Approach.” *Proceedings of the 28th International Conference on Information Systems (ICIS).* December 2007, Montreal,(Canada).

Jia, X., Huang, B., Dong, Q., Zhu, D., and Maxwell, J. (2016). “Influence of Pavement Condition Data Variability on Network-Level Maintenance decision.” *Transportation Research Record: Journal of the Transportation Research Board,* (2589), 20-31.

Kinoshita, K., Teshima, T., Ohno, Y., Inoue, T., Yamashita, T., Hiraoka, M., and Japanese PCS Working Group. (2003). “Logical Checking Function Increases the Accuracy of Data Entry in the Patterns of Care Study.” *Strahlentherapie und Onkologie*, 179(2), 107-112.

Kim, D. Y., Menches, C. L., Kim, D., Kweon, T., and Huh, Y. (2014). “Comparative Simulation Analysis of Pavement Technology for A Decision Support System in the US Road Renewal Industry.” *KSCE Journal of Civil Engineering,*18(4), 920-926.

Larson, C. D., Sami, N., and Luhr, D. R. (2000). *"Structured Approach to Managing Quality of Pavement Distress Data: Virginia Department of Transportation Experience." Transportation Research Record,* 1699, 72-80.

Lazarevic, A., Kumar, V., and Srivastava, J. (2005). *Intrusion Detection: A Survey*. In Managing Cyber Threats (pp. 19-78). Springer US.

Livneh, M. (1994). “Repeatability and Reproducibility of Manual Pavement Distress Survey Methods.” *3rd International Conference on Managing Pavements*. San Antonio, Texas.

Loshin, D. (2006). “Monitoring Data Quality Performance Using Data Quality Metrics: A White Paper.” *Informatica.* November.

Lytton, R. L. (1987). “Concepts of Pavement Performance Prediction and Modeling.” *In Proc., 2nd North American Conference on Managing Pavements* (Vol. 2).

Malik, K., Sadawarti, H., and G S, K. (2014). “Comparative Analysis of Outlier Detection Techniques.” *International Journal of Computer Applications*, 97(8), 12-21.

Maletic, J. I., and Marcus, A. (2000). “Data Cleansing: Beyond Integrity Analysis.” *Information Quality* (pp. 200-209).

Matteoli, S., Diani, M., and Corsini, G. (2010). “A Tutorial Overview of Anomaly Detection in Hyperspectral Images.” *Aerospace and Electronic Systems Magazine,* IEEE, 25(7), 5-28.

McNeil, S., and Humplick, F. (1991). *"*Evaluation of Errors in Automated Pavement-Distress Data Acquisition.*" Journal of Transportation Engineering,* 117(2), 224-241.

Menendez, J., Siabil, S., Narciso, P., and Gharaibeh, N. (2013). "Prioritizing Infrastructure Maintenance and Rehabilitation Activities Under Various Budgetary Scenarios: Evaluation of Worst-First and Benefit-Cost Analysis Approaches." *Transportation Research Record: Journal of the Transportation Research Board,* (2361), 56-62.

Migliaccio, G. C., Bogus, S. M., and Cordova, A. A. (2011). "Continuous Quality Improvement Techniques for Data Collection in Asset Management Systems." *ASCE Journal of Construction Engineering and Management*.

Morian, D., Stoeffels, S., and Firth, D. J. (2002). "Quality Management of Pavement Performance Data." *Pavement Evaluation 2002,* Al-Qadi, and T. Clarck, eds. Roanoke, Va.

Motro, A., and Rakov, I. (1998). "Estimating the Quality of Databases." *Flexible query answering systems* (pp. 298-307). Springer Berlin Heidelberg.

Mukherjee, B., Heberlein, L. T., and Levitt, K. N. (1994). *Network intrusion detection*. Network, IEEE, 8.3, 26-41.

NCHRP Report 814, "Data to Support Transportation Agency Business Needs: A Self-Assessment Guide", 2015.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., and Sun, X. (2011). "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature." *Decision Support Systems,* 50(3), 559-569

Patcha, A., and Park, J. M. (2007). "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends." *Computer networks,* 51(12), 3448-3470.

Phillips, D. L., and Marks, D. G. (1996). "Spatial Uncertainty Analysis: Propagation of Interpolation Errors in Spatially Distributed Models." *Ecological Modelling,* 91(1), 213-229.

Pierce, L. M., McGovern, G., and Zimmerman, K. A. (2013). Practical Guide for Quality Management of Pavement condition data Collection.

Pierce, L. M., and Zimmerman, K. A. (2015). "Quality Management for Pavement Condition Data Collection." *In 9th International Conference on Managing Pavement Assets,* May, Washington D.C.

Ramaswamy, S., Rastogi, R., and Shim, K. (2000). “Efficient Algorithms for Mining Outliers from Large Datasets.” *In ACM SIGMOD Record* (Vol. 29, No. 2, pp. 427-438). ACM.

Ratsch, G., Mika, S., Scholkopf, B., and Müller, K. R. (2002). “Constructing Boosting Algorithms from Svms: An Application to One-Class Classification.” *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 24(9), 1184-1199.

Reddy, A. M., and Kumar, P. (2006). “Outlier Detection in Wireless Sensor Networks Using Bayesian Belief Networks.” *In 1st International Conference on Communication Systems Software & Middleware* (pp. 1-6).

Redman, T. C., (1996). *Data Quality for the Information Age*, Artech House, Boston.

Saliminejad, S. (2012).*GIS-Based Probabilistic Approach for Assessing and Enhancing Infrastructure Data Quality*, PhD’s Thesis, Texas A&M University, December 2012.

Saliminejad, S., and Gharaibeh, N. G. (2013). “Impact of Error in Pavement condition data on the Output of Network-Level Pavement Management Systems.*”* *Transportation Research Record: Journal of the Transportation Research Board,* 2366(1), 110-119.

Scannapieco, M., and Catarci, T. (2002). “Data Quality under A Computer Science Perspective.” *Archivi & Computer*, 2, 1-15.

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

Shahin, M. Y., and Kohn, S. D. (1979). “Development of a Pavement Condition Rating Procedure for Roads, Streets, and Parking Lots. Volume II. Distress Identification Manual (No. CERL-TR-M-268-VOL-2).” *Construction Engineering Research Lab* (Army) Champaign Il.

Shekharan, R., Frith, D., Chowdhury, T., Larson, C., and Morian, D. (2007). “Effects of Comprehensive Quality Assurance/Quality Control Plan on Pavement Management.” *Transportation Research Record*(1990), 65-71.

Siabil, S. Z., and Gharaibeh, N. G. (2016). “A Computational Technique for Detecting Errors in Network-Level Pavement Condition Data.” *Transportation Research Record: Journal of the Transportation Research Board,* (2589), 14-19.

Smith, R., Freeman, T., and Pendleton, O. (1998). "Evaluation of Automated Pavement Distress Data Collection Procedures for Local Agency Pavement Management." *4th International Conference on Managing Pavements.* Durban, South Africa.

Spiegelman, C. H., Park, E. S., and Rilett, L. R. (2011). *Transportation Statistics and Microsimulation,* Chapman & Hall/CRC.

Tan, S. G., and Cheng, D. (2015). “Improving Data Quality for Pavement Management System.” *In 9th International Conference on Managing Pavement Assets*, Washington D.C.

TxDOT (2015). *Pavement Management Information System (PMIS) User’s Manual*. Texas Department of Transportation, Austin, Texas.

U.S. Environmental Protection Agency (EPA). (2006). *Data Quality Assessment: Statistical Methods for Practitioners.* (EPA/240/B-06/003).

U.S. Office of Management and Budget (USOMB). (2015). “Discount Rates for Cost-Effectiveness, Lease, Purchase and Related Analyses.” *Appendix C to guidelines and discount rates for benefit-cost analysis of federal programs,* Washington DC.

Vanier, D. D. (2001). “Why Industry Needs Asset Management Tools.” *Journal of computing in civil engineering*, 15(1), 35-43.

Vasconcelos, G. C., Fairhurst, M. C., and Bisset, D. L. (1995). “Investigating Feedforward Neural Networks with Respect to the Rejection of Spurious Patterns.” *Pattern Recognition Letters*, 16(2), 207-212.

Walls III, J., and Smith, M. R. (1998). “Life-Cycle Cost Analysis in Pavement Design-Interim Technical Bulletin.” (No. FHWA-SA-98-079,).

Wand, Y., and Wang, R. Y. (1996). “Anchoring Data Quality Dimensions in Ontological Foundations.” *Communications of the ACM*, 39(11), 86-95.

Wang, R.Y., and Strong, D.M., (1996). “Beyond Accuracy: What Data Quality Means to Data Consumers.” *Journal of Management Information Systems,* 12(4), 5–33.

Wang, R. Y., Storey, V. C., and Firth, C. P. (1995). “A Framework for Analysis of Data Quality Research.” *Knowledge and Data Engineering, IEEE Transactions on,* 7(4), 623-640.

Wong, W. K., Moore, A., Cooper, G., and Wagner, M. (2002). “Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks.” *In AAAI/IAAI* (pp. 217-223).

Yu, J., Chou, E. Y., and Luo, Z. (2007). “Development of Linear Mixed Effects Models for Predicting Individual Pavement Conditions.” *Journal of transportation engineering*, 133(6), 347-354

APPENDIX A

COEFFICIENTS OF PERFORMANCE PREDICTION MODELS

Table A.1 Model coefficients for Zone 1-pavement family A

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.41	75.16	100.00	0.43	102.38	100.00	0.39	58.34
	LR	100.00	0.47	79.75	100.00	0.47	107.18	100.00	0.42	66.85
	MR	100.00	0.52	80.38	100.00	0.55	121.09	100.00	0.47	67.14
	HR	100.00	0.53	91.69	100.00	0.58	122.99	100.00	0.55	70.69
Deep Rutting	PM	100.00	0.54	88.24	100.00	0.76	60.35	100.00	0.58	95.02
	LR	100.00	0.55	101.18	100.00	0.80	68.37	100.00	0.60	113.20
	MR	100.00	0.56	115.81	100.00	0.88	80.79	100.00	0.65	116.07
	HR	100.00	0.57	133.23	100.00	1.01	83.07	100.00	0.73	123.10
Failures	PM	20.00	1.11	23.48	20.00	1.30	19.85	20.00	3.61	8.86
	LR	20.00	1.17	24.55	20.00	1.33	20.51	20.00	3.88	9.10
	MR	20.00	1.26	27.30	20.00	1.37	21.50	20.00	4.19	9.14
	HR	20.00	1.40	30.05	20.00	1.40	21.49	20.00	4.54	9.18
Block Cracking	PM	100.00	3.73	114.51	100.00	0.96	45.92	100.00	6.75	83.46
	LR	100.00	3.81	130.91	100.00	1.83	47.93	100.00	7.69	94.98
	MR	100.00	4.46	142.20	100.00	2.58	48.74	100.00	8.80	108.82
	HR	100.00	4.98	146.76	100.00	3.14	58.32	100.00	10.10	125.49
Alligator Cracking	PM	100.00	0.58	101.42	100.00	0.49	96.93	100.00	4.24	8.20
	LR	100.00	0.62	104.61	100.00	0.53	113.11	100.00	5.10	9.67
	MR	100.00	0.72	115.98	100.00	0.58	133.61	100.00	5.73	11.28
	HR	100.00	0.73	135.90	100.00	0.65	159.49	100.00	6.06	11.90
Longitudinal Cracking	PM	500.00	0.52	116.51	500.00	0.53	90.24	500.00	0.44	69.52
	LR	500.00	0.60	133.63	500.00	0.54	104.52	500.00	0.50	71.55
	MR	500.00	0.67	146.86	500.00	0.56	123.32	500.00	0.51	81.25
	HR	500.00	0.71	153.66	500.00	0.59	146.45	500.00	0.58	84.37
Transverse Cracking	PM	20.00	0.71	95.12	20.00	0.49	68.47	20.00	0.88	20.33
	LR	20.00	1.11	109.50	20.00	0.54	68.87	20.00	0.92	21.07
	MR	20.00	1.52	125.33	20.00	0.55	77.01	20.00	0.99	22.61
	HR	20.00	1.95	143.04	20.00	0.61	78.23	20.00	1.09	25.68
Patching	PM	100.00	0.38	101.23	100.00	0.64	49.65	100.00	0.52	87.67
	LR	100.00	0.41	105.68	100.00	0.65	53.60	100.00	0.52	100.95
	MR	100.00	0.48	119.25	100.00	0.65	57.65	100.00	0.53	115.41
	HR	100.00	0.50	119.67	100.00	0.78	61.64	100.00	0.54	131.59

Table A.2 Model coefficients for Zone 1-pavement family B

Distress Type	Treatment Type	Low Traffic			Medium Traffic			High Traffic		
		α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.39	90.81	100.00	0.73	51.56	100.00	0.30	99.51
	LR	100.00	0.41	106.60	100.00	0.74	56.02	100.00	0.33	103.93
	MR	100.00	0.43	127.73	100.00	0.75	60.38	100.00	0.38	115.99
	HR	100.00	0.46	130.80	100.00	0.76	64.74	100.00	0.39	137.75
Deep Rutting	PM	100.00	0.60	101.99	100.00	0.73	82.60	100.00	0.51	101.42
	LR	100.00	0.67	121.58	100.00	0.86	91.14	100.00	0.58	120.33
	MR	100.00	0.78	124.74	100.00	0.98	96.09	100.00	0.68	122.26
	HR	100.00	0.94	131.24	100.00	1.08	97.65	100.00	0.83	127.33
Failures	PM	20.00	0.42	118.33	20.00	0.68	97.50	20.00	0.57	109.25
	LR	20.00	0.62	129.27	20.00	0.72	98.24	20.00	1.18	126.57
	MR	20.00	0.66	153.80	20.00	0.79	102.39	20.00	1.71	144.27
	HR	20.00	0.89	167.58	20.00	0.90	110.55	20.00	2.12	160.90
Block Cracking	PM	100.00	0.63	118.81	100.00	3.89	50.61	100.00	7.79	25.39
	LR	100.00	0.80	133.78	100.00	4.21	55.19	100.00	9.31	26.85
	MR	100.00	0.90	140.54	100.00	4.58	59.86	100.00	9.62	29.82
	HR	100.00	1.18	165.32	100.00	4.99	65.57	100.00	10.32	34.55
Alligator Cracking	PM	100.00	0.46	74.29	100.00	0.54	53.38	100.00	3.34	9.15
	LR	100.00	0.53	78.08	100.00	0.56	58.71	100.00	3.64	9.28
	MR	100.00	0.57	93.06	100.00	0.59	66.42	100.00	4.03	9.49
	HR	100.00	0.57	104.61	100.00	0.63	75.32	100.00	4.56	9.71
Longitudinal Cracking	PM	500.00	0.57	25.48	500.00	0.49	67.22	500.00	0.54	72.19
	LR	500.00	0.62	27.71	500.00	0.58	78.01	500.00	0.61	74.44
	MR	500.00	0.71	33.43	500.00	0.74	80.44	500.00	0.64	85.46
	HR	500.00	0.73	34.89	500.00	0.74	87.33	500.00	0.75	91.01
Transverse Cracking	PM	20.00	1.58	6.58	20.00	0.29	107.15	20.00	6.98	6.29
	LR	20.00	1.81	7.59	20.00	0.32	116.81	20.00	7.87	6.95
	MR	20.00	1.92	8.34	20.00	0.33	118.06	20.00	8.78	7.89
	HR	20.00	2.25	9.50	20.00	0.38	133.00	20.00	9.69	9.66
Patching	PM	100.00	0.36	55.57	100.00	0.59	77.35	100.00	1.08	10.30
	LR	100.00	0.38	63.16	100.00	0.68	82.08	100.00	1.27	11.72
	MR	100.00	0.41	74.20	100.00	0.75	82.76	100.00	1.46	13.35
	HR	100.00	0.47	75.09	100.00	0.78	95.87	100.00	1.62	14.48

Table A.3 Model coefficients for Zone 1-pavement family C

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.30	111.93	Not Enough Data					
	LR	100.00	0.35	125.82						
	MR	100.00	0.38	134.17						
	HR	100.00	0.39	134.69						
Deep Rutting	PM	100.00	0.51	97.89						
	LR	100.00	0.54	99.55						
	MR	100.00	0.61	106.95						
	HR	100.00	0.71	119.97						
Failures	PM	20.00	0.60	85.56						
	LR	20.00	0.72	96.57						
	MR	20.00	0.85	107.31						
	HR	20.00	0.99	116.75						
Block Cracking	PM	100.00	5.51	112.31						
	LR	100.00	6.28	113.75						
	MR	100.00	7.34	119.19						
	HR	100.00	8.79	127.92						
Alligator Cracking	PM	100.00	0.76	48.31						
	LR	100.00	0.91	50.79						
	MR	100.00	1.07	52.81						
	HR	100.00	1.23	53.91						
Longitudinal Cracking	PM	500.00	0.64	84.90						
	LR	500.00	0.77	94.96						
	MR	500.00	0.90	104.07						
	HR	500.00	1.03	111.74						
Transverse Cracking	PM	20.00	9.15	58.53						
	LR	20.00	10.19	60.41						
	MR	20.00	10.63	69.24						
	HR	20.00	12.48	74.90						
Patching	PM	100.00	0.41	69.14						
	LR	100.00	0.46	70.51						
	MR	100.00	0.47	80.20						
	HR	100.00	0.54	83.32						

Table A.4 Model coefficients for Zone 2-pavement family A

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.42	110.20	100.00	0.50	91.77	100.00	0.52	71.62
	LR	100.00	0.47	121.74	100.00	0.52	107.61	100.00	0.58	74.26
	MR	100.00	0.50	125.66	100.00	0.55	129.57	100.00	0.61	85.16
	HR	100.00	0.59	145.28	100.00	0.59	132.43	100.00	0.71	90.54
Deep Rutting	PM	100.00	0.62	85.47	100.00	0.70	76.20	100.00	0.89	44.97
	LR	100.00	0.75	95.17	100.00	0.83	83.33	100.00	1.05	46.10
	MR	100.00	0.89	104.45	100.00	0.88	89.09	100.00	1.18	54.08
	HR	100.00	1.03	112.31	100.00	1.53	93.10	100.00	1.28	61.40
Failures	PM	20.00	0.87	21.95	20.00	0.55	111.45	20.00	0.78	69.14
	LR	20.00	0.91	23.32	20.00	0.63	122.81	20.00	0.86	82.97
	MR	20.00	1.00	25.56	20.00	0.66	126.71	20.00	1.02	90.45
	HR	20.00	1.12	29.47	20.00	0.78	145.99	20.00	1.10	105.81
Block Cracking	PM	100.00	0.57	97.50	100.00	0.54	89.19	100.00	0.88	85.46
	LR	100.00	0.60	113.66	100.00	0.55	102.78	100.00	1.04	94.11
	MR	100.00	0.64	131.51	100.00	0.57	119.77	100.00	1.20	101.08
	HR	100.00	0.66	151.23	100.00	0.58	139.97	100.00	1.34	105.18
Alligator Cracking	PM	100.00	0.49	68.28	100.00	0.54	35.51	100.00	0.95	19.85
	LR	100.00	0.55	68.53	100.00	0.62	42.29	100.00	0.99	20.93
	MR	100.00	0.56	76.35	100.00	0.66	46.25	100.00	1.05	22.33
	HR	100.00	0.62	76.19	100.00	0.78	47.48	100.00	1.13	23.74
Longitudinal Cracking	PM	500.00	0.65	41.91	500.00	0.39	75.34	500.00	0.47	115.36
	LR	500.00	0.76	42.71	500.00	0.45	79.93	500.00	0.54	131.06
	MR	500.00	0.85	50.70	500.00	0.50	80.72	500.00	0.60	142.35
	HR	500.00	0.91	57.23	500.00	0.52	93.19	500.00	0.63	146.90
Transverse Cracking	PM	20.00	0.66	50.51	20.00	0.51	68.85	20.00	0.63	60.35
	LR	20.00	0.67	54.85	20.00	1.21	81.97	20.00	0.67	65.17
	MR	20.00	0.67	59.12	20.00	1.38	86.76	20.00	0.69	69.94
	HR	20.00	0.68	63.36	20.00	1.95	98.37	20.00	0.69	74.61
Patching	PM	100.00	0.58	54.80	100.00	0.42	110.20	100.00	0.60	67.13
	LR	100.00	0.61	61.13	100.00	0.47	121.74	100.00	0.66	79.75
	MR	100.00	0.64	69.17	100.00	0.50	125.66	100.00	0.79	86.73
	HR	100.00	0.70	80.26	100.00	0.59	145.28	100.00	0.85	101.63

Table A.5 Model coefficients for Zone 2-pavement family B

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.49	93.49	100.00	0.74	46.21	Not Enough Data		
	LR	100.00	0.51	110.67	100.00	1.01	53.71			
	MR	100.00	0.55	112.98	100.00	1.09	57.91			
	HR	100.00	0.61	118.76	100.00	1.40	68.24			
Deep Rutting	PM	100.00	0.82	55.66	100.00	0.59	118.61			
	LR	100.00	0.84	61.69	100.00	0.69	137.57			
	MR	100.00	0.88	69.12	100.00	0.78	153.66			
	HR	100.00	0.94	78.38	100.00	0.85	164.70			
Failures	PM	20.00	2.87	10.68	20.00	1.22	19.56			
	LR	20.00	3.20	11.67	20.00	1.25	20.57			
	MR	20.00	3.75	13.25	20.00	1.29	21.63			
	HR	20.00	3.86	15.73	20.00	1.33	22.69			
Block Cracking	PM	100.00	4.58	37.62	100.00	4.64	34.94			
	LR	100.00	5.44	44.08	100.00	5.43	39.55			
	MR	100.00	5.88	46.13	100.00	5.69	40.32			
	HR	100.00	6.93	52.43	100.00	6.41	43.43			
Alligator Cracking	PM	100.00	0.74	53.56	100.00	0.73	39.14			
	LR	100.00	0.76	59.13	100.00	0.83	46.76			
	MR	100.00	0.79	65.12	100.00	0.90	52.78			
	HR	100.00	0.82	72.21	100.00	0.92	56.47			
Longitudinal Cracking	PM	500.00	0.28	111.35	500.00	0.37	86.61			
	LR	500.00	0.32	124.55	500.00	0.37	99.37			
	MR	500.00	0.34	132.67	500.00	0.38	113.92			
	HR	500.00	0.35	159.94	500.00	0.39	131.58			
Transverse Cracking	PM	20.00	1.08	22.90	20.00	0.79	16.22			
	LR	20.00	1.14	24.49	20.00	0.81	16.46			
	MR	20.00	1.23	26.20	20.00	0.84	16.62			
	HR	20.00	1.36	28.91	20.00	0.89	17.79			
Patching	PM	100.00	0.34	102.28	100.00	0.71	94.44			
	LR	100.00	0.37	107.66	100.00	0.74	111.07			
	MR	100.00	0.43	123.83	100.00	0.78	111.41			
	HR	100.00	0.46	127.92	100.00	0.85	114.06			

Table A.6 Model coefficients for Zone 2-pavement family C

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.58	49.93	100.00	0.46	96.74	Not Enough Data		
	LR	100.00	0.59	53.87	100.00	0.49	98.00			
	MR	100.00	0.60	58.23	100.00	0.54	104.16			
	HR	100.00	0.60	62.57	100.00	0.63	115.82			
Deep Rutting	PM	100.00	0.60	90.24	100.00	0.65	103.52			
	LR	100.00	0.60	101.52	100.00	0.71	108.86			
	MR	100.00	0.80	112.77	100.00	0.82	122.33			
	HR	100.00	1.01	123.06	100.00	0.86	145.53			
Failures	PM	20.00	0.78	100.18	20.00	4.36	90.05			
	LR	20.00	0.84	101.50	20.00	4.70	104.11			
	MR	20.00	0.93	108.45	20.00	5.11	121.38			
	HR	20.00	1.08	120.54	20.00	5.59	141.92			
Block Cracking	PM	100.00	3.17	42.00	100.00	9.87	33.50			
	LR	100.00	3.38	44.69	100.00	11.39	34.21			
	MR	100.00	3.61	48.17	100.00	12.65	39.69			
	HR	100.00	3.87	51.67	100.00	13.49	44.76			
Alligator Cracking	PM	100.00	0.62	65.51	100.00	0.47	92.92			
	LR	100.00	0.67	77.95	100.00	0.49	110.43			
	MR	100.00	0.79	82.99	100.00	0.52	111.54			
	HR	100.00	0.83	94.98	100.00	0.58	116.91			
Longitudinal Cracking	PM	500.00	0.48	117.56	500.00	0.48	105.14			
	LR	500.00	0.56	136.53	500.00	0.53	112.14			
	MR	500.00	0.63	152.26	500.00	0.63	129.81			
	HR	500.00	0.68	162.75	500.00	0.69	137.24			
Transverse Cracking	PM	20.00	0.84	49.84	20.00	0.77	111.83			
	LR	20.00	0.84	53.07	20.00	0.86	122.68			
	MR	20.00	1.00	55.46	20.00	0.90	124.05			
	HR	20.00	1.16	56.55	20.00	1.03	138.96			
Patching	PM	100.00	1.16	28.63	100.00	0.37	101.70			
	LR	100.00	1.25	31.85	100.00	0.44	120.93			
	MR	100.00	1.41	36.46	100.00	0.54	122.24			
	HR	100.00	1.67	36.09	100.00	0.68	125.78			

Table A.7 Model coefficients for Zone 3-pavement family A

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.60	38.00	100.00	0.72	49.36	100.00	0.39	93.20
	LR	100.00	0.69	45.74	100.00	0.72	52.28	100.00	0.41	111.05
	MR	100.00	0.75	51.83	100.00	0.86	54.69	100.00	0.45	113.92
	HR	100.00	0.76	55.28	100.00	1.01	57.32	100.00	0.50	121.25
Deep Rutting	PM	100.00	0.60	38.00	100.00	0.66	93.30	100.00	1.47	19.47
	LR	100.00	0.69	45.74	100.00	0.68	109.59	100.00	1.50	19.24
	MR	100.00	0.75	51.83	100.00	0.72	131.53	100.00	1.50	22.99
	HR	100.00	0.76	55.28	100.00	0.78	134.19	100.00	1.79	27.66
Failures	PM	20.00	0.95	45.35	20.00	0.68	87.47	20.00	9.56	100.94
	LR	20.00	1.11	46.06	20.00	0.69	98.27	20.00	10.69	110.36
	MR	20.00	1.25	53.97	20.00	0.82	110.18	20.00	11.17	112.24
	HR	20.00	1.35	61.19	20.00	0.97	122.22	20.00	13.09	128.18
Block Cracking	PM	100.00	0.91	52.14	100.00	6.17	57.67	100.00	7.31	14.32
	LR	100.00	0.92	55.67	100.00	6.34	69.79	100.00	8.45	14.39
	MR	100.00	0.92	59.56	100.00	7.18	77.09	100.00	9.86	14.54
	HR	100.00	1.10	63.36	100.00	7.46	77.46	100.00	11.59	15.69
Alligator Cracking	PM	100.00	0.60	94.44	100.00	0.50	95.69	100.00	0.70	73.43
	LR	100.00	0.63	111.68	100.00	0.53	96.07	100.00	0.79	75.55
	MR	100.00	0.67	113.84	100.00	0.58	100.16	100.00	0.83	87.17
	HR	100.00	0.74	119.43	100.00	0.66	108.68	100.00	0.96	92.43
Longitudinal Cracking	PM	500.00	0.39	84.13	500.00	0.44	57.29	500.00	0.30	70.76
	LR	500.00	0.47	94.78	500.00	0.46	64.57	500.00	0.33	74.03
	MR	500.00	0.56	106.02	500.00	0.51	76.98	500.00	0.34	86.91
	HR	500.00	0.66	117.05	500.00	0.58	79.33	500.00	0.40	95.59
Transverse Cracking	PM	20.00	0.41	84.04	20.00	0.55	33.12	20.00	0.32	113.93
	LR	20.00	0.49	94.64	20.00	0.62	38.69	20.00	0.37	129.94
	MR	20.00	0.58	105.72	20.00	0.64	42.05	20.00	0.41	141.41
	HR	20.00	0.68	116.56	20.00	0.74	49.59	20.00	0.43	146.17
Patching	PM	100.00	0.55	93.11	100.00	0.59	60.92	100.00	0.58	97.59
	LR	100.00	0.59	96.04	100.00	0.63	70.20	100.00	0.62	98.90
	MR	100.00	1.12	106.52	100.00	0.71	70.67	100.00	0.69	104.70
	HR	100.00	2.26	126.90	100.00	0.84	75.60	100.00	0.79	116.03

Table A.8 Model coefficients for Zone 3-pavement family B

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.56	49.36	100.00	0.57	76.78	100.00	0.51	87.86
	LR	100.00	0.57	52.75	100.00	0.65	81.89	100.00	0.52	101.12
	MR	100.00	0.57	57.01	100.00	0.72	82.44	100.00	0.52	115.75
	HR	100.00	0.57	60.89	100.00	0.74	93.67	100.00	0.53	133.15
Deep Rutting	PM	100.00	0.56	49.36	100.00	0.76	76.78	100.00	0.92	51.94
	LR	100.00	0.57	52.75	100.00	0.87	81.03	100.00	0.92	55.58
	MR	100.00	0.57	57.01	100.00	0.93	95.61	100.00	1.11	59.38
	HR	100.00	0.57	60.89	100.00	0.93	106.60	100.00	1.32	62.07
Failures	PM	20.00	0.78	83.75	20.00	0.65	84.70	20.00	0.71	80.98
	LR	20.00	0.92	92.56	20.00	0.78	94.81	20.00	0.86	88.21
	MR	20.00	1.06	99.13	20.00	0.90	103.75	20.00	1.67	92.33
	HR	20.00	1.17	101.83	20.00	1.03	111.23	20.00	2.42	92.55
Block Cracking	PM	100.00	5.92	91.20	100.00	6.49	47.84	100.00	2.83	93.88
	LR	100.00	6.00	95.02	100.00	6.87	54.13	100.00	3.25	98.62
	MR	100.00	6.64	108.23	100.00	7.75	63.40	100.00	3.40	113.26
	HR	100.00	6.66	111.74	100.00	9.23	64.24	100.00	3.87	117.22
Alligator Cracking	PM	100.00	0.58	84.99	100.00	0.51	104.29	100.00	0.65	97.12
	LR	100.00	0.69	95.36	100.00	0.56	110.31	100.00	0.69	97.17
	MR	100.00	0.81	104.94	100.00	0.66	126.22	100.00	0.75	101.11
	HR	100.00	0.94	114.17	100.00	0.71	130.35	100.00	0.86	109.48
Longitudinal Cracking	PM	500.00	0.31	72.86	500.00	0.39	91.01	500.00	0.47	54.04
	LR	500.00	0.35	77.01	500.00	0.40	106.77	500.00	0.49	60.22
	MR	500.00	0.38	91.16	500.00	0.42	107.08	500.00	0.52	68.45
	HR	500.00	0.39	102.18	500.00	0.46	109.77	500.00	0.57	79.80
Transverse Cracking	PM	20.00	0.79	8.48	20.00	0.31	66.46	20.00	1.50	20.04
	LR	20.00	1.04	9.98	20.00	0.34	66.58	20.00	1.53	20.29
	MR	20.00	1.28	11.90	20.00	0.35	74.21	20.00	1.65	20.07
	HR	20.00	1.50	13.83	20.00	0.39	74.97	20.00	1.76	21.27
Patching	PM	100.00	2.55	9.44	100.00	0.81	65.89	100.00	0.51	69.04
	LR	100.00	2.86	9.78	100.00	0.88	77.37	100.00	0.57	70.04
	MR	100.00	3.41	11.48	100.00	1.01	80.31	100.00	0.58	78.29
	HR	100.00	3.62	11.18	100.00	1.04	89.06	100.00	0.65	79.75

Table A.9 Model coefficients for Zone 3-pavement family C

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.47	92.92	100.00	0.42	116.22	Not Enough Data		
	LR	100.00	0.49	110.43	100.00	0.49	133.09			
	MR	100.00	0.52	111.54	100.00	0.55	147.01			
	HR	100.00	0.58	116.91	100.00	0.59	155.54			
Deep Rutting	PM	100.00	0.60	89.57	100.00	0.97	40.67			
	LR	100.00	0.61	103.70	100.00	1.11	48.60			
	MR	100.00	0.63	120.53	100.00	1.19	54.91			
	HR	100.00	0.64	140.51	100.00	1.20	57.54			
Failures	PM	20.00	0.65	90.72	20.00	1.47	23.67			
	LR	20.00	0.67	105.15	20.00	1.52	25.05			
	MR	20.00	0.69	122.36	20.00	1.60	26.29			
	HR	20.00	0.71	142.82	20.00	1.69	27.52			
Block Cracking	PM	100.00	2.55	35.71	100.00	0.60	91.96			
	LR	100.00	2.69	38.13	100.00	0.62	107.20			
	MR	100.00	2.87	40.74	100.00	0.65	126.58			
	HR	100.00	3.08	44.03	100.00	0.69	127.51			
Alligator Cracking	PM	100.00	0.58	84.99	100.00	0.49	87.29			
	LR	100.00	0.69	95.36	100.00	0.49	99.18			
	MR	100.00	0.81	104.94	100.00	0.50	113.35			
	HR	100.00	0.94	114.17	100.00	0.50	129.52			
Longitudinal Cracking	PM	500.00	0.59	60.92	500.00	0.36	61.59			
	LR	500.00	0.63	70.20	500.00	0.39	72.32			
	MR	500.00	0.71	70.67	500.00	0.45	75.19			
	HR	500.00	0.84	75.60	500.00	0.46	83.94			
Transverse Cracking	PM	20.00	0.53	83.84	20.00	0.36	99.89			
	LR	20.00	0.63	93.93	20.00	0.39	103.99			
	MR	20.00	0.73	103.18	20.00	0.45	116.06			
	HR	20.00	0.84	111.02	20.00	0.46	137.82			
Patching	PM	100.00	0.51	104.29	100.00	0.37	86.61			
	LR	100.00	0.56	110.31	100.00	0.37	99.37			
	MR	100.00	0.66	126.22	100.00	0.38	113.92			
	HR	100.00	0.71	130.35	100.00	0.39	131.58			

Table A.10 Model coefficients for Zone 4-pavement family A

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.43	95.69	100.00	0.47	92.92	100.00	0.55	94.44
	LR	100.00	0.46	96.07	100.00	0.49	110.43	100.00	0.58	111.97
	MR	100.00	0.51	100.16	100.00	0.52	111.54	100.00	0.62	114.49
	HR	100.00	0.59	108.68	100.00	0.58	116.91	100.00	0.69	120.47
Deep Rutting	PM	100.00	0.57	116.22	100.00	0.86	54.62	100.00	0.88	60.35
	LR	100.00	0.74	126.95	100.00	0.88	60.15	100.00	0.92	65.17
	MR	100.00	0.78	151.64	100.00	0.91	66.05	100.00	0.94	69.94
	HR	100.00	0.98	166.84	100.00	0.94	73.03	100.00	0.94	74.61
Failures	PM	20.00	0.63	90.15	20.00	0.77	60.53	20.00	4.54	102.37
	LR	20.00	0.64	103.93	20.00	0.81	69.42	20.00	5.11	102.51
	MR	20.00	0.66	120.95	20.00	0.90	81.88	20.00	5.94	107.28
	HR	20.00	0.68	141.16	20.00	1.03	84.21	20.00	7.12	115.03
Block Cracking	PM	100.00	3.46	19.85	100.00	1.32	22.04	100.00	6.71	93.11
	LR	100.00	3.94	21.92	100.00	1.37	22.78	100.00	8.00	93.95
	MR	100.00	3.98	25.31	100.00	1.43	24.21	100.00	8.38	100.35
	HR	100.00	4.29	25.89	100.00	1.51	25.45	100.00	9.26	112.37
Alligator Cracking	PM	100.00	0.61	56.90	100.00	0.58	50.13	100.00	1.20	17.56
	LR	100.00	0.64	63.85	100.00	0.58	53.97	100.00	1.22	18.31
	MR	100.00	0.69	74.32	100.00	0.59	58.41	100.00	1.23	18.18
	HR	100.00	0.77	88.47	100.00	0.60	62.86	100.00	1.33	18.05
Longitudinal Cracking	PM	500.00	0.47	70.47	500.00	0.33	93.68	500.00	0.47	86.52
	LR	500.00	0.53	71.91	500.00	0.34	112.74	500.00	0.47	98.81
	MR	500.00	0.55	81.85	500.00	0.38	117.63	500.00	0.47	112.61
	HR	500.00	0.63	86.34	500.00	0.43	128.00	500.00	0.47	128.33
Transverse Cracking	PM	20.00	1.37	20.99	20.00	0.79	34.85	20.00	1.20	27.59
	LR	20.00	1.41	21.59	20.00	0.85	41.99	20.00	1.28	30.55
	MR	20.00	1.45	22.62	20.00	1.43	47.89	20.00	1.43	34.93
	HR	20.00	1.49	23.81	20.00	1.89	52.75	20.00	1.66	41.19
Patching	PM	100.00	0.34	61.02	100.00	0.35	78.21	100.00	0.53	61.40
	LR	100.00	0.36	71.17	100.00	0.41	85.05	100.00	0.57	70.64
	MR	100.00	0.42	73.91	100.00	0.46	89.16	100.00	0.64	72.63
	HR	100.00	0.43	82.53	100.00	0.51	89.88	100.00	0.77	79.45

Table A.11 Model coefficients for Zone 4-pavement family B

Distress Type	Treatment Type	Low Traffic			Medium Traffic			High Traffic		
		α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.36	96.83	100.00	0.42	97.79	Not Enough Data		
	LR	100.00	0.38	98.63	100.00	0.45	100.03			
	MR	100.00	0.43	106.47	100.00	0.50	107.93			
	HR	100.00	0.51	119.87	100.00	0.60	123.21			
Deep Rutting	PM	100.00	0.60	101.99	100.00	0.66	79.17			
	LR	100.00	0.65	105.86	100.00	0.83	87.75			
	MR	100.00	0.75	117.40	100.00	0.94	96.08			
	HR	100.00	0.77	137.53	100.00	0.98	102.67			
Failures	PM	20.00	9.95	99.32	20.00	1.11	23.48			
	LR	20.00	10.88	104.94	20.00	1.17	24.55			
	MR	20.00	12.95	121.25	20.00	1.26	27.30			
	HR	20.00	14.05	127.01	20.00	1.40	30.05			
Block Cracking	PM	100.00	9.31	97.21	100.00	0.71	46.78			
	LR	100.00	10.33	104.85	100.00	0.84	49.16			
	MR	100.00	10.64	105.51	100.00	0.98	50.18			
	HR	100.00	12.27	119.35	100.00	1.12	59.95			
Alligator Cracking	PM	100.00	0.65	47.36	100.00	0.58	49.93			
	LR	100.00	0.93	54.93	100.00	0.59	53.87			
	MR	100.00	1.03	59.34	100.00	0.60	58.23			
	HR	100.00	1.38	69.90	100.00	0.60	62.57			
Longitudinal Cracking	PM	500.00	0.46	45.45	500.00	0.29	107.15			
	LR	500.00	0.55	47.42	500.00	0.32	116.81			
	MR	500.00	0.65	48.67	500.00	0.33	118.06			
	HR	500.00	0.75	50.03	500.00	0.38	133.00			
Transverse Cracking	PM	20.00	0.44	114.98	20.00	0.42	70.47			
	LR	20.00	0.50	130.98	20.00	0.54	80.58			
	MR	20.00	0.56	142.26	20.00	0.75	96.06			
	HR	20.00	0.59	146.80	20.00	1.05	98.42			
Patching	PM	100.00	0.34	89.86	100.00	0.88	10.30			
	LR	100.00	0.35	105.32	100.00	1.05	11.84			
	MR	100.00	0.36	125.24	100.00	1.23	13.58			
	HR	100.00	0.39	126.92	100.00	1.40	15.99			

Table A.12 Model coefficients for Zone 4-pavement family C

Distress	Treatment	Low Traffic			Medium Traffic			High Traffic		
Type	Type	α	β	A	α	β	A	α	β	A
Shallow Rutting	PM	100.00	0.42	97.79	100.00	0.39	56.14	Not Enough Data		
	LR	100.00	0.45	100.03	100.00	0.41	63.30			
	MR	100.00	0.50	107.93	100.00	0.44	74.46			
	HR	100.00	0.60	123.21	100.00	0.51	75.50			
Deep Rutting	PM	100.00	0.68	97.50	100.00	0.49	116.13			
	LR	100.00	0.72	98.24	100.00	0.56	132.55			
	MR	100.00	0.79	102.39	100.00	0.63	145.77			
	HR	100.00	0.90	110.55	100.00	0.67	152.57			
Failures	PM	20.00	0.59	70.95	20.00	0.42	100.84			
	LR	20.00	0.67	72.62	20.00	0.83	113.32			
	MR	20.00	0.68	82.87	20.00	1.00	122.16			
	HR	20.00	0.78	86.41	20.00	2.16	126.34			
Block Cracking	PM	100.00	3.39	84.04	100.00	0.33	114.03			
	LR	100.00	3.65	96.43	100.00	0.50	120.85			
	MR	100.00	3.96	110.73	100.00	0.81	136.54			
	HR	100.00	4.34	129.03	100.00	0.90	163.92			
Alligator Cracking	PM	100.00	0.65	55.86	100.00	0.28	111.35			
	LR	100.00	0.68	62.31	100.00	0.32	124.55			
	MR	100.00	0.72	70.46	100.00	0.34	132.67			
	HR	100.00	0.79	81.70	100.00	0.35	159.94			
Longitudinal Cracking	PM	500.00	0.48	82.69	500.00	0.29	72.09			
	LR	500.00	0.57	92.51	500.00	0.33	75.44			
	MR	500.00	0.66	101.42	500.00	0.36	88.61			
	HR	500.00	0.75	107.54	500.00	0.36	99.25			
Transverse Cracking	PM	20.00	5.71	16.32	20.00	1.76	23.96			
	LR	20.00	5.89	18.04	20.00	1.81	24.66			
	MR	20.00	6.75	19.05	20.00	1.85	25.50			
	HR	20.00	7.18	22.38	20.00	1.87	26.34			
Patching	PM	100.00	0.39	71.42	100.00	0.32	40.19			
	LR	100.00	0.46	76.60	100.00	0.37	41.03			
	MR	100.00	1.05	80.33	100.00	0.42	48.60			
	HR	100.00	1.62	80.71	100.00	0.47	56.73			

APPENDIX B

CODE OF COMPUTER PROGRAM DEVELOPED FOR CALCULATING REMAINING SERVICE LIFE

```
## LAST TREATMENT

# Data for the year of 2014

        PMIS_CONDITION_SUMMARY_14 < - read.csv ("C:/RR-
        PMIS/PMIS_CONDITION_SUMMARY_14.txt");

        PMIS_CONDITION_SUMMARY_14$Unique = paste
        (PMIS_CONDITION_SUMMARY_14$SIGNED_HIGHWAY_RDBD_ID,PMIS_CONDITION
        _SUMMARY_14$BEG_REF_MARKER_NBR,format(PMIS_CONDITION_SUMMARY_14$
        BEG_REF_MARKER_DISP,digit=2),sep ="");

        PMIS_CONDITION_SUMMARY_14 =
        PMIS_CONDITION_SUMMARY_14[PMIS_CONDITION_SUMMARY_14$RATING_CYCL
        E_CODE=='P',]

        PMIS_DATA_COLLECTION_SECTION_14 <- read.csv("C:/RR-
        PMIS/PMIS_DATA_COLLECTION_SECTION_14.txt");

        PMIS_DATA_COLLECTION_SECTION_14$Unique = paste
        (PMIS_DATA_COLLECTION_SECTION_14$SIGNED_HIGHWAY_RDBD_ID,PMIS_DATA
        _COLLECTION_SECTION_14$BEG_REF_MARKER_NBR,format(PMIS_DATA_COLLECT
        ION_SECTION_14$BEG_REF_MARKER_DISP,digit=2),sep ="");

        PMIS_COLLECTION_14= PMIS_DATA_COLLECTION_SECTION_14 [, c("Unique",
        "COUNTY_NBR","PVMNT_TYPE_DTL_RD_LIFE_CODE","CURRENT_18KIP_MEAS","A
        ADT_CURRENT_YEAR")];

        PMIS_JOIN_COLLECTION_SUMMARY_14= merge
        (PMIS_CONDITION_SUMMARY_14,PMIS_COLLECTION_14, "Unique");

        Condition_Sum_14=
        PMIS_JOIN_COLLECTION_SUMMARY_14[PMIS_JOIN_COLLECTION_SUMMARY_14$
        CONDITION_SCORE>4 &
        PMIS_JOIN_COLLECTION_SUMMARY_14$PVMNT_TYPE_DTL_RD_LIFE_CODE>3,];

        names (Condition_Sum_14)[names(Condition_Sum_14)=="CONDITION_SCORE"]<-"CS_14"

        Condition_Sum_14 =
        Condition_Sum_14[,c(setdiff(names(Condition_Sum_14),"CS_14"),"CS_14")]

# Data for the year of 2013

        PMIS_CONDITION_SUMMARY_13 <- read.csv("C:/RR-
        PMIS/PMIS_CONDITION_SUMMARY_13.txt");
```

```
    Condition_Sum_13=
    PMIS_CONDITION_SUMMARY_13[PMIS_CONDITION_SUMMARY_13$RATING_CYCL
    E_CODE=='P',];

    Condition_Sum_13$Unique = paste
    (Condition_Sum_13$SIGNED_HIGHWAY_RDBD_ID,Condition_Sum_13$BEG_REF_MARK
    ER_NBR,format(Condition_Sum_13$BEG_REF_MARKER_DISP,digit=2),sep ="");

    Distress_13 = Condition_Sum_13[,c("CONDITION_SCORE","Unique")];

    names(Distress_13)[names(Distress_13)=="CONDITION_SCORE"]<-"CS_13"

# Data for the year of 2012

    PMIS_CONDITION_SUMMARY_12 <- read.csv("C:/RR-
    PMIS/PMIS_CONDITION_SUMMARY_12.txt");

    Condition_Sum_12=
    PMIS_CONDITION_SUMMARY_12[PMIS_CONDITION_SUMMARY_12$RATING_CYCL
    E_CODE=='P',];

    Condition_Sum_12$Unique = paste
    (Condition_Sum_12$SIGNED_HIGHWAY_RDBD_ID,Condition_Sum_12$BEG_REF_MARK
    ER_NBR,format(Condition_Sum_12$BEG_REF_MARKER_DISP,digit=2),sep ="")

    Distress_12 = Condition_Sum_12[,c("CONDITION_SCORE","Unique")];

    names(Distress_12)[names(Distress_12)=="CONDITION_SCORE"]<-"CS_12"

# Data for the year of 2011

    PMIS_CONDITION_SUMMARY_11 <- read.csv("C:/RR-
    PMIS/PMIS_CONDITION_SUMMARY_11.txt")

    Condition_Sum_11=
    PMIS_CONDITION_SUMMARY_11[PMIS_CONDITION_SUMMARY_11$RATING_CYCL
    E_CODE=='P',];

    Condition_Sum_11$Unique = paste
    (Condition_Sum_11$SIGNED_HIGHWAY_RDBD_ID,Condition_Sum_11$BEG_REF_MARK
    ER_NBR,format(Condition_Sum_11$BEG_REF_MARKER_DISP,digit=2),sep ="")

    Distress_11 = Condition_Sum_11[,c("CONDITION_SCORE","Unique")];

    names(Distress_11)[names(Distress_11)=="CONDITION_SCORE"] <- "CS_11"

# Data for the year of 2010

    PMIS_CONDITION_SUMMARY_10 <- read.csv("C:/RR-
    PMIS/PMIS_CONDITION_SUMMARY_10.txt")

    Condition_Sum_10=
    PMIS_CONDITION_SUMMARY_10[PMIS_CONDITION_SUMMARY_10$RATING_CYCL
    E_CODE=='P',];
```

```
        Condition_Sum_10$Unique = paste(Condition_Sum_10$SIGNED_HIGHWAY_RDBD_ID,Condition_Sum_10$BEG_REF_MARKER_NBR,format(Condition_Sum_10$BEG_REF_MARKER_DISP,digit=2),sep ="")

        Distress_10 = Condition_Sum_10[,c("CONDITION_SCORE","Unique")];

        names(Distress_10)[names(Distress_10)=="CONDITION_SCORE"]<-"CS_10"

# Data for the year of 2009

        PMIS_CONDITION_SUMMARY_09 <- read.csv("C:/RR-PMIS/PMIS_CONDITION_SUMMARY_09.txt")

        Condition_Sum_09= PMIS_CONDITION_SUMMARY_09[PMIS_CONDITION_SUMMARY_09$RATING_CYCLE_CODE=='P',];

        Condition_Sum_09$Unique = paste(Condition_Sum_09$SIGNED_HIGHWAY_RDBD_ID,Condition_Sum_09$BEG_REF_MARKER_NBR,format(Condition_Sum_09$BEG_REF_MARKER_DISP,digit=2),sep ="")

        Distress_09 = Condition_Sum_09[,c("CONDITION_SCORE","Unique")];

        names(Distress_09)[names(Distress_09)=="CONDITION_SCORE"]<-"CS_09"

# Data for the year of 2008

        PMIS_CONDITION_SUMMARY_08 <- read.csv("C:/RR-PMIS/PMIS_CONDITION_SUMMARY_08.txt")

        Condition_Sum_08= PMIS_CONDITION_SUMMARY_08[PMIS_CONDITION_SUMMARY_08$RATING_CYCLE_CODE=='P',];

        #Condition_Sum_08= Condition_Sum_08[Condition_Sum_08$DISTRESS_SCORE>10,]

        Condition_Sum_08$Unique = paste(Condition_Sum_08$SIGNED_HIGHWAY_RDBD_ID,Condition_Sum_08$BEG_REF_MARKER_NBR,format(Condition_Sum_08$BEG_REF_MARKER_DISP,digit=2),sep ="")

        Distress_08 = Condition_Sum_08[,c("CONDITION_SCORE","Unique")];

        names(Distress_08)[names(Distress_08)=="CONDITION_SCORE"]<-"CS_08"



# Merge CS of different years together

        CS14_13 = merge(Condition_Sum_14,Distress_13,by.x="Unique",by.y="Unique", all.x=TRUE);

        CS14_12 = merge(CS14_13,Distress_12, by.x="Unique",by.y="Unique", all.x=TRUE);

        CS14_11 = merge(CS14_12,Distress_11, by.x="Unique",by.y="Unique", all.x=TRUE);

        CS14_10 = merge(CS14_11,Distress_10, by.x="Unique",by.y="Unique", all.x=TRUE);
```

```
              CS14_09 = merge(CS14_10,Distress_09, by.x="Unique",by.y="Unique", all.x=TRUE);

              CS14_08 = merge(CS14_09,Distress_08, by.x="Unique",by.y="Unique", all.x=TRUE);

# Last treatment and Year of treatment on CS14_08

For (i in 1:nrow(CS14_08)){

  CS14_08$Last_TR[i]= "HR"

  CS14_08$Year_TR[i]=2008

    if (!is.na(CS14_08$CS_09[i]) && !is.na(CS14_08$CS_08[i]) ) {

                  if (CS14_08$CS_09[i]>4 && CS14_08$CS_08[i] >4){

                                  if (CS14_08$CS_09[i] - CS14_08$CS_08[i] >40){

                                    CS14_08$Last_TR[i]= "HR"
                                    CS14_08$Year_TR[i]=2009
                                  }else if(CS14_08$CS_09[i] - CS14_08$CS_08[i] >30){

                                    CS14_08$Last_TR [i]= "MR"
                                    CS14_08$Year_TR[i]=2009
                                  }else if(CS14_08$CS_09[i] - CS14_08$CS_08[i] >20){

                                    CS14_08$Last_TR[i]= "LR"
                                    CS14_08$Year_TR[i]=2009
                                  }else if(CS14_08$CS_09[i] - CS14_08$CS_08[i] >=5){

                                    CS14_08$Last_TR[i]= "PM"
                                    CS14_08$Year_TR[i]=2009
                                  }

                  }

  }



  if (!is.na(CS14_08$CS_10[i]) && !is.na(CS14_08$CS_09[i]) ) {

                  if (CS14_08$CS_10[i]>4 && CS14_08$CS_09[i] >4){

                                  if (CS14_08$CS_10[i] - CS14_08$CS_09[i] >40){

                                    CS14_08$Last_TR[i]= "HR"
                                    CS14_08$Year_TR[i]=2010
                                  }else if(CS14_08$CS_10[i] - CS14_08$CS_09[i] >30){

                                    CS14_08$Last_TR [i]= "MR"
                                    CS14_08$Year_TR[i]=2010
                                  }else if(CS14_08$CS_10[i] - CS14_08$CS_09[i] >20){

                                    CS14_08$Last_TR[i]= "LR"
```

```
                    CS14_08$Year_TR[i]=2010
                  }else if(CS14_08$CS_10[i] - CS14_08$CS_09[i] >=5){

                    CS14_08$Last_TR[i]= "PM"
                    CS14_08$Year_TR[i]=2010
                  }

                 } else if (CS14_08$CS_10[i]>4 && !is.na(CS14_08$CS_08[i]) &&
               CS14_08$CS_08[i] >4 && CS14_08$CS_09[i]==0){

                  if (CS14_08$CS_10[i] - CS14_08$CS_08[i] >40){

                    CS14_08$Last_TR[i]= "HR"
                    CS14_08$Year_TR[i]=2009
                  }else if(CS14_08$CS_10[i] - CS14_08$CS_08[i] >30){

                    CS14_08$Last_TR [i]= "MR"
                    CS14_08$Year_TR[i]=2009
                  }else if(CS14_08$CS_10[i] - CS14_08$CS_08[i] >20){

                    CS14_08$Last_TR[i]= "LR"
                    CS14_08$Year_TR[i]=2009
                  }else if(CS14_08$CS_10[i] - CS14_08$CS_08[i] >=5){

                    CS14_08$Last_TR[i]= "PM"
                    CS14_08$Year_TR[i]=2009
                  }

         }

}

if (!is.na(CS14_08$CS_11[i]) && !is.na(CS14_08$CS_10[i]) ) {

         if (CS14_08$CS_11[i]>4 && CS14_08$CS_10[i] >4){

                  if (CS14_08$CS_11[i] - CS14_08$CS_10[i] >40){

                    CS14_08$Last_TR[i]= "HR"
                    CS14_08$Year_TR[i]=2011
                  }else if(CS14_08$CS_11[i] - CS14_08$CS_10[i] >30){

                    CS14_08$Last_TR [i]= "MR"
                    CS14_08$Year_TR[i]=2011
                  }else if(CS14_08$CS_11[i] - CS14_08$CS_10[i] >20){

                    CS14_08$Last_TR[i]= "LR"
                    CS14_08$Year_TR[i]=2011
                  }else if(CS14_08$CS_11[i] - CS14_08$CS_10[i] >=5){

                    CS14_08$Last_TR[i]= "PM"
                    CS14_08$Year_TR[i]=2011
                  }
               } else if (CS14_08$CS_11[i]>4 && !is.na(CS14_08$CS_09[i]) &&
             CS14_08$CS_09[i] >4 && CS14_08$CS_10[i] == 0){

                if (CS14_08$CS_11[i] - CS14_08$CS_09[i] >40){

                 CS14_08$Last_TR[i]= "HR"
                 CS14_08$Year_TR[i]=2010
                }else if(CS14_08$CS_11[i] - CS14_08$CS_09[i] >30){

                 CS14_08$Last_TR [i]= "MR"
                 CS14_08$Year_TR[i]=2010
                }else if(CS14_08$CS_11[i] - CS14_08$CS_09[i] >20){

                 CS14_08$Last_TR[i]= "LR"
                 CS14_08$Year_TR[i]=2010
                }else if(CS14_08$CS_11[i] - CS14_08$CS_09[i] >=5){

                 CS14_08$Last_TR[i]= "PM"
                 CS14_08$Year_TR[i]=2010
                }

        }

}

 if (!is.na(CS14_08$CS_12[i]) && !is.na(CS14_08$CS_11[i]) ) {

        if (CS14_08$CS_12[i]>4 && CS14_08$CS_11[i] >4){

                if (CS14_08$CS_12[i] - CS14_08$CS_11[i] >40){

                 CS14_08$Last_TR[i]= "HR"
                 CS14_08$Year_TR[i]=2012
                }else if(CS14_08$CS_12[i] - CS14_08$CS_11[i] >30){

                 CS14_08$Last_TR [i]= "MR"
                 CS14_08$Year_TR[i]=2012
                }else if(CS14_08$CS_12[i] - CS14_08$CS_11[i] >20){

                 CS14_08$Last_TR[i]= "LR"
                 CS14_08$Year_TR[i]=2012
                }else if(CS14_08$CS_12[i] - CS14_08$CS_11[i] >=5){

                 CS14_08$Last_TR[i]= "PM"
                 CS14_08$Year_TR[i]=2012
                }

               } else if (CS14_08$CS_12[i]>4 && !is.na(CS14_08$CS_10[i])&&
             CS14_08$CS_10[i] >4 && CS14_08$CS_11[i]== 0){

                if (CS14_08$CS_12[i] - CS14_08$CS_10[i] >40){

                 CS14_08$Last_TR[i]= "HR"
                 CS14_08$Year_TR[i]=2011
                }else if(CS14_08$CS_12[i] - CS14_08$CS_10[i] >30){
                      CS14_08$Last_TR [i]= "MR"
                      CS14_08$Year_TR[i]=2011
                     }else if(CS14_08$CS_12[i] - CS14_08$CS_10[i] >20){

                      CS14_08$Last_TR[i]= "LR"
                      CS14_08$Year_TR[i]=2011
                     }else if(CS14_08$CS_12[i] - CS14_08$CS_10[i] >=5){

                      CS14_08$Last_TR[i]= "PM"
                      CS14_08$Year_TR[i]=2011
                     }

          }

}

 if (!is.na(CS14_08$CS_13[i]) && !is.na(CS14_08$CS_12[i]) ) {

          if (CS14_08$CS_13[i]>4 && CS14_08$CS_12[i] >4){

                     if (CS14_08$CS_13[i] - CS14_08$CS_12[i] >40){

                      CS14_08$Last_TR[i]= "HR"
                      CS14_08$Year_TR[i]=2013
                     }else if(CS14_08$CS_13[i] - CS14_08$CS_12[i] >30){

                      CS14_08$Last_TR [i]= "MR"
                      CS14_08$Year_TR[i]=2013
                     }else if(CS14_08$CS_13[i] - CS14_08$CS_12[i] >20){

                      CS14_08$Last_TR[i]= "LR"
                      CS14_08$Year_TR[i]=2013
                     }else if(CS14_08$CS_13[i] - CS14_08$CS_12[i] >= 5){

                      CS14_08$Last_TR[i]= "PM"
                      CS14_08$Year_TR[i]=2013
                     }

                    }else if (CS14_08$CS_13[i]>4 && !is.na(CS14_08$CS_11[i])&&
                 CS14_08$CS_11[i] >4 && CS14_08$CS_12[i] ==0){

                     if (CS14_08$CS_13[i] - CS14_08$CS_11[i] >40){

                      CS14_08$Last_TR[i]= "HR"
                      CS14_08$Year_TR[i]=2012
                     }else if(CS14_08$CS_13[i] - CS14_08$CS_11[i] >30){

                      CS14_08$Last_TR [i]= "MR"
                      CS14_08$Year_TR[i]=2012
                     }else if(CS14_08$CS_13[i] - CS14_08$CS_11[i] >20){

                      CS14_08$Last_TR[i]= "LR"
                      CS14_08$Year_TR[i]=2012
                     }else if(CS14_08$CS_13[i] - CS14_08$CS_11[i] >= 5){
```

```
                    CS14_08$Last_TR[i]= "PM"
                    CS14_08$Year_TR[i]=2012
                   }

       }

}

 if (!is.na(CS14_08$CS_14[i]) && !is.na(CS14_08$CS_13[i]) ) {

         if (CS14_08$CS_14[i]>4 && CS14_08$CS_13[i] >4){

                   if (CS14_08$CS_14[i] - CS14_08$CS_13[i] >40){

                            CS14_08$Last_TR[i]= "HR"
                            CS14_08$Year_TR[i]=2014
                           }else if(CS14_08$CS_14[i] - CS14_08$CS_13[i] >30){

                            CS14_08$Last_TR [i]= "MR"
                            CS14_08$Year_TR[i]=2014
                           }else if(CS14_08$CS_14[i] - CS14_08$CS_13[i] >20){

                            CS14_08$Last_TR[i]= "LR"
                            CS14_08$Year_TR[i]=2014
                           }else if(CS14_08$CS_14[i] - CS14_08$CS_13[i] >= 5){

                            CS14_08$Last_TR[i]= "PM"
                            CS14_08$Year_TR[i]=2014
                   }

                  }else if (CS14_08$CS_14[i]>4 && !is.na(CS14_08$CS_12[i]) &&
               CS14_08$CS_12[i] >4 && CS14_08$CS_13[i] ==0){

                           if (CS14_08$CS_14[i] - CS14_08$CS_12[i] >40){

                            CS14_08$Last_TR[i]= "HR"
                            CS14_08$Year_TR[i]=2013
                           }else if(CS14_08$CS_14[i] - CS14_08$CS_12[i] >30){

                            CS14_08$Last_TR [i]= "MR"
                            CS14_08$Year_TR[i]=2013
                           }else if(CS14_08$CS_14[i] - CS14_08$CS_12[i] >20){

                            CS14_08$Last_TR[i]= "LR"
                            CS14_08$Year_TR[i]=2013
                           }else if(CS14_08$CS_14[i] - CS14_08$CS_12[i] >=5){

                            CS14_08$Last_TR[i]= "PM"
                            CS14_08$Year_TR[i]=2013
                   }

                  }

        }
```

```
}

# Define climate zone

ClimateZone <- read.csv("C:/RR-PMIS/ClimateZone.csv");

ClimateZone = ClimateZone[,c(1,2)]

CS14_08= merge(CS14_08,ClimateZone, by.x = "COUNTY_NBR",by.y="CountyNum", all.x=TRUE);



# Define pavement families

        PvFamily <- read.csv("C:/RR-PMIS/PvFamily.csv");

        CS14_08= merge(CS14_08,PvFamily, by.x =
        "PVMNT_TYPE_DTL_RD_LIFE_CODE",by.y="PVMNT_TYPE_DTL_RD_LIFE_CODE",
        all.x=TRUE);

# Define loading type

        CS14_08$Loading =
        ifelse(CS14_08$CURRENT_18KIP_MEAS<1000,'Low',ifelse(CS14_08$CURRENT_18KIP_M
        EAS<10000,'Medium','High'));

#Write the output

        write.csv(CS14_08,"C:/RR-PMIS/CS14_08.txt");

## RSL ESTIMATION

# Bring Coeficient table and merge it to the current CS14_08 and make TX_14

        CS14_08 <- read.csv("C:/RR-PMIS/CS14_08.txt");

        Coeficients <- read.csv("C:/RR-PMIS/Coeficients.csv")

        CS14_08$Unique_Coef =
        paste(CS14_08$ClimateZone,CS14_08$PvFamily,CS14_08$Loading,CS14_08$Last_TR,sep
        ="");

        Coeficients$Unique_Coef =
        paste(Coeficients$ClimateZone,Coeficients$PvFamily,Coeficients$Loading,Coeficients$Last_T
        R,sep ="");

        Coeficients =
        Coeficients[,c("AC1","AC2","AC3","LC1","LC2","LC3","TC1","TC2","TC3","ShRUT1","ShR
        UT2","ShRUT3","DRUT1","DRUT2","DRUT3","Fail1","Fail2","Fail3","BC1","BC2","BC3","P
        atch1","Patch2","Patch3","DS2","DS3","Unique_Coef" )];

        TX_14= merge (CS14_08,Coeficients,by.x="Unique_Coef",by.y="Unique_Coef");

        TX_14 =
        TX_14[complete.cases(TX_14[,c("ACP_ALLIGATOR_CRACKING_PCT","ACP_LONGITUD
        E_CRACKING_PCT","ACP_TRANSVERSE_CRACKING_QTY",
        "ACP_RUT_AUTO_SHALLOW_AVG_PCT","ACP_RUT_AUTO_DEEP_AVG_PCT","ACP_
        FAILURE_QTY","ACP_BLOCK_CRACKING_PCT","ACP_PATCHING_PCT",
        "DISTRESS_SCORE")]),];

# Drop some unimportant columns

        drops =
        c("AC_Cur_Age","ACP_RUT_VISUAL_DEEP_PCT","ACP_RUT_VISUAL_SEVERE_PCT","
        PCC_AVG_CRACK_SPACING_AUTO_QTY","AC_CurAge","SSI_SCORE","SCI_ADJ","SSI
        _DEFLECT_7_ADJ","CRCP_PCC_PATCHES_QTY","CRCP_AVG_CRACK_SPACING_QT
        Y","PCC_PUNCHOUT_AUTO_SMRY_QTY","PCC_CRCK_OUT_WP_AVG_AUTO_PCT",
        "PCC_CRCK_RWP_AVG_AUTO_PCT","PCC_CRCK_BET_WP_AVG_AUTO_PCT",
        "PCC_SPALLED_CRACKS_AUTO_SMRY_QTY","PCC_LONG_CRACKS_AUTO_SMRY_
        MEAS", "JCP_SHATTERED_SLABS_QTY", "JCP_APPARENT_JNT_SPACE_MEAS",
        "JCP_FAILURES_QTY","JCP_PCC_PATCHES_QTY",
        "JCP_FAILED_JNTS_CRACKS_QTY","JCP_LONGITUDE_CRACKS_QTY",
        "PCC_CRCK_LWP_AVG_AUTO_PCT", "ACP_CRCK_OUT_WP_AVG_AUTO_PCT",
        "ACP_CRCK_BET_WP_AVG_AUTO_PCT", "ACP_CRCK_LWP_AVG_AUTO_PCT",
        "ACP_LONG_CRACKS_AUTO_SMRY_MEAS", "ACP_RUT_VISUAL_SHALLOW_PCT",
        "ACP_RUT_VISUAL_FAILURE_PCT", "CRCP_SPALLED_CRACKS_QTY",
        "CRCP_ACP_PATCHES_QTY", "CRCP_PUNCHOUT_QTY", "TEXTURE_RIGHT_SCORE",
        "TEXTURE_LEFT_SCORE","ACP_ALLIG_CRACKS_AUTO_SMRY_PCT",
        "ACP_CRCK_RWP_AVG_AUTO_PCT", "ACP_TRANS_CRACKS_AUTO_SMRY_QTY");

        TX_14 = TX_14[,!(names(TX_14) %in% drops)];

# Calculate Current Age based on each indicator

TX_14$AC_CurAge =
ifelse(TX_14$ACP_ALLIGATOR_CRACKING_PCT==100,(TX_14$AC3)*((log(TX_14$AC1/99))^(-
1/(TX_14$AC2))),(TX_14$AC3)*((log(TX_14$AC1/TX_14$ACP_ALLIGATOR_CRACKING_PCT))^(
-1/(TX_14$AC2))));

TX_14$LC_CurAge = ifelse(TX_14$ACP_LONGITUDE_CRACKING_PCT>400,
(TX_14$LC3)*((log(TX_14$LC1/400))^(-
1/(TX_14$LC2))),(TX_14$LC3)*((log(TX_14$LC1/TX_14$ACP_LONGITUDE_CRACKING_PCT))^(-
1/(TX_14$LC2))));

TX_14$TC_CurAge = ifelse(TX_14$ACP_TRANSVERSE_CRACKING_QTY>19,
(TX_14$TC3)*((log(TX_14$TC1/19))^(-
1/(TX_14$TC2))),(TX_14$TC3)*((log(TX_14$TC1/TX_14$ACP_TRANSVERSE_CRACKING_QTY))
^(-1/(TX_14$TC2))));

TX_14$ShRUT_CurAge =
(TX_14$ShRUT3)*((log(TX_14$ShRUT1/TX_14$ACP_RUT_AUTO_SHALLOW_AVG_PCT))^(-
1/(TX_14$ShRUT2)));

TX_14$DRUT_CurAge =
(TX_14$DRUT3)*((log(TX_14$DRUT1/TX_14$ACP_RUT_AUTO_DEEP_AVG_PCT))^(-
1/(TX_14$DRUT2)));

TX_14$Failure_CurAge = ifelse(TX_14$ACP_FAILURE_QTY>19,
(TX_14$Fail3)*((log(TX_14$Fail1/19))^(-
1/(TX_14$Fail2))),(TX_14$Fail3)*((log(TX_14$Fail1/TX_14$ACP_FAILURE_QTY))^(-
1/(TX_14$Fail2))));

TX_14$BC_CurAge = ifelse(TX_14$ACP_BLOCK_CRACKING_PCT>99,
(TX_14$BC3)*((log(TX_14$BC1/99))^(-
1/(TX_14$BC2))),(TX_14$BC3)*((log(TX_14$BC1/TX_14$ACP_BLOCK_CRACKING_PCT))^(-
1/(TX_14$BC2))));

TX_14$Pathc_CurAge = ifelse(TX_14$ACP_PATCHING_PCT>99,
(TX_14$Patch3)*((log(TX_14$Patch1/99))^(-1/(TX_14$Patch2))),
(TX_14$Patch3)*((log(TX_14$Patch1/TX_14$ACP_PATCHING_PCT))^(-1/(TX_14$Patch2))));

TX_14$DS_CurAge = (TX_14$DS2)*((log(100/(100-TX_14$DISTRESS_SCORE)))^(-
1/(TX_14$DS3)));

# Define AADT groups

        TX_14$AADT_Group =
        ifelse(TX_14$AADT_CURRENT<100,1,ifelse(TX_14$AADT_CURRENT<1000,2,ifelse(TX_1
        4$AADT_CURRENT<5000,3,4)));

# Read Threshold Matrix and add it to the TX

        Thresholds <- read.csv("C:/RR-PMIS/Thresholds.csv")

        TX_14= merge (TX_14,Thresholds,"AADT_Group");

#Calculate the age for Thresholds

TX_14$AC_Thresh_Age = (TX_14$AC3)*((log(TX_14$AC1/TX_14$AC_Threshold))^(-
1/(TX_14$AC2)));

TX_14$LC_Thresh_Age = (TX_14$LC3)*((log(TX_14$LC1/TX_14$LC_Threshold))^(-
1/(TX_14$LC2)));

TX_14$TC_Thresh_Age = (TX_14$TC3)*((log(TX_14$TC1/TX_14$TC_Threshold))^(-
1/(TX_14$TC2)));

TX_14$ShRUT_Thresh_Age =
(TX_14$ShRUT3)*((log(TX_14$ShRUT1/TX_14$ShRUT_Threshold))^(-1/(TX_14$ShRUT2)));

TX_14$DRUT_Thresh_Age = (TX_14$DRUT3)*((log(TX_14$DRUT1/TX_14$DRUT_Threshold))^(-
1/(TX_14$DRUT2)));

TX_14$Failure_Threshold_Age = (TX_14$Fail3)*((log(TX_14$Fail1/TX_14$Failure_Threshold))^(-
1/(TX_14$Fail2)));

TX_14$BC_Threshold_Age = (TX_14$BC3)*((log(TX_14$BC1/TX_14$BC_Threshold))^(-
1/(TX_14$BC2)));

TX_14$Pathc_Threshold_Age = (TX_14$Patch3)*((log(TX_14$Patch1/TX_14$Patch_Threshold))^(-
1/(TX_14$Patch2)));

# TX_14$RS_Threshold_Age = (TX_14$RS3)*((log(TX_14$RS1/TX_14$RS_Threshold))^(-
1/(TX_14$RS2)));
```

```
TX_14$DS_Threshold_Age = (TX_14$DS2)*((log(100/(100-TX_14$DS_Threshold)))^(-
1/(TX_14$DS3)));

# Calculate RSL for each indicators

        TX_14$AC_RSL = pmax(TX_14$AC_Thresh_Age - TX_14$AC_CurAge,0);

        TX_14$LC_RSL = pmax(TX_14$LC_Thresh_Age - TX_14$LC_CurAge,0);

        TX_14$TC_RSL = pmax(TX_14$TC_Thresh_Age - TX_14$TC_CurAge,0);

        TX_14$ShRUT_RSL = pmax(TX_14$ShRUT_Thresh_Age - TX_14$ShRUT_CurAge,0);

        TX_14$DRUT_RSL = pmax(TX_14$DRUT_Thresh_Age - TX_14$DRUT_CurAge,0);

        TX_14$Failure_RSL = pmax(TX_14$Failure_Threshold_Age - TX_14$Failure_CurAge,0);

        TX_14$BC_RSL = pmax(TX_14$BC_Threshold_Age - TX_14$BC_CurAge,0);

        TX_14$Pathc_RSL = pmax(TX_14$Pathc_Threshold_Age - TX_14$Pathc_CurAge,0);

        TX_14$DS_RSL = pmax(TX_14$DS_Threshold_Age - TX_14$DS_CurAge,0);

# Calculate the minimum RSL

TX_14$Section_RSL = pmin(TX_14$AC_RSL, TX_14$LC_RSL, TX_14$TC_RSL,
TX_14$ShRUT_RSL, TX_14$DRUT_RSL,TX_14$Failure_RSL, TX_14$BC_RSL, TX_14$Pathc_RSL,
TX_14$DS_RSL );

TX_14$Section_RSL_Indicator =
ifelse(TX_14$Section_RSL==TX_14$AC_RSL,"AC",ifelse(TX_14$Section_RSL==TX_14$LC_RSL,"L
C",ifelse(TX_14$Section_RSL==TX_14$TC_RSL,"TC",ifelse(TX_14$Section_RSL==TX_14$ShRUT_
RSL,"ShRUT",ifelse(TX_14$Section_RSL==TX_14$DRUT_RSL,"DRUT",ifelse(TX_14$Section_RSL=
=TX_14$Failure_RSL,"FAIL",ifelse(TX_14$Section_RSL==TX_14$BC_RSL,"BC",ifelse(TX_14$Secti
on_RSL==TX_14$Pathc_RSL,"Patch",ifelse(TX_14$Section_RSL==TX_14$DS_RSL,"DS","NN"))))))))
);

# Filter NA data and Take necessary columns

TX14_RSL =
TX_14[complete.cases(TX_14[,c("ACP_ALLIGATOR_CRACKING_PCT","ACP_LONGITUDE_CRA
CKING_PCT","ACP_TRANSVERSE_CRACKING_QTY",
"ACP_RUT_AUTO_SHALLOW_AVG_PCT","ACP_RUT_AUTO_DEEP_AVG_PCT","ACP_FAILUR
E_QTY","ACP_BLOCK_CRACKING_PCT","ACP_PATCHING_PCT", "DISTRESS_SCORE")]),];

TX14_RSL = TX14_RSL[,c("Unique", "AC_RSL", "LC_RSL", "TC_RSL", "ShRUT_RSL",
"DRUT_RSL", "Failure_RSL", "BC_RSL", "Pathc_RSL", "DS_RSL", "Section_RSL")];

# Output

write.csv(TX14_RSL,"C:/RR-PMIS/TX14_RSL.txt");
```

APPENDIX C

EFFECT OF CONSIDERING MULTIPLE DIMENSIONS OF ERROR DETECTIONS IN PAVEMENT CONDITION DATA

As mentioned in Section 6, the influence of considering multiple error detection dimensions of pavement conditions data was investigated. Three error detection techniques were defined such that each considered a different number of these dimensions when used to identify likely errors. The results were then assessed and compared to one another to determine the general impact of using these multiple dimensions in error detection. Following figures present results for AC, TC, IRI, and rutting for Brownwood District roadway network.

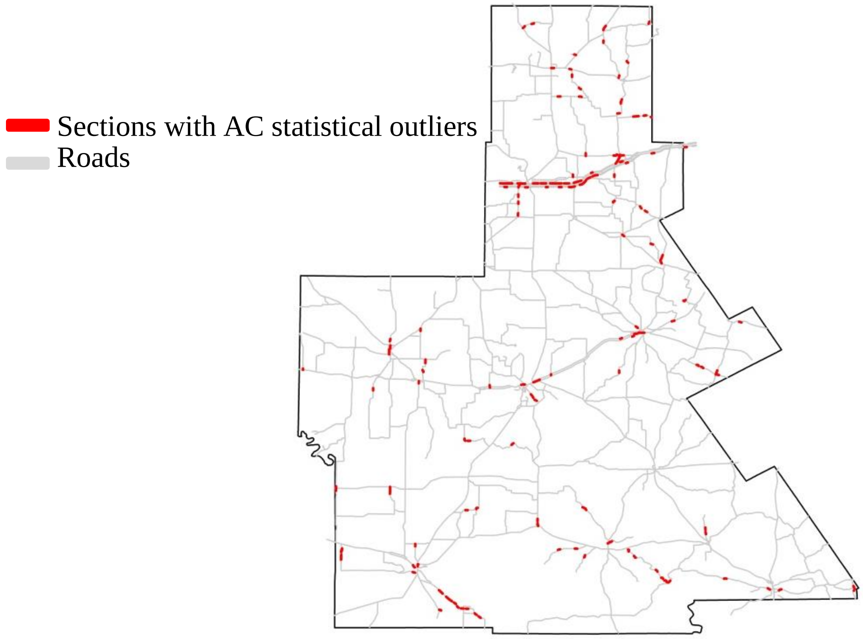


Figure C1 Results of Case 1 analysis: sections with erroneous AC values (Brownwood District)

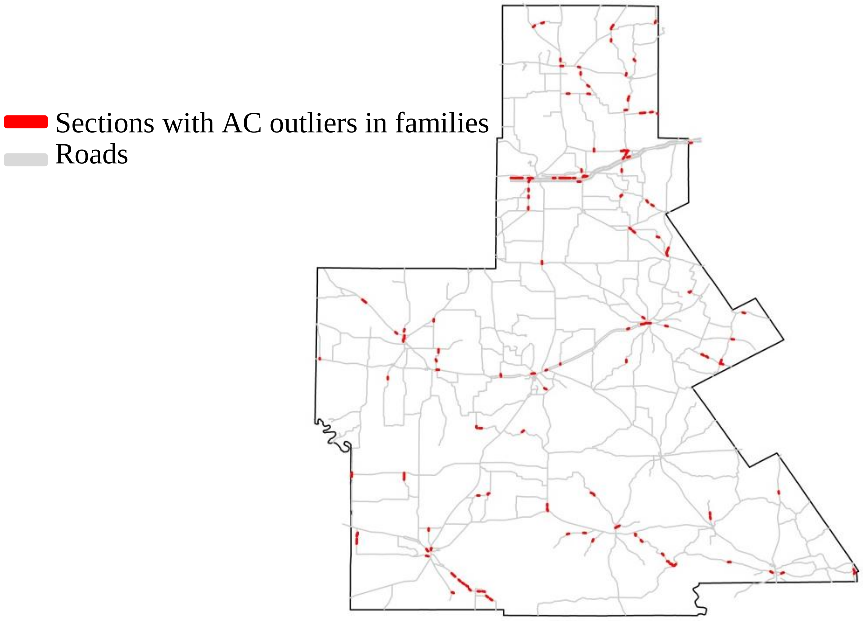


Figure C2 Results of Case 2 analysis: sections with erroneous AC values (Brownwood District)

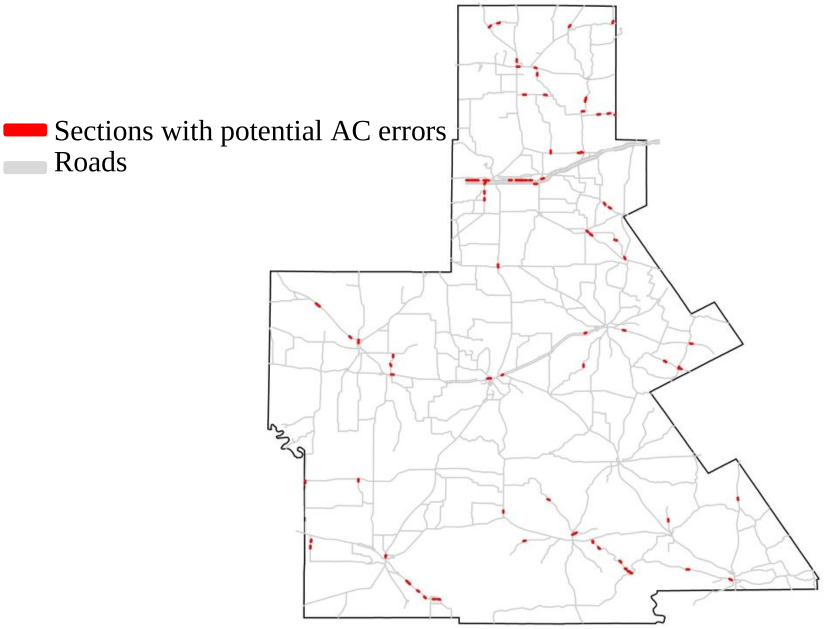


Figure C3 Results of Case 3 analysis: sections with erroneous AC values (Brownwood District)

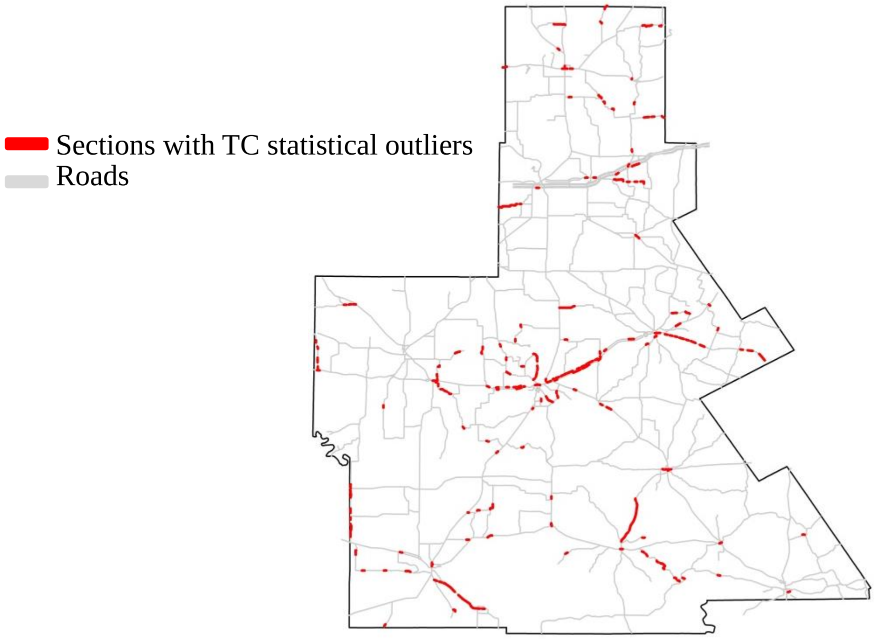


Figure C4 Results of Case 1 analysis: sections with erroneous TC values (Brownwood District)

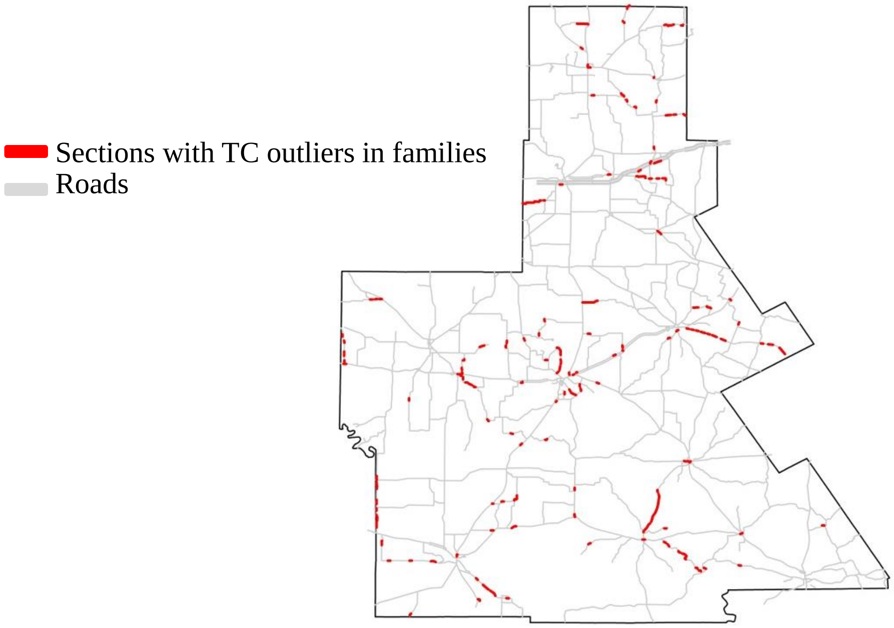


Figure C5 Results of Case 2 analysis: sections with erroneous TC values (Brownwood District)

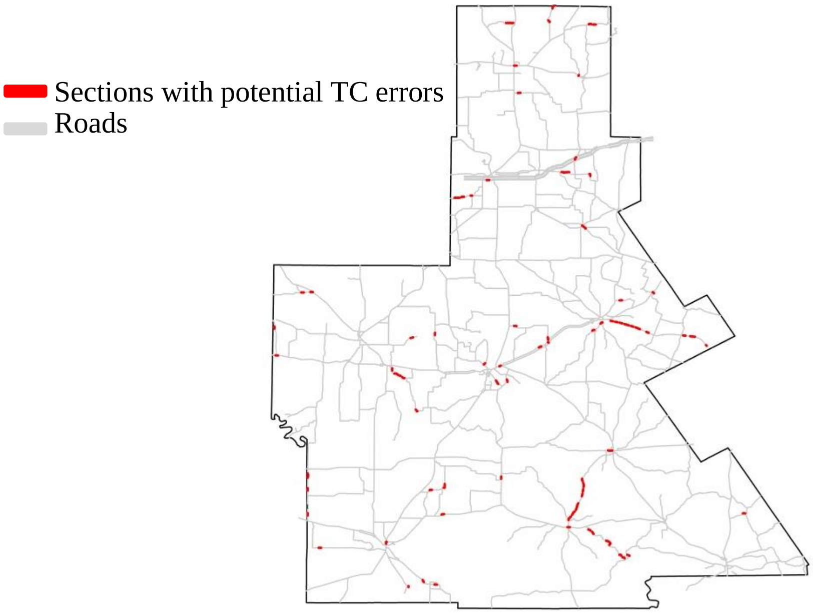


Figure C6 Results of Case 3 analysis: sections with erroneous TC values (Brownwood District)

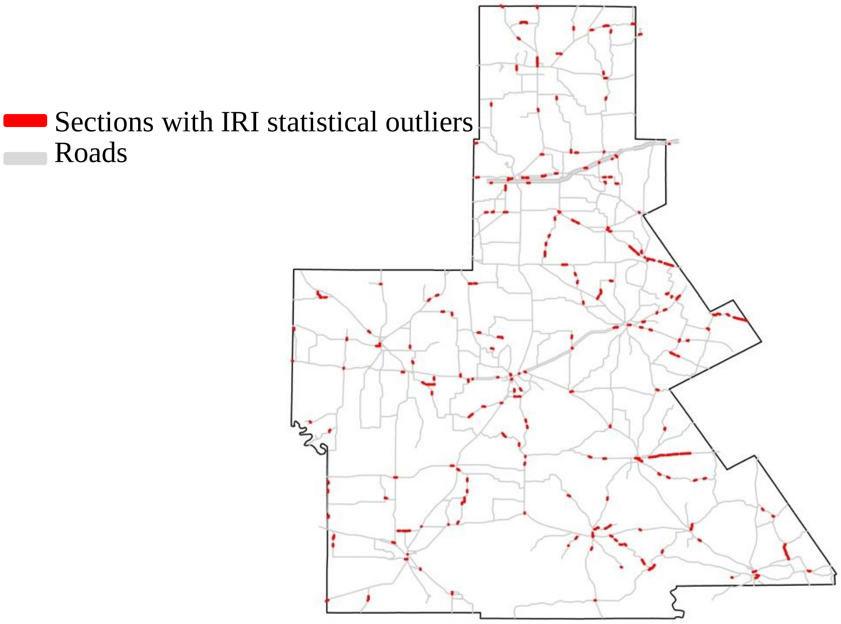


Figure C7 Results of Case 1 analysis: sections with erroneous IRI values (Brownwood District)

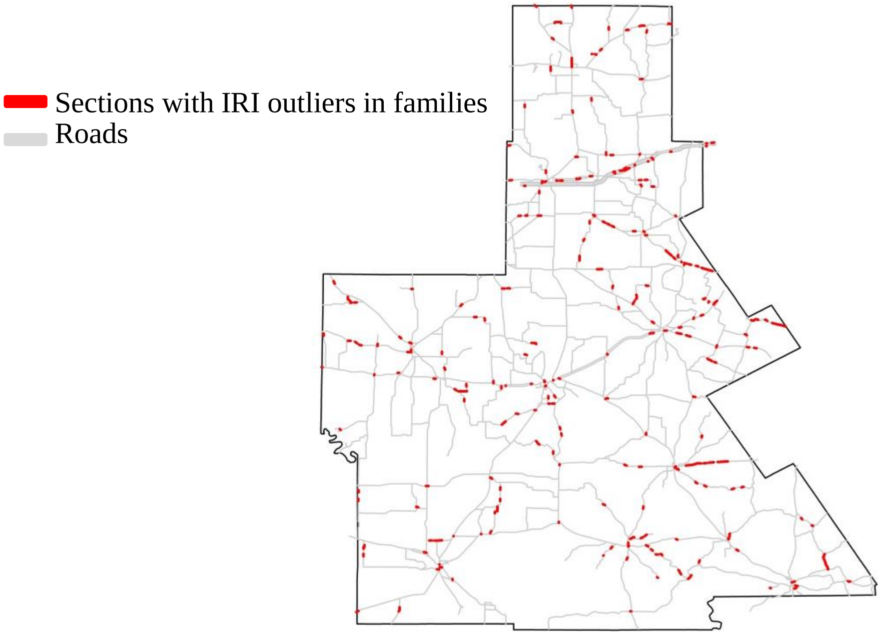


Figure C8 Results of Case 2 analysis: sections with erroneous IRI values (Brownwood District)

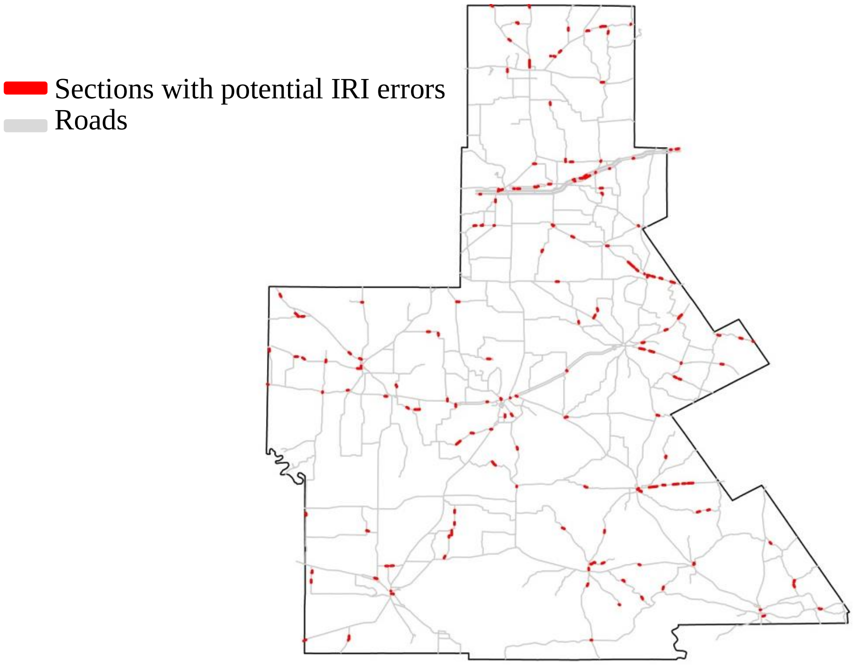


Figure C9 Results of Case 3 analysis: sections with erroneous IRI values (Brownwood District)

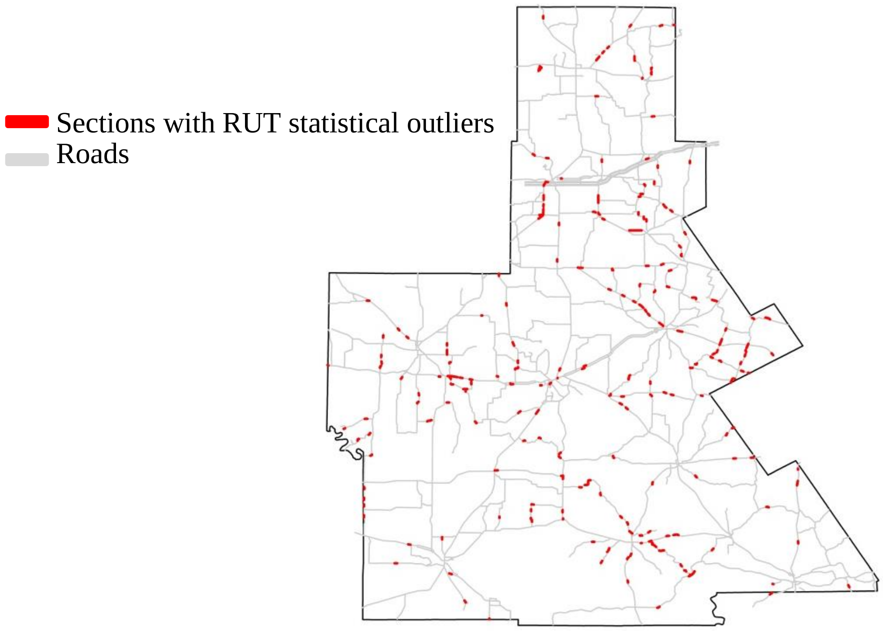


Figure C10 Results of Case 1 analysis: sections with erroneous RUT values (Brownwood District)

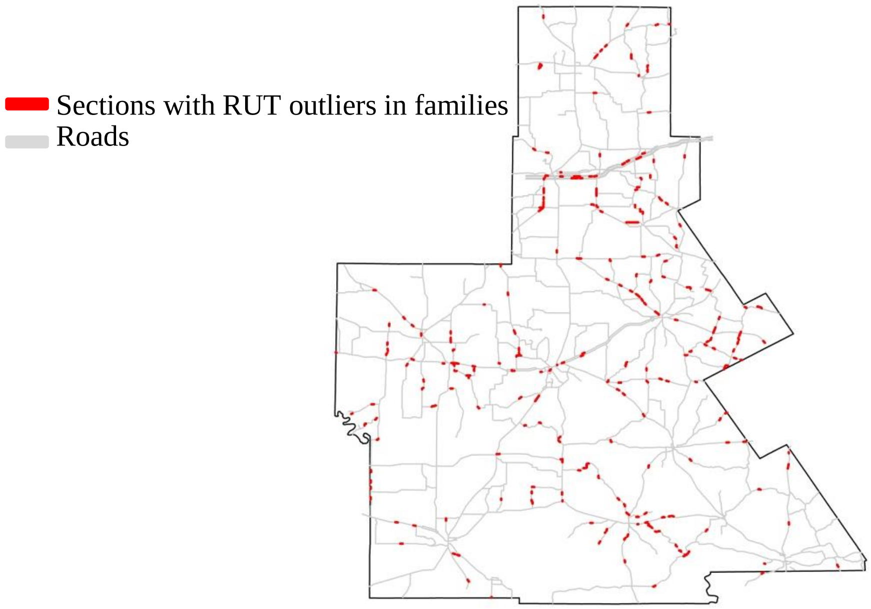


Figure C11 Results of Case 2 analysis: sections with erroneous RUT values (Brownwood District)

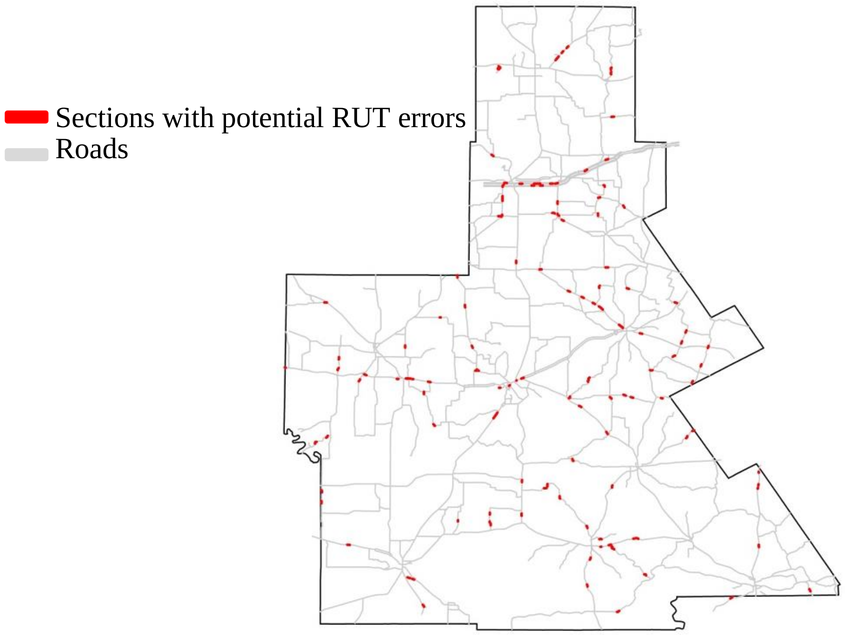


Figure C12 Results of Case 3 analysis: sections with erroneous RUT values (Brownwood District)

APPENDIX D

CODE OF COMPUTER PROGRAM DEVELOPED FOR DETECTING ERRORS CONSIDERING DIFFERENT PROPERTIES OF PAVEMENT CNDITION DATA

```
# Get data

        Condition14_13 = read.csv("C:/R_Dissertation/Condition14_13.csv");

# This code should be activated when it is only considering BrownWood District data

        Condition14_13 = Condition14_13[Condition14_13$RESPONSIBLE_DISTRICT==23,];


# Calculate the Diff

        Condition14_13$D_AC = Condition14_13$ACP_ALLIGATOR_CRACKING_PCT -
        Condition14_13$ACP_ALLIGATOR_CRACKING_PCT_13;

        Condition14_13$D_LC = Condition14_13$ACP_LONGITUDE_CRACKING_PCT -
        Condition14_13$ACP_LONGITUDE_CRACKING_PCT_13;

        Condition14_13$D_TC = Condition14_13$ACP_TRANSVERSE_CRACKING_QTY -
        Condition14_13$ACP_TRANSVERSE_CRACKING_QTY_13;

        Condition14_13$D_Failure = Condition14_13$ACP_FAILURE_QTY -
        Condition14_13$ACP_FAILURE_QTY_13;

        Condition14_13$D_Flush = Condition14_13$ACP_FLUSHING_CODE -
        Condition14_13$ACP_FLUSHING_CODE_13;

        Condition14_13$D_Patch = Condition14_13$ACP_PATCHING_PCT -
        Condition14_13$ACP_PATCHING_PCT_13;

        Condition14_13$D_Ravel = Condition14_13$ACP_RAVELING_CODE -
        Condition14_13$ACP_RAVELING_CODE_13;

        Condition14_13$D_RUT = (Condition14_13$ACP_RUT_AUTO_DEEP_AVG_PCT +
        Condition14_13$ACP_RUT_AUTO_SEVERE_AVG_PCT +
        Condition14_13$ACP_RUT_AUTO_FAILURE_AVG_PCT +
        Condition14_13$ACP_RUT_AUTO_SHALLOW_AVG_PCT) -
        (Condition14_13$ACP_RUT_AUTO_DEEP_AVG_PCT_13 +
        Condition14_13$ACP_RUT_AUTO_SEVERE_AVG_PCT_13 +
        Condition14_13$ACP_RUT_AUTO_FAILURE_AVG_PCT_13 +
        Condition14_13$ACP_RUT_AUTO_SHALLOW_AVG_PCT_13);
```

```
    Condition14_13$D_IRI = ((Condition14_13$IRI_LEFT_SCORE +
    Condition14_13$IRI_RIGHT_SCORE) - (Condition14_13$IRI_LEFT_SCORE_13 +
    Condition14_13$IRI_RIGHT_SCORE_13))/2;

    Condition14_13$D_BC = Condition14_13$ACP_BLOCK_CRACKING_PCT -
    Condition14_13$ACP_BLOCK_CRACKING_PCT_13;

# Merge 2013 and 2014 data to calculate normalization coeficients

    Condition14_13$My_IRI_14 = (Condition14_13$IRI_LEFT_SCORE +
    Condition14_13$IRI_RIGHT_SCORE)/2

    Condition14_13$My_RUT_14 = Condition14_13$ACP_RUT_AUTO_DEEP_AVG_PCT +
    Condition14_13$ACP_RUT_AUTO_SEVERE_AVG_PCT +
    Condition14_13$ACP_RUT_AUTO_FAILURE_AVG_PCT +
    Condition14_13$ACP_RUT_AUTO_SHALLOW_AVG_PCT;

    Condition14_13$My_IRI_13 = (Condition14_13$IRI_LEFT_SCORE_13 +
    Condition14_13$IRI_RIGHT_SCORE_13)/2

    Condition14_13$My_RUT_13 = Condition14_13$ACP_RUT_AUTO_DEEP_AVG_PCT_13 +
    Condition14_13$ACP_RUT_AUTO_SEVERE_AVG_PCT_13 +
    Condition14_13$ACP_RUT_AUTO_FAILURE_AVG_PCT_13 +
    Condition14_13$ACP_RUT_AUTO_SHALLOW_AVG_PCT_13;

    ForNorm_IRI = c(Condition14_13$My_IRI_14,Condition14_13$My_RUT_13);

    ForNorm_RUT = c(Condition14_13$My_RUT_14,Condition14_13$My_RUT_13);

    ForNorm_AC =
    c(Condition14_13$ACP_ALLIGATOR_CRACKING_PCT,Condition14_13$ACP_ALLIGATO
    R_CRACKING_PCT_13);

    ForNorm_LC =
    c(Condition14_13$ACP_LONGITUDE_CRACKING_PCT,Condition14_13$ACP_LONGITUD
    E_CRACKING_PCT_13);

    ForNorm_TC =
    c(Condition14_13$ACP_TRANSVERSE_CRACKING_QTY,Condition14_13$ACP_TRANSV
    ERSE_CRACKING_QTY_13);

    ForNorm_Fail= c(Condition14_13$ACP_FAILURE_QTY,
    Condition14_13$ACP_FAILURE_QTY_13);

    ForNorm_Flush = c(Condition14_13$ACP_FLUSHING_CODE,
    Condition14_13$ACP_FLUSHING_CODE_13);

    ForNorm_Ravel =
    c(Condition14_13$ACP_RAVELING_CODE,Condition14_13$ACP_RAVELING_CODE_13);

    ForNorm_BC = c(Condition14_13$ACP_BLOCK_CRACKING_PCT,
    Condition14_13$ACP_BLOCK_CRACKING_PCT_13);

# Coeficient to normalize Changes
```

```
        Norm_Coef= matrix(0,nrow = 9,ncol=2);

        colnames(Norm_Coef)=c("Mean","Std");

        rownames(Norm_Coef)=
        c("AC_D","LC_D","TC_D","FAIL_D","FLUSH_D","RAV_D","RUT_D","IRI_D","BC_D");

        Norm_Coef[1,1]= mean(ForNorm_AC);

        Norm_Coef[1,2]= sd(ForNorm_AC);

        Norm_Coef[2,1]= mean(ForNorm_LC);

        Norm_Coef[2,2]= sd(ForNorm_LC);

        Norm_Coef[3,1]= mean(ForNorm_TC);

        Norm_Coef[3,2]= sd(ForNorm_TC);

        Norm_Coef[4,1]= mean(ForNorm_Fail);

        Norm_Coef[4,2]= sd(ForNorm_Fail);

        Norm_Coef[5,1]= mean(ForNorm_Flush);

        Norm_Coef[5,2]= sd(ForNorm_Flush);

        Norm_Coef[6,1]= mean(ForNorm_Ravel);

        Norm_Coef[6,2]= sd(ForNorm_Ravel);

        Norm_Coef[7,1]= mean(ForNorm_RUT);

        Norm_Coef[7,2]= sd(ForNorm_RUT);

        Norm_Coef[8,1]= mean(ForNorm_IRI);

        Norm_Coef[8,2]= sd(ForNorm_IRI);

        Norm_Coef[9,1]= mean(ForNorm_BC);

        Norm_Coef[9,2]= sd(ForNorm_BC);



#  Normalize annual changes

        Condition14_13$N_AC_D = ((Condition14_13$ACP_ALLIGATOR_CRACKING_PCT-
        Norm_Coef[1,1])/Norm_Coef[1,2])-
        ((Condition14_13$ACP_ALLIGATOR_CRACKING_PCT_13-
        Norm_Coef[1,1])/Norm_Coef[1,2]);

        Condition14_13$N_LC_D = ((Condition14_13$ACP_LONGITUDE_CRACKING_PCT-
        Norm_Coef[2,1])/Norm_Coef[2,2])-
        ((Condition14_13$ACP_LONGITUDE_CRACKING_PCT_13-
        Norm_Coef[2,1])/Norm_Coef[2,2])
        Condition14_13$N_TC_D = ((Condition14_13$ACP_TRANSVERSE_CRACKING_QTY-
        Norm_Coef[3,1])/Norm_Coef[3,2])-
        ((Condition14_13$ACP_TRANSVERSE_CRACKING_QTY_13-
        Norm_Coef[3,1])/Norm_Coef[3,2]);

        Condition14_13$N_FAIL_D = ((Condition14_13$ACP_FAILURE_QTY-
        Norm_Coef[4,1])/Norm_Coef[4,2])-((Condition14_13$ACP_FAILURE_QTY_13-
        Norm_Coef[4,1])/Norm_Coef[4,2]);

        Condition14_13$N_FLUSH_D = ((Condition14_13$ACP_FLUSHING_CODE-
        Norm_Coef[5,1])/Norm_Coef[5,2])-((Condition14_13$ACP_FLUSHING_CODE_13-
        Norm_Coef[5,1])/Norm_Coef[5,2]);

        Condition14_13$N_RAV_D = ((Condition14_13$ACP_RAVELING_CODE-
        Norm_Coef[6,1])/Norm_Coef[6,2])-((Condition14_13$ACP_RAVELING_CODE_13-
        Norm_Coef[6,1])/Norm_Coef[6,2]);

        Condition14_13$N_RUT_D = ((Condition14_13$My_RUT-Norm_Coef[7,1])/Norm_Coef[7,2])-
        ((Condition14_13$My_RUT_13 - Norm_Coef[7,1])/Norm_Coef[7,2]);

        Condition14_13$N_IRI_D = ((Condition14_13$My_IRI-Norm_Coef[8,1])/Norm_Coef[8,2])-
        ((Condition14_13$My_IRI_13-Norm_Coef[8,1])/Norm_Coef[8,2]);

        Condition14_13$N_BC_D =  ((Condition14_13$ACP_BLOCK_CRACKING_PCT-
        Norm_Coef[9,1])/Norm_Coef[9,2]) - ((Condition14_13$ACP_BLOCK_CRACKING_PCT_13-
        Norm_Coef[9,1])/Norm_Coef[9,2]);

# Get Median and Q1 and Q3

        Z=2;

        require(plyr);

        LIMIT_Measures =
        ddply(Condition14_13,c("ClimateZone","PvFamily","Loading"),summarize,Count=
        length(D_AC),

                    Mean_AC=mean(D_AC),SD_AC=sd(D_AC),UL_AC=Mean_AC + (Z*SD_AC),
        LL_AC= Mean_AC - (Z*SD_AC),

                    Mean_LC=mean(D_LC),SD_LC=sd(D_LC),UL_LC=Mean_LC + (Z*SD_LC),
        LL_LC= Mean_LC - (Z*SD_LC),

                    Mean_TC=mean(D_TC),SD_TC=sd(D_TC),UL_TC=Mean_TC + (Z*SD_TC),
        LL_TC= Mean_TC - (Z*SD_TC),

                    Mean_IRI=mean(D_IRI),SD_IRI=sd(D_IRI),UL_IRI=Mean_IRI + (Z*SD_IRI),
        LL_IRI= Mean_IRI - (Z*SD_IRI),

                    Mean_RUT=mean(D_RUT),SD_RUT=sd(D_RUT),UL_RUT=Mean_RUT +
        (Z*SD_RUT), LL_RUT= Mean_RUT - (Z*SD_RUT))
        Cond14_13=merge(Condition14_13,LIMIT_Measures,by=c("ClimateZone","PvFamily","Loadin
        g"));

 # Check for Not Errors

                        Cond14_13$pos_surf =
        (Cond14_13$D_AC>0)+(Cond14_13$D_TC>0)+(Cond14_13$D_LC>0)+(Cond14_13$D_Flush
        >0)+(Cond14_13$D_Failure>0)+(Cond14_13$D_Ravel>0)+(Cond14_13$D_BC>0);

                        Cond14_13$Neg_surf =
        (Cond14_13$D_AC<0)+(Cond14_13$D_TC<0)+(Cond14_13$D_LC<0)+(Cond14_13$D_Flush
        <0)+(Cond14_13$D_Failure<0)+(Cond14_13$D_Ravel<0)+(Cond14_13$D_BC<0);

                        Cond14_13$D_surf = (Cond14_13$pos_surf > Cond14_13$Neg_surf) -
        (Cond14_13$Neg_surf> Cond14_13$pos_surf);

                        Cond14_13$SurfCrack_NotError_pos = (Cond14_13$D_surf>0 &
        Cond14_13$D_IRI >= 0 & Cond14_13$D_RUT >=0);

                        Cond14_13$SurfCrack_NotError_Neg = (Cond14_13$D_surf<0 &
        ((Cond14_13$D_IRI >= 0 & Cond14_13$D_RUT >=0) | (Cond14_13$D_IRI < 0 &
        Cond14_13$D_RUT < 0)));

                        Cond14_13$IRI_RUT_NotError = (Cond14_13$D_IRI>0 &
        Cond14_13$D_RUT>0) | (Cond14_13$D_IRI<0 & Cond14_13$D_RUT<0 &
        Cond14_13$D_surf<0)|(Cond14_13$D_IRI>=0 & Cond14_13$D_RUT>=0 &
        Cond14_13$D_surf>0);

# Select Outlier and detect Errors as TRUE variables

        Cond14_13$LC_Error = ((Cond14_13$D_LC > Cond14_13$UL_LC)&
        (Cond14_13$SurfCrack_NotError_pos==FALSE)) | ((Cond14_13$D_LC <
        Cond14_13$LL_LC)&(Cond14_13$SurfCrack_NotError_Neg==FALSE));

        Cond14_13$AC_Error = ((Cond14_13$D_AC > Cond14_13$UL_AC)&
        (Cond14_13$SurfCrack_NotError_pos==FALSE)) | ((Cond14_13$D_AC <
        Cond14_13$LL_AC)&(Cond14_13$SurfCrack_NotError_Neg==FALSE));

        Cond14_13$TC_Error = ((Cond14_13$D_TC > Cond14_13$UL_TC)&
        (Cond14_13$SurfCrack_NotError_pos==FALSE)) | ((Cond14_13$D_TC <
        Cond14_13$LL_TC)&(Cond14_13$SurfCrack_NotError_Neg==FALSE));

        Cond14_13$IRI_Error = ((Cond14_13$D_IRI > Cond14_13$UL_IRI)|(Cond14_13$D_IRI <
        Cond14_13$LL_IRI)) & (Cond14_13$IRI_RUT_NotError==FALSE);

        Cond14_13$RUT_Error = ((Cond14_13$D_RUT > Cond14_13$UL_RUT)|(Cond14_13$D_RUT
        < Cond14_13$LL_RUT)) & (Cond14_13$IRI_RUT_NotError==FALSE);

        Cond14_13$Errroneous = (Cond14_13$LC_Error==TRUE)| (Cond14_13$AC_Error==TRUE) |
        (Cond14_13$TC_Error==TRUE) | (Cond14_13$IRI_Error==TRUE) |
        (Cond14_13$RUT_Error==TRUE);

        Cond14_13$LC_Outlier = (Cond14_13$D_LC > Cond14_13$UL_LC)|(Cond14_13$D_LC <
        Cond14_13$LL_LC);
        Cond14_13$AC_Outlier = (Cond14_13$D_AC > Cond14_13$UL_AC)|(Cond14_13$D_AC <
        Cond14_13$LL_AC);

        Cond14_13$TC_Outlier = (Cond14_13$D_TC > Cond14_13$UL_TC)|(Cond14_13$D_TC <
        Cond14_13$LL_TC);

        Cond14_13$IRI_Outlier = (Cond14_13$D_IRI > Cond14_13$UL_IRI)|(Cond14_13$D_IRI <
        Cond14_13$LL_IRI);

        Cond14_13$RUT_Outlier = (Cond14_13$D_RUT >
        Cond14_13$UL_RUT)|(Cond14_13$D_RUT < Cond14_13$LL_RUT);

        Cond14_13$Outlier = Cond14_13$LC_Outlier>0
        |Cond14_13$AC_Outlier>0|Cond14_13$TC_Outlier>0|Cond14_13$IRI_Outlier>0 |
        Cond14_13$RUT_Outlier>0;

#CurrentTechniques

        Mean_AC_Current=mean(Cond14_13$D_AC);SD_AC_Current=sd(Cond14_13$D_AC);UL_A
        C_Current=Mean_AC_Current + (Z*SD_AC_Current); LL_AC_Current= Mean_AC_Current -
        (Z*SD_AC_Current);

        Mean_LC_Current=mean(Cond14_13$D_LC);SD_LC_Current=sd(Cond14_13$D_LC);UL_LC_
        Current=Mean_LC_Current + (Z*SD_LC_Current); LL_LC_Current= Mean_LC_Current -
        (Z*SD_LC_Current);

        Mean_TC_Current=mean(Cond14_13$D_TC);SD_TC_Current=sd(Cond14_13$D_TC);UL_TC_
        Current=Mean_TC_Current + (Z*SD_TC_Current); LL_TC_Current= Mean_TC_Current -
        (Z*SD_TC_Current);

        Mean_IRI_Current=mean(Cond14_13$D_IRI);SD_IRI_Current=sd(Cond14_13$D_IRI);UL_IRI
        _Current=Mean_IRI_Current + (Z*SD_IRI_Current); LL_IRI_Current= Mean_IRI_Current -
        (Z*SD_IRI_Current);

        Mean_RUT_Current=mean(Cond14_13$D_RUT);SD_RUT_Current=sd(Cond14_13$D_RUT);U
        L_RUT_Current=Mean_RUT_Current + (Z*SD_RUT_Current); LL_RUT_Current=
        Mean_RUT_Current - (Z*SD_RUT_Current);

        Cond14_13$AC_Outlier_Current = Cond14_13$D_AC > UL_AC_Current | Cond14_13$D_AC
        < LL_AC_Current ;

        Cond14_13$LC_Outlier_Current = Cond14_13$D_LC > UL_LC_Current | Cond14_13$D_LC <
        LL_LC_Current ;

        Cond14_13$TC_Outlier_Current = Cond14_13$D_TC > UL_TC_Current | Cond14_13$D_TC <
        LL_TC_Current ;

        Cond14_13$IRI_Outlier_Current = Cond14_13$D_IRI > UL_IRI_Current | Cond14_13$D_IRI <
        LL_IRI_Current ;

         Cond14_13$RUT_Outlier_Current = Cond14_13$D_RUT > UL_RUT_Current |
        Cond14_13$D_RUT < LL_RUT_Current ;
        Cond14_13$Outlier_Current =
        (Cond14_13$AC_Outlier_Current>0)|(Cond14_13$LC_Outlier_Current>0)|(Cond14_13$TC_Outlier_Current>0)|(Cond14_13$IRI_Outlier_Current>0)|(Cond14_13$RUT_Outlier_Current>0);

# Define Output

        Cond14_13$AC_14 = Cond14_13$ACP_ALLIGATOR_CRACKING_PCT;

         Cond14_13$LC_14 = Cond14_13$ACP_LONGITUDE_CRACKING_PCT;

        Cond14_13$TC_14 = Cond14_13$ACP_TRANSVERSE_CRACKING_QTY;

        Cond14_13$FAIL_14 = Cond14_13$ACP_FAILURE_QTY;

        Cond14_13$FLUSH_14 = Cond14_13$ACP_FLUSHING_CODE;

        Cond14_13$RAV_14 = Cond14_13$ACP_RAVELING_CODE;

        Cond14_13$RUT_14 = Cond14_13$ACP_RUT_AUTO_DEEP_AVG_PCT+
        Cond14_13$ACP_RUT_AUTO_SEVERE_AVG_PCT +
        Cond14_13$ACP_RUT_AUTO_FAILURE_AVG_PCT +
        Cond14_13$ACP_RUT_AUTO_SHALLOW_AVG_PCT;

        Cond14_13$IRI_14 = (Cond14_13$IRI_LEFT_SCORE+Cond14_13$IRI_RIGHT_SCORE)/2

        Cond14_13$BC_14 = Cond14_13$ACP_BLOCK_CRACKING_PCT



         Errors =
        Cond14_13[Cond14_13$Errroneous==1,c("Unique","LC_Error","AC_Error","TC_Error","IRI_Error","RUT_Error","Errroneous","LC_Outlier","AC_Outlier","TC_Outlier","IRI_Outlier","RUT_Outlier","Outlier","LC_Outlier_Current","AC_Outlier_Current","TC_Outlier_Current","IRI_Outlier_Current","RUT_Outlier_Current","Unique_GIS","D_AC","D_LC","D_TC","D_Failure","D_Flush","D_Ravel","D_RUT","D_IRI","D_BC")];

          Outliers =
        Cond14_13[Cond14_13$Outlier==1,c("Unique","LC_Error","AC_Error","TC_Error","IRI_Error","RUT_Error","Errroneous","LC_Outlier","AC_Outlier","TC_Outlier","IRI_Outlier","RUT_Outlier","Outlier","LC_Outlier_Current","AC_Outlier_Current","TC_Outlier_Current","IRI_Outlier_Current","RUT_Outlier_Current","Unique_GIS","D_AC","D_LC","D_TC","D_Failure","D_Flush","D_Ravel","D_RUT","D_IRI","D_BC")];

        Outliers_Current =
        Cond14_13[Cond14_13$Outlier_Current==1,c("Unique","LC_Error","AC_Error","TC_Error","IRI_Error","RUT_Error","Errroneous","LC_Outlier","AC_Outlier","TC_Outlier","IRI_Outlier","RUT_Outlier","Outlier","LC_Outlier_Current","AC_Outlier_Current","TC_Outlier_Current","IRI_Outlier_Current","RUT_Outlier_Current","Unique_GIS","D_AC","D_LC","D_TC","D_Failure","D_Flush","D_Ravel","D_RUT","D_IRI","D_BC")];

        General =
        Cond14_13[( Cond14_13$Outlier_Current==1|Cond14_13$Outlier==1|Cond14_13$Errroneous==1),c("Unique","LC_Error","AC_Error","TC_Error","IRI_Error","RUT_Error","Errroneous","LC_Outlier","AC_Outlier","TC_Outlier","IRI_Outlier","RUT_Outlier","Outlier","LC_Outlier_Current","AC_Outlier_Current","TC_Outlier_Current","IRI_Outlier_Current","RUT_Outlier_Curren
        t","Unique_GIS","D_AC","D_LC","D_TC","D_Failure","D_Flush","D_Ravel","D_RUT","D_IR
        I","D_BC","LL_LC","UL_LC","LL_AC","UL_LC","LL_TC","UL_TC","LL_IRI","UL_IRI","L
        L_RUT","UL_RUT", "AC_14", "LC_14", "TC_14", "FAIL_14", "FLUSH_14", "RAV_14",
        "RUT_14","IRI_14","BC_14","ClimateZone","PvFamily","Loading",
        "N_AC_D","N_LC_D","N_TC_D","N_FAIL_D","N_FLUSH_D","N_RAV_D","N_RUT_D","N
        _IRI_D","N_BC_D")];


# Output

        library(foreign)

        write.dbf(Errors,"C:/Dissertation/Effect of multiple dimensions/Errors.dbf");

        write.dbf(Outliers,"C:/Dissertation/Effect of multiple dimensions/Outliers.dbf");

        write.dbf(Outliers_Current,"C:/Dissertation/Effect of multiple dimensions/Outliers_Current.dbf");

        write.dbf(General,"C:/Dissertation/Effect of multiple dimensions/General.dbf");
```