A DATA MINING APPROACH FOR IDENTIFYING PAVEMENT DISTRESS SIGNATURES

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Megan Sue Bouret

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science

November 2015

Fargo, North Dakota

NORTH DAKOTA STATE UNIVERSITY

Graduate School

Title

A DATA MINING APPROACH FOR IDENTIFYING PAVEMENT DISTRESS SIGNATURES

By

Megan Sue Bouret

The supervisory committee certifies that this thesis complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Anne Denton
Chair

Dr. Saeed Salem

Dr. Changhui Yan

Dr. Pan Lu

Approved:

11/20/2015
Date

Dr. Brian Slator
Department Chair

ABSTRACT

This work introduces signature-based data mining of pavement distress data. The goal is to understand the factors that influence pavement distress. The presented approach maintains multiple types of flexible pavement distress scores throughout the analysis and considers them as signatures. The signatures are used to establish the relationship between distress score increases and overweight truck characteristics. Hierarchical clustering of pavement distress signatures provides insights into similarities among road segments. The use of signatures, rather than composite distress scores, is consistent with a data mining approach to the pavement distress problem. One set of experiments showed a relationship between the discovered signature groups and a difference between overweight truck traffic. Group validation has been implemented with Fisher's exact test. Future work related to algorithm improvements have been identified and considered.

ACKNOWLEDGEMENTS

I would like to acknowledge the part that my adviser Dr. Anne Denton, Terry Woehl and Stephanie Weigel from the NDDOT, and the thesis supervisory committee members had in my accomplishment. Dr. Denton's advising style encouraged me to own my power and my project. I would also like to recognize the many people in the NDSU Computer Science and Upper Great Plains Transportation Institute departments that provided me so much support and encouragement over the years. I am also indebted to the NDSU student chapter of the Association for Computing Machinery for the friendship and opportunity it bestowed upon me. I could not have finished without the incredible service provided by the NDSU Counseling Center. There are so many supportive people in NDSU IT, Facilities Management, etc. who all provided me encouragement and kindness. This has been a tremendous and challenging learning opportunity and would have never happened without your assistance. Thank you.

I am also thankful for the strengths and challenges provided to me by my parents. I also express deep gratitude for the love and support provided by my boyfriend, my sponsor, my goal witness, the sisterhood, my recovery friends, and VRii. The commonality with those recognized is that you loved and believed in me and taught me to love and believe in myself. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The recent oil boom in western North Dakota has changed the traffic characteristics in that region. An opportunity exists to gain understanding regarding the relationship between pavement distress scores and truck traffic characteristics. The focus of this work is on flexible pavement surface sections of U.S. highways, near weigh-in-motion (WIM) sites. It is shown that truck traffic characteristics can be recognized by hierarchical clustering to separate pavement segments in specific circumstances.

By understanding how truck traffic impacts pavement distress, additional knowledge is gained regarding how traffic, road construction and rehabilitation, and environmental factors all influence pavement life. Greater understanding leads to more cost-effective roadway infrastructure maintenance. Increases in flexible pavement segment distress scores were used to determine whether hierarchical clustering could recognize categories in the pavement distress data that corresponded to truck traffic characteristics observed at nearby weigh-in-motion (WIM) sites.

In one experiment set, construction, rehabilitation, and weather could be disregarded to highlight overweight truck traffic's impact near North Dakota's western oil producing region with lane division correctly identified between 81%-87% of the selected pavement segments. Related work concerning how different truck configuration types affect pavement differently was partially verified by the presented results. While the results are related to a specific span of time with specific pavement segments, there exists considerable work in this area.

1.1. Problem Statement

1.1.1. Problem

Two specific problems exist concerning pavement performance maintenance: data complexity and data accessibility. There are multiple influence categories that not only affect pavement distress but also each other. On one end is the environmental influences, including weather and soil. The deterioration caused by these external factors are counteracted by the pavement's base and surface construction and rehabilitation. Expected traffic influences the construction and rehabilitation.

In order to accurately model the complexity, accurate data would have to be collected for the different influence categories. It would then have to be available for model generation. Much of the data is already being accurately collected and maintained by different organizations with different priorities.

1.1.2. Objective

Even if the data can be brought together, the sheer amount and complexity of it prevents development of a timely and efficient model representation. Understanding how specific influence categories affect pavement distress can improve pavement distress model representation. In a sense, the problem is being decomposed. Traffic is an external influence category that can fluctuate and vary greatly in a local area. Truck traffic causes more distress due to its load-bearing capabilities. The objective of this work is to determine whether pavement distress scores can constitute a signature that corresponds to truck traffic characteristics.

1.1.3. Concept

By selecting pavement segments near WIM sites and are similar in construction, the problem is simplified further. The pavement segments are then clustered by their pavement distress signature, defined by the pavement distress score increases experienced. The resulting clusters are then visually and statistically analyzed for a correspondence with the truck traffic characteristics determined by data collected at WIM sites.

1.2. Related Work

Through regression analysis, [22] concluded that trucks with single or tandem axles affected cracking more, while trucks with more complicated axle configurations affected rutting more. [5] observed significant pavement distress prediction errors when national or Arizona state default factors were substituted for WIM data as mechanistic-empirical pavement design guide (MEPDG) input. The MEPDG was developed by the American Association of State Highway and Transportation Officials (AASHTO) in 2004 to provide states with a means to predict pavement life based on various inputs [24].

Pavement distress score patterns were found by [8]. With the aid of GIS, a regression model was developed that showed parallels between New Mexico pavement distress scores and truck traffic characteristics. Increased deterioration was found in northern New Mexico, where it was more urbanized with more truck traffic [8]. By implementing a data mining approach that could accurately classify all overweight trucks, legal or not, [12] was able to utilize WIM data to identify non-permitted overweight trucks with less than a 10% standard error.

By studying the relationship between overweight truck traffic and flexible pavement service life in Poland, [21] observed pavement service life was reduced as much as 50% when overweight truck traffic grew by 20%. They also recognized that when enforcement reduced overweight truck traffic by 10%, pavement service life could be extended 4 to 6 years [21]. [24] determined a relationship between pavement distress scores and overweight traffic data through the use of the MEPDG. Rutting and alligator cracking were the primary distress types regarded by [24]. They also discovered a linear relationship between the percentage of overweight traffic and the decrease in pavement life [24].

1.3. Related Work Challenges

Pavement distress data varies from year to year, heavily influenced by rehabilitation. Pavement condition composite indices can get outdated relatively quickly even when the weighted factors used to calculate them are based on multi-variate optimization analysis. While MEPDG has had success predicting pavement life, it does require many precise data inputs to achieve accurate results. Though prediction relates to an understanding of the relationship between pavement distress and its causes, exact results require additional inputs beyond truck traffic characteristics.

In general, statistics utilizes a small number of variables, even in multi-variate analysis. The approach presented in this work is multi-dimensional in nature but different from the pavement condition composite indices used by some jurisdictions. The selected subset of pavement distress scores is considered equal in the hierarchical clustering approach being implemented. Data mining methods like hierarchical clustering show the relationship found between the pavement distress and truck traffic data sets by calculating tree-like representations called dendrograms that show similar pavement segments together.

1.4. Organization of Thesis

Chapter 2 provides an overview of data mining concepts such as clustering, proximity utilization, and statistical analysis. Information describing the pavement distress and truck traffic data sets used can be found in Chapter 3. Specific implementation details related to the data mining concepts explained in Chapter 2 are located in Chapter 4.

Exploratory experiments are summarized in Chapter 5, including an introduction to the vehicle re-identification problem. Classification by regression was attempted and related work later discovered that compared the use Bayesian and neural network models to solve the vehicle re-identification problem.

How dendrograms (tree diagrams) and significance testing show promise in increasing knowledge concerning pavement distress and truck traffic characteristics is located in Chapter 6. The conclusion of the main material is in Chapter 7.2.2 with the references at the end.

2. DATA MINING CONCEPTS

2.1. Data Mining and Computer Science

[18] summarizes the development of data mining as a computer science field. From artificial intelligence's need to understand an agent's environment arose machine learning. Data mining emerged as data analysis approaches developed in machine learning. There are numerous topics associated with data mining which may also be related to machine learning or statistics [18]. There are also data mining areas that do not apply to either, such as data warehousing and database querying [18].

2.2. Proximity Utilization

The proximity between data subsets can be utilized in different ways, such as prediction or clustering. With clustering, subsets can be discovered in the data depending on attribute values. Data points can be associated with data subsets with classification. Data point attribute values can be predicted by regression.

The introduction referenced the following classification approaches: Bayesian, neural networks, and linear regression. Bayesian approaches use the famous Bayes' Rule to determine the probability of the hypothesis given the evidence [18]. The probability's value determines the observation's class assignment. [18] explains that classification by mathematical functions include neural networks and linear regression. These approaches are separated by the fact that neural networks can be used with nonlinear functions [18].

2.2.1. Clustering

Clustering can be further divided into hierarchical clustering and centroid-based clustering. K-means is the most often utilized centroid-based clustering algorithm [18]. K-means finds the k subsets in the data by iteratively calculating the arithmetic mean for the k estimated clusters. With each calculation, it adjusts the clusters until they no longer change.

With hierarchical clustering, the goal is to use similarity to link data points to identify and represent the resulting subsets in a tree-like diagram called a dendrogram. Hierarchical clustering can be further divided into agglomerative or divisive approaches. The difference between these methods is that a divisive splits one set into n sets of one data point while an agglomerative merges n sets of one data point to one set [6].

In order to determine which pavement segments are similar, a proximity matrix is utilized to store proximity values between each cluster pair. Most agglomerative hierarchical clustering can be categorized as linkage clustering, centroid clustering, and weighted clustering [6].

[6] describes various agglomerative methods, including unweighted pair-group method using the average (UPGMA). Also known as group average linkage, UPGMA merges clusters by selecting the average distance calculated between all cluster members in each cluster pair [9] [6].

2.3. Proximity Measurements

There are numerous ways to measure proximity between data subsets. [6] recognizes proximity as a general term that include dissimilarity, distance, and similarity. Euclidean distance and cosine similarity are popular proximity measures used not only in clustering but also other data mining areas [9].

One calculation utilized in multiple proximity measures is the cosine similarity, which is illustrated in Fig. 2.1. [9] defines $\overrightarrow{v}d_1$, $\overrightarrow{v}d_2$, and $\overrightarrow{v}d_3$ as vectors representing words and their relationship with the terms "gossip" and "jealous." The between $\overrightarrow{v}d_1$ and $\overrightarrow{v}d_2$ represents the relative difference between the two words. Θ or cosine similarity is defined by Eq. 2.1.

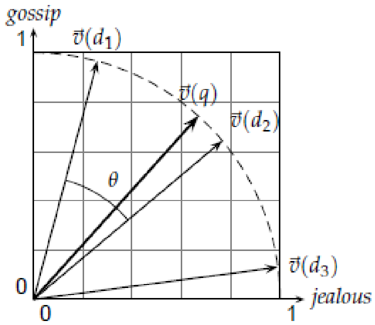


Figure 2.1. Cosine Similarity Example [9]

[9] defines the cosine similarity as the dot product of the segments s_1 and s_2 being divided by the product of the Euclidean lengths of s_1 and s_2. Its inverse, cosine distance is defined by Eq. 2.2. [6] describes $\frac{dis_c}{2}$ is another proximity measure called angular separation. Angular separation uses the cosine distance, dis_c in its definition.

$$sim_c(s_1, s_2) = \frac{(\sum_{i=1}^{m} s_{1i} * s_{2i})}{\sqrt{(\sum_{i=1}^{m} s_{1i}^2)(\sum_{i=1}^{m} s_{2i}^2)}} \text{ [4]} \tag{2.1}$$

$$dis_c(s_1, s_2) = 1 - sim_c(s_1, s_2) \text{ [18]} \tag{2.2}$$

Euclidean distance is defined by Eq. 2.3. It is the most popular proximity measure utilized in cluster analysis [6].

$$dis_e(s_1, s_2) = \sqrt{\sum_{i=1}^{m}(s_{1i} - s_{2i})^2} \text{ [6]} \tag{2.3}$$

2.4. Dendrogram Utilization

Not only do dendrograms visually represent the clusters, but they are also used in statistical significance testing to evaluate the results. Contingency tables can be constructed with branch membership of two selected branches defining one variable as the columns. The expected categories represent the second variable as the rows. The expected categories need to be exclusive, as a branch member can only belong to one category [10].

The null hypothesis, H_0, is that there is no meaningful difference found between the two most suggestive segment subsets discovered by the hierarchical clustering implementation. In order to reject or fail to reject the null hypothesis, a Fisher's exact test would be applied on the contingency table to determine the statistical significance of the resulting dendrogram [10].

2.4.1. Fisher's Exact Test

The Fisher's exact test first appeared in 1934 in R.A. Fisher's work concerning 2x2 contingency tables [3]. Due to the calculation complexity, Fisher's exact test was not practical to use with larger contingency tables advancements were made with computing [10]. Fisher's exact test works by calculating the the probability of each possible contingency table using the hypergeometric distribution [10]. The possible contingency table combinations are determined by any one of the four cells and its respective row [17]. Table 2.1 will be used to explain how Fisher's exact test works and its results. A, B, C, and D are cell notations used in Table 2.1. For example, if $A = 3$ and $A + B = 5$, the possible contingency tables would have A less than or equal to three with the total of $A + B$ equal to five. These contingency tables all have the same marginal totals as the one being tested [23].

[17] describes the method for calculating one-tailed and two-tailed probabilities. The one-tailed probability is calculated by summing the probability of the current contingency table with the extremely low probabilities that have A less than three, the A value in the current example. With a two-tailed

probability, the current contingency table probability is summed up with probabilities from both extreme ends, meaning with A less than three and A greater than three that have extremely low probabilities.

Table 2.1. Example Contingency Table Frequencies [17]

	Group 1	Group 2
Class 1	A	B
Class 2	C	D

2.4.1.1. P-values

A calculated p-value is the probability that the resulting contingency table could exist with H_0 and is generally considered significant when less than 0.05 [10].

2.4.1.2. Confidence Intervals

Another output provided by the Fisher's exact test is the confidence interval. With a standard p-value of 0.05, the 95% confidence interval is defined as $(A - B) \pm 1.96\sqrt{\frac{A(1-A)}{A+C} + \frac{B(1-B)}{B+D}}$ [10].

2.4.1.3. Odds Ratio

The odds ratio formula is defined as $\Theta = \frac{A/C}{B/D}$ [3].

[10] shows:

$$\begin{aligned} \Theta = & \quad \frac{A/C}{B/D} \\ = & \quad \frac{AD}{BC} \end{aligned}$$

If Θ equals one, then the contingency table's frequencies are all equal, which would be expected in H_0. Hence, the farther from one Θ is, the closer the clusters discovered are to representing the expected classification.

3. TRANSPORTATION CONCEPTS

3.1. Truck Traffic

3.1.1. Data Collection

The North Dakota Department of Transportation (NDDOT) started installing WIM sites in late 2003. As of now, they have 15 sites in operation with another one to be constructed by late 2015. [26]. Each of the current sites incorporates quartz pierzoelectric sensors that are embedded into U.S. and interstate highways with a nearby controller to collect vehicle transversal data [26] [19]. The traffic data that is collected and stored consists of details related to the vehicle itself as well as information related to the specific transversal. Attributes collected include WIM site, timestamp, vehicle class, gross weight, speed, axle weights, and distances between axles.

The weight data provided by the WIM sites is mainly used for the 20-year equivalent single axle load (ESAL) forecasts that assist the NDDOT Design Division in drafting pavement plans that include a proper base and appropriate pavement type and thickness [26]. The WIM sites are also used by North Dakota Highway Patrol (NDHP) to monitor overweight truck traffic. Patrol officers are able to receive a signal from nearby WIM sites to view real-time transversals on their issued laptops [26].

3.1.2. Selected WIM Sites

WIM sites were selected in the oil-producing region and in other parts of the state to compare truck traffic. WIM sites on U.S. highways were selected to keep the pavement classification consistent between segments. Pavement segments were restricted further by selecting divided highway segments. The WIM sites included were near Williston (U.S. 2), Devils Lake (U.S. 2), Buchanan (U.S. 52), and Washburn (U.S. 83). The Washburn and Buchanan sites are able to collect bi-directional traffic data while the Williston and Devils Lake WIM sites only collect eastbound traffic data. On U.S. 2, the Williston and Devils Lake WIM sites were installed just west of each site's namesake, approximately 250 miles apart. Both sites only collect eastbound traffic data. On U.S. 83, the Washburn site was installed approximately eight miles south of its namesake. On U.S. 52, the Buchanan site was installed approximately four miles north of its namesake.

3.1.3. Truck Traffic Characteristics at WIM Sites

The Federal Highway Administration (FHWA) provides a means of classifying vehicles, which was presented in [12] (Fig. 3.1). Each row represents a vehicle classification and provides vehicle configuration examples for each class. Classes 5-13 include trucks.

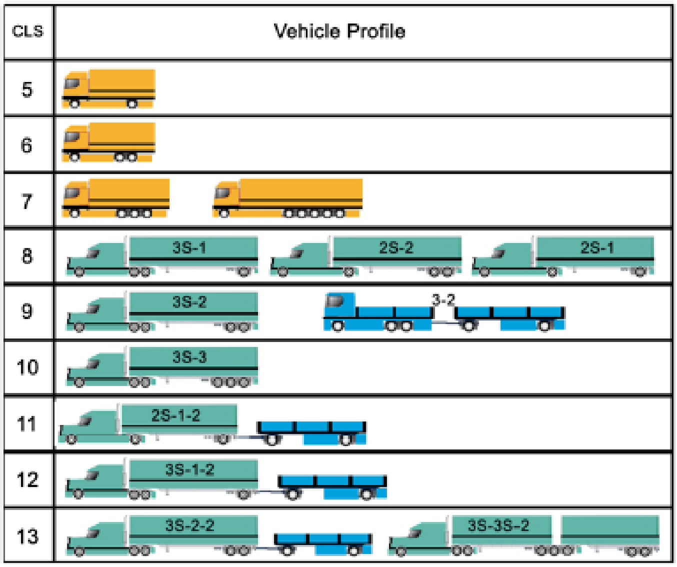


Figure 3.1. FHWA Vehicle Classification

Tables 3.1 and 3.2 depict truck traffic characteristics for class 9 and class 13 trucks. For each site and direction collected, a number of attributes are displayed. The total counts for the specified class, the class percentage of all counts with a steering axle weighing 3.5 kips or more are shown. Used in structural engineering, kips are a unit of measure representing 1,000 pounds-force [2]. Class 9 trucks constitute the largest truck class population at the selected WIM sites, considered overweight when they weigh more than 80 kips [26].

Table 3.1. Site Totals for Class 9 Trucks between 2011 and 2010

Direction-Site	Totals	Class Percentage	Mean Gross (in Kips)	Overweight Percentage
EB Williston	205278	30.4%	48	4.5%
NB Washburn	85994	39.3%	60	10.4%
SB Washburn	91512	38.2%	47	8.8%
EB Devils Lake	75793	42.2%	50	10.1%
EB Buchanan	152735	60.9%	57	6.8%
WB Buchanan	127345	56.3%	53	7.6%

Class 13 trucks, which have seven or more axles, were selected for their higher gross weight potential [1]. 105.5 kips is the maximum legal weight for class 13 trucks, according to [26]. The data from the years 2010 and 2011 was selected for this work due to complications with missing data from other years.

Table 3.2. Site Totals for Class 13 Trucks between 2011 and 2010

Direction-Site	Totals	Class Percentage	Mean Gross (in Kips)	Overweight Percentage
EB Williston	70094	10.4%	67	35.7%
NB Washburn	20263	9.3%	94	69.4%
SB Washburn	24441	10.2%	79	52.0%
EB Devils Lake	10648	5.9%	50	59.2%
EB Buchanan	9657	3.9%	89	64.9%
WB Buchanan	14582	6.4%	85	61.7%

3.2. Pavement Distress

3.2.1. Data Collection

According to the NDDOT, a Pathway Services van drives along the North Dakota highways to collect pavement distress data and imagery [25] [11]. The two primary technologies used on the van include lasers for roughness (IRI) and rutting measurements and cameras to collect surface images [11]. The surface images are taken every 26.4 feet. Before 2013, the first tenth of a mile of each one-mile pavement segment was manually scored by NDDOT employees for different pavement distress types and their severity. In 2013, an automated system was implemented to score whole one-mile pavement segments. NDDOT employees verify the output from the automated scoring system.

The resulting pavement distress data is utilized by the NDDOT in a number of ways. It is used by NDDOT's pavement management system to recommend needed projects through a cost-benefit analysis. The pavement distress data also illustrates road condition for project managers, district engineers, and NDDOT's upper management [25].

3.2.2. Flexible Pavement Distress Types

Fig. 3.2 shows the pavement distress types and advice for manually scoring flexible pavement segments. [25] shared Fig. 3.2 to illustrate how manual scoring for flexible pavement is defined. Each condition is a distress type with additional information on how to score the segment depending on the extent and severity of the distress.

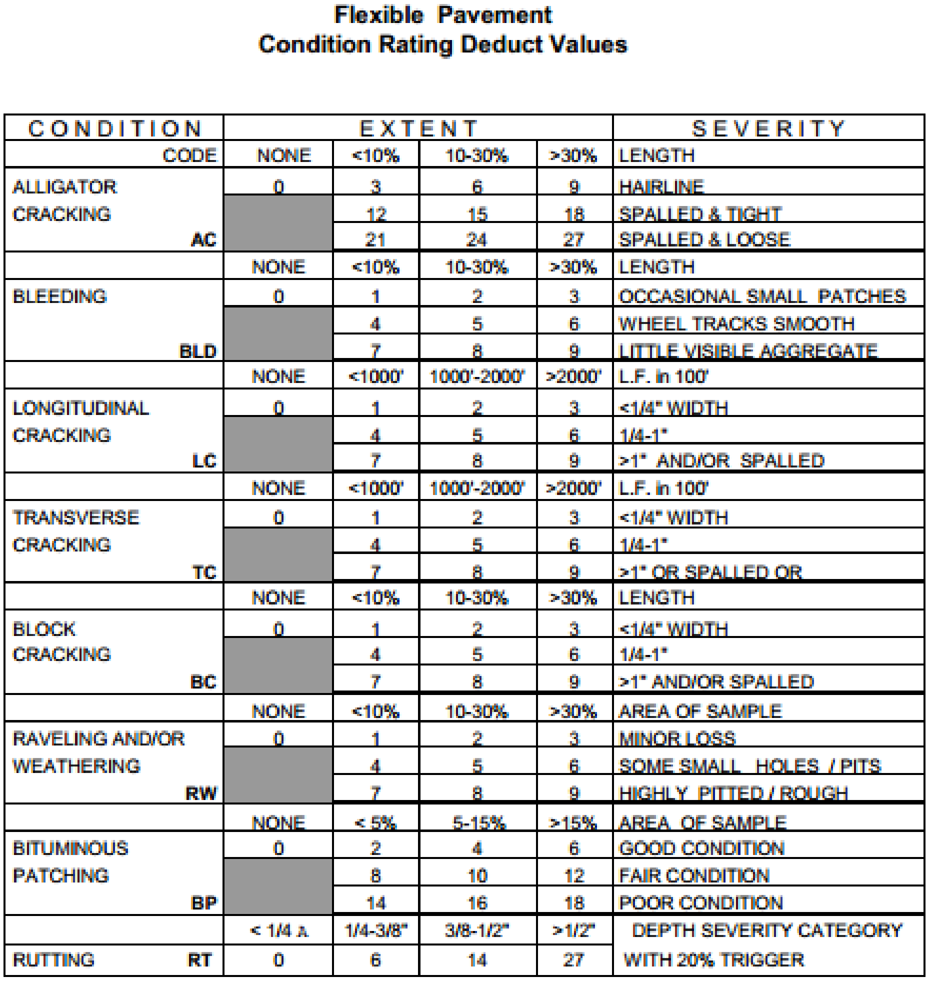
Alligator cracking, known as as fatigue cracking in [15], has a cracking pattern similar to the reptile's skin in its more severe state. It occurs in wheel paths along the pavement. The presence of excessive asphalt binder is known as bleeding and is recognized by wheel path discoloration [15]. Longitudinal cracking runs parallel to the pavement while transverse cracking runs perpendicular [15]. Block cracking is similar to alligator cracking but is more consistently rectangular. It also occurs beyond the wheel paths. Weathering, also known as raveling, occurs when aggregate is lost due to surface wear [15]. Rutting is depression in the wheel paths [15].

Pavement patches created with bituminous material are used as a method of pavement rehabilitation. When the patch area is at or over 0.1 m^2, it is considered a distress type known as bituminous patching [15]. In this work, bituminous patching is excluded from distress types used in the clustering. By excluding it, the algorithm could not be influenced by NDDOT pavement rehabilitation policy. As a result, seven different flexible distress types are being utilized by the hierarchical clustering implementation.

3.3. Linking WIM Sites with Flexible Pavement Sections

3.3.1. Data Set Selection

As stated above, there are a number of determinants concerning pavement deterioration. To focus on truck traffic's role, the other influences had to be considered.The goal was to reduce the number of factors complicating the truck traffic versus pavement distress analysis. Some factors are needed to compare segments and their different truck traffic characteristics. Comparing segments from the same U.S. highway eliminates weather and most construction complexities. Fig. 3.3 highlights in yellow which pavement sections were selected for this analysis. The Williston and Devils Lake sites are along the western section edges while the Buchanan site is on the northern section edge. The Washburn site is near the section's midpoint. Highway numbers, endpoint descriptions, and what lane directions were used to compare pavement distress signatures are detailed in Table 3.3.



Flexible Pavement
Condition Rating Deduct Values

CONDITION	EXTENT				SEVERITY
CODE	NONE	<10%	10-30%	>30%	LENGTH
ALLIGATOR	0	3	6	9	HAIRLINE
CRACKING		12	15	18	SPALLED & TIGHT
AC		21	24	27	SPALLED & LOOSE
	NONE	<10%	10-30%	>30%	LENGTH
BLEEDING	0	1	2	3	OCCASIONAL SMALL PATCHES
		4	5	6	WHEEL TRACKS SMOOTH
BLD		7	8	9	LITTLE VISIBLE AGGREGATE
	NONE	<1000'	1000'-2000'	>2000'	L.F. in 100'
LONGITUDINAL	0	1	2	3	<1/4" WIDTH
CRACKING		4	5	6	1/4-1"
LC		7	8	9	>1" AND/OR SPALLED
	NONE	<1000'	1000'-2000'	>2000'	L.F. in 100'
TRANSVERSE	0	1	2	3	<1/4" WIDTH
CRACKING		4	5	6	1/4-1"
TC		7	8	9	>1" OR SPALLED OR
	NONE	<10%	10-30%	>30%	LENGTH
BLOCK	0	1	2	3	<1/4" WIDTH
CRACKING		4	5	6	1/4-1"
BC		7	8	9	>1" AND/OR SPALLED
	NONE	<10%	10-30%	>30%	AREA OF SAMPLE
RAVELING AND/OR	0	1	2	3	MINOR LOSS
WEATHERING		4	5	6	SOME SMALL HOLES / PITS
RW		7	8	9	HIGHLY PITTED / ROUGH
	NONE	< 5%	5-15%	>15%	AREA OF SAMPLE
BITUMINOUS	0	2	4	6	GOOD CONDITION
PATCHING		8	10	12	FAIR CONDITION
BP		14	16	18	POOR CONDITION
	< 1/4 ʌ	1/4-3/8"	3/8-1/2"	>1/2"	DEPTH SEVERITY CATEGORY
RUTTING RT	0	6	14	27	WITH 20% TRIGGER

Figure 3.2. NDDOT Flexible Deduct Table provides guidance when manually scoring flexible pavement segments.

Table 3.3. Selected Sections Near Selected WIM Sites

Highway	Endpoint Description	Directions
U.S. 2	Williston - US 85 Merge End	East-West
U.S. 2	Devils Lake - Lakota	East-West
U.S. 52	Pingree - Jamestown	East-West
U.S. 83	Wilton - Underwood	North-South

Pavement distress signatures were created when subtracting 2010 pavement distress scores from 2011 pavement distress scores for each one-mile segment in selected highway sections. Segments with at least one of the selected pavement distress scores less than zero were excluded from clustering. Some pavement distress types do not change between 2010 and 2011. The lack of change in specific types may indicate a relationship between pavement distress and truck traffic characteristics.

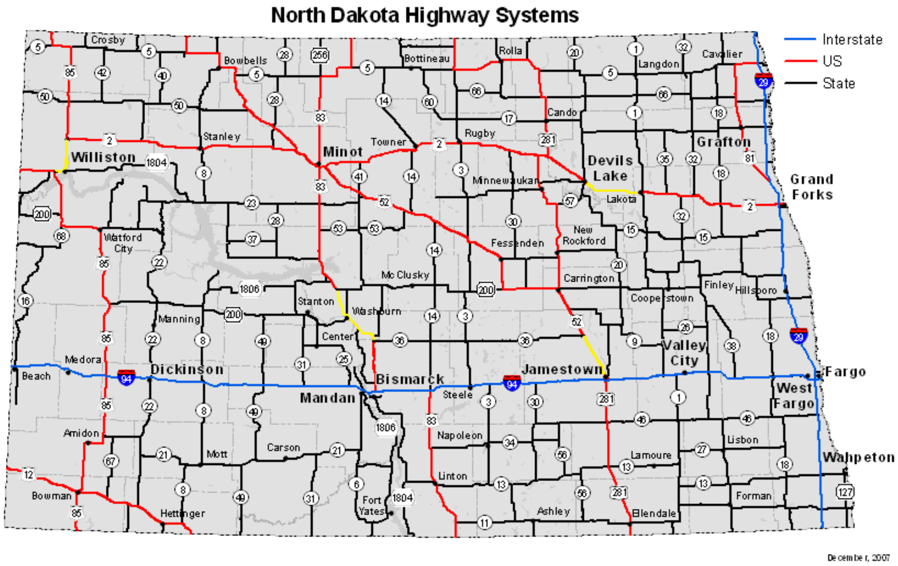


Figure 3.3. Selected Sections in Yellow [16]

One observation made during initial analysis was the apparent pavement distress difference between northbound and southbound U.S. 83 sections. Thus, an additional experiment was conducted to determine whether the northbound and southbound sections would separate into clusters.

4. METHOD

4.1. Distress Signature Generation

The NDDOT-provided pavement distress data was provided as collection year-specific files. Each file was a spreadsheet with each row representing a one-mile segment's pavement distress scores. In this analysis, the 2010 and 2011 score files were converted to comma-separated values (CSV) files.

An R script was created to read in the two files and generated a signature table storing the selected pavement segments and their distress score increases [20]. A vector subset is created for each year, WIM site, and lane direction by selecting the highway, lane direction, reference points limits, surface type as flexible. An intersection between the 2010 and 2011 reference points to ensure the difference between the 2010 and 2011 distress scores is taken between the same mile segments.

A vector is created for each mile if all distress scores are greater than or equal to zero. Values less than zero may occur due to invalid values or some sort of pavement rehabilitation occurring in between collection times. Each vector represents a distress signature for each one-mile segment.

4.1.1. Problem Decomposition with Pavement Segment Subsets

In order to eliminate as many construction-related factors as possible, pavement segments were selected by highway class and surface type. In addition, rehabilitation efforts that are also construction-related were excluded by only selecting segments without any negative pavement distress scores. Environmental factors were eliminated in the experiment set that clustered northbound and southbound Washburn segments.

4.2. Distress Signature Identification

To determine whether a clear relationship exists between pavement distress and truck traffic characteristics, a comparison of pavement distress score increases of selected pavement segments from different areas around the state was needed. Plotting these differences in parallel coordinate plots, a clear distinction between oil-producing regions and non-oil-producing regions was recognized, particularly with pavement rutting score increases. The significant increase in truck traffic near Williston was noted in tables 3.1 and 3.2.

Williston was initially compared with the Ellendale site with similar plots. Ellendale is located near the South Dakota border along U.S. 281. Being farther away from the oil-producing region with less traffic, it was considered a reasonable choice for such a comparison. The initial plots led to more plots with more pavement sections chosen. Discovering noticeable differences, the analysis then continued with the hierarchical clustering implementation with the signature table. The Ellendale section was no longer used in the quantitative experiments as it was not constructed with separated lanes.

4.3. Distance Measures

The pavement distress signatures are vectors in a seven-dimensional space, with each dimension being the selected flexible pavement distress types. For this analysis, the cosine distance was selected in part due to the normalization property provided by the denominator defined by Eq. 2.1, which is the cosine similarity [9]. The cosine similarity itself could have been used as it is cosine distance's inverse. In this analysis, the cosine distance results were compared with Euclidean distance results. The comparison was made for the following reasons: the Euclidean distance is the most popular distance measure and it does measure the size difference between pavement distress signatures [6].

Using the signature table, two distance matrices were created. Using the proxy package, a cosine distance matrix was created [14]. The Euclidean distance matrix was created using dist function in the stats R package [20].

4.4. Using the UPGMA Agglomerative Hierarchical Clustering Method

As noted in Subsection 2.2.1, UPGMA is an agglomerative hierarchical clustering linkage technique. UPGMA was selected for being less responsive to outliers and as middle ground between single linkage and complete linkage.

[20] provided the hclust function in their stats package. By providing a distance matrix and the method (UPGMA), the hclust function provides a tree description. In the stats package, [20] provides the dendrogram class. Plotting or cutting dendrograms are available in this class which were used for results validation [20].

4.5. Results Validation

A 2x2 contingency table is needed by Fisher's exact test for testing significance. Dendrogram branches with outlier segments and segments with no deterioration increase were not used in contingency table creation. The remaining dendrogram portion was used if it represented two major branches. Three main experiments are described in the Quantitative Experiments chapter. Two experiments used data from all four selected WIM sites. The third experiment described in Chapter 6 attempted to identify dendrogram branches as Northbound and Southbound Washburn.

For each experiment, each selected dendrogram branch was processed by a function using R developed by [20]. R provides operations that were used to return the counts for each experiment based on their highway and direction. The output would then populate the contingency table column representing the processed dendrogram branch.

[20] provided the Fisher's exact test with the fisher.test function found in the stats package of R. By providing the contingency table as a parameter, the fisher.test function returns the p-value, 95% confidence interval, and odds ratio. In Chapter 6, the Fisher's exact test results and discussion are located.

5. EXPLORATORY EXPERIMENTS

5.1. Vehicle Re-identification

Several attempts were made to work with the pavement distress score and truck traffic data sets. More of the initial focus was on the truck traffic data. Attempts were made to re-identify vehicles crossing the state. The axle configuration and individual axle weights were used to calculate the root mean square between trucks.

One of the obstacles experienced vehicle re-identification was that the majority of WIM sites were uni-directional for traffic coming in but not leaving the state. Trucks crossing the state may change in weight for a couple of different reasons. A change in load or fuel consumption would change the truck weight. Many truck axle configurations are standard between truck classifications and truck manufacturers so axle spacing alone would result in too many false positives.

It was later determined that the vehicle re-identification classification problem has been successfully considered with Bayesian and neural network approaches. [7] were able to re-identify trucks in an analysis completed by the Oregon Transportation Research and Education Consortium (OTREC) using the same input but also using additional traffic data collectors called AVC sensors. A Bayesian model was applied to match downstream trucks with upstream trucks. Additional processing was necessary to handle trucks entering in between the upstream and downstream sites. [7] compared their Bayesian model with the pre-existing neural network vehicle re-identification approach. Their Bayesian model outperformed the neural network model in all but one case. [7] were also able to verify their results using transponder-equipped trucks in their analysis, seeing 91% accuracy when using both WIM and AVC data.

5.2. IRI Comparison between WIM Sites

Once the OTREC analysis was discovered with such excellent results, interest was redirected to studying the relationship between the NDDOT pavement distress and WIM truck traffic data sets. To confirm a difference between the oil-producing region and elsewhere in the state, international roughness index values (IRI) were compared between years in scatterplots. IRI Pavement sections with or near WIM sites were plotted. Fig. 5.1 and Fig. 5.2 are two example plots showing how the IRI values of the pavement segments near the Williston and Wahpeton WIM sites changed between 2010 and 2011. The red line is a trend line. More of the Wahpeton segments lie along the trend line than in the Williston segment plot, suggesting that there was little change between 2010 and 2011 near the Wahpeton WIM site.

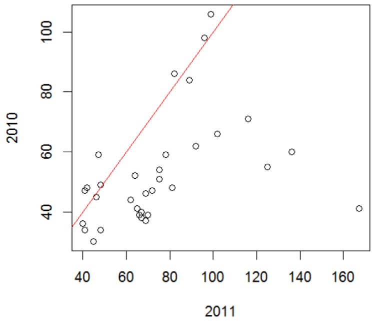


Figure 5.1. U.S. 2 Segments Near Williston WIM Site

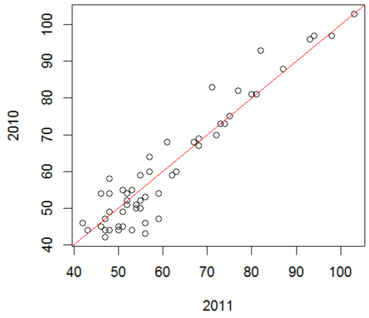


Figure 5.2. I-29 Segments Near Wahpeton WIM Site

A majority of the Williston segments had noticeable IRI increases. With the x-axis representing 2011 and the y-axis representing 2010, many data points have higher x-axis values than y-axis values shows an IRI increase. The Williston 2011 x-axis also has a larger range of values than the Wahpeton 2011 x-axis. The difference in change and value ranges suggests that traffic may be influencing Williston more than Wahpeton. However, it should be noted that environmental differences may exist between Williston and Wahpeton that could also have an influence.

During this work, state highways near the Williston and Wahpeton sites were also plotted with similar results. While the analysis was showing differences between the regions, there are a few points to note. The highways do not have consistent surface types along their length. Over time, projects occur that warrant a different surface type being used. Initially, this was not considered in these plots. It is possible that the data points do not share a common surface type. The comparison between Williston and Wahpeton is skewed by the fact that the Williston WIM site is in U.S. 2 while the Wahpeton segment is in I-29. Interstates and U.S. highways are built to different specifications.

5.3. IRI Prediction Concerning WIM Sites

Recognizing a difference in IRI in different regions, there was interest in determining whether IRI could be predicted using prior IRI history. As part of the work leading up to IRI prediction attempts, the IRI change of pavement sections was plotted in bar plots as in Fig. 5.3 and Fig. 5.4.

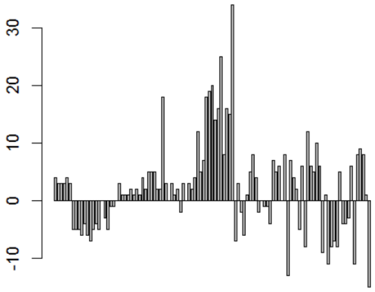


Figure 5.3. IRI Change Between 2012 and 2013 for I-94 Segments Near West Fargo WIM Site

These figures are an example of expanded data selection as the y-axis is measuring IRI change between 2012 and 2013 and Ellendale and West Fargo were evaluated among other WIM sites. Y-axis values below zero indicate an IRI improvement. Such values may indicate a recent rehabilitation. The x-axis represents the road section itself. Each bar is a one-mile segment and the height of the bar is the IRI change experienced by the segment.

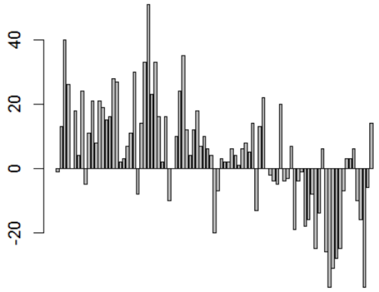


Figure 5.4. IRI Change Between 2012 and 2013 for U.S. 281 Segments Near Ellendale WIM Site

While the different pavement classifications were recognized, the segments were still being considered in the analysis. Different years were being evaluated without regard for recent rehabilitation done to pavement segments. Another adjustment made was the decision to exclude IRI scores from the quantitative experiments. The concern was that if IRI was included in the distress signature definition, its higher variance range may overshadow the other distress types and their contribution.

6. QUANTITATIVE EXPERIMENTS

6.1. Selected WIM Sites Results

Table 6.1 displays a signature table with a subset of WIM site segments. The site information is provided as an aid for the reader as the algorithm does not need that information. The segment labels are used to identify segments in the dendrogram.

Table 6.1. Partial Distress Signature Table. Each signature is identified by their segment mile and nearby WIM site.

Segment	Site	RT	AC	BLD	LC	TC	BC	RW
2 E 12	Williston	0	0	0	3	0	0	0
2 E 13	Williston	9	0	0	3	3	0	0
2 E 14	Williston	9	0	0	0	3	0	0
2 E 15	Williston	4	0	0	0	3	0	0
2 E 16	Williston	4	0	0	0	4	0	0
2 W 271	Devils Lake	0	0	0	0	3	0	0
2 W 272	Devils Lake	0	0	0	0	0	0	0
2 W 273	Devils Lake	4	0	0	0	0	0	0
2 W 274	Devils Lake	4	0	0	0	0	0	0
2 W 275	Devils Lake	0	0	0	0	0	0	0
52 W 253	Buchanan	0	0	0	0	0	0	0
52 W 254	Buchanan	0	0	0	1	3	0	0
52 W 255	Buchanan	0	0	0	1	0	0	0
52 W 256	Buchanan	0	0	0	2	0	0	0
52 W 257	Buchanan	0	0	0	1	3	0	0
83 N 90	Washburn	4	0	0	0	1	0	0
83 N 91	Washburn	4	0	0	0	0	0	0
83 N 92	Washburn	4	0	0	0	1	0	0
83 N 93	Washburn	4	0	0	0	1	0	0
83 N 94	Washburn	4	0	0	0	1	0	0

Fig. 6.1 is a parallel coordinates plot that illustrates the pavement distress signatures found for the selected WIM sites. Red represents Williston segments and green is associated with Washburn segments. Buchanan segments are blue and Devils Lake segments are yellow. The variables displayed are the flexible pavement distress types, previously described in Subsection 3.2.2. In this specific sample, there were no deterioration increases for alligator cracking or weathering (raveling). Only one segment experienced an increase in block cracking. There are less Buchanan and Devils Lake segments selected as

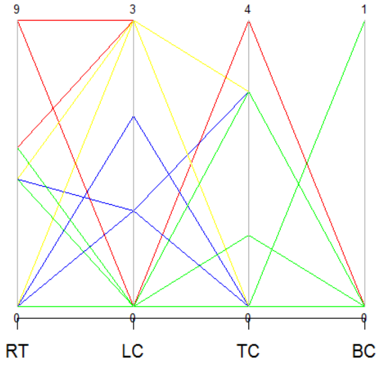


Figure 6.1. Parallel Coordinates Plot of Pavement Distress Signatures for Selected WIM Sites

that section had less sections without a distress decrease between 2010 and 2011 data collection times. As a result, less Buchanan and Devils Lake segments were included in the analysis. As it appears that not many segments were appearing in the plot, a verification was done by organizing the distinct pavement distress signatures and their frequency in Table 6.2. Out of the 20 distinct signatures discovered, 35% of them were only observed once. Over 55% of the selected pavement segments exhibited the distress signature with no distress score increase observed or a rutting score increase of four.

Table 6.2. Distinct Pavement Distress Signatures for Selected WIM Sites

RT	LC	TC	BC	CNT	MEM
0	0	0	1	1	WASH
0	2	0	0	1	BUCH
0	3	3	0	1	DL
4	1	0	0	1	BUCH
4	3	0	0	1	DL
5	3	3	0	1	WILL
9	3	3	0	1	WILL
4	0	4	0	2	WILL
9	0	3	0	2	WILL
0	0	1	0	3	WASH & WILL
0	1	0	0	3	WILL & BUCH
0	3	0	0	3	WILL & BUCH & DL
5	0	0	0	3	WASH & WILL
0	1	3	0	4	WILL & BUCH
5	0	3	0	4	WILL
4	0	3	0	11	WILL
4	0	1	0	13	WASH
0	0	3	0	17	ALL, WILL (53%)
4	0	0	0	31	ALL, SB WASH (52%)
0	0	0	0	58	ALL, SB WASH (58%)
				161	

The most common signature was one with no deterioration increase. For the most part, Williston tended to have the highest rutting increases. It is interesting to note that most of the southbound Washburn segments either had no deterioration increase or a rutting score increased by four.

Fig. 6.2 and Fig. 6.3 depict the leftmost and rightmost dendrogram branches generated by UPGMA clustering calculated with the four WIM site selected segments' cosine distance values. The segment with a block cracking increase and the branch containing segments with no deterioration increase were not included in the statistical analysis.

Fig. 6.2 branch membership largely consisted of Williston, eastbound Devils Lake, and northbound Washburn. Buchanan and Williston were the primary members clustered in Fig. 6.3. In both figures, longer segments can be recognized sequentially which illustrates a general distress trend in longer pavement sections.

There was interest in determining whether the two suggestive branches represented the following categories: west vs. east and Williston vs. others.

6.1.1. Cosine Distance

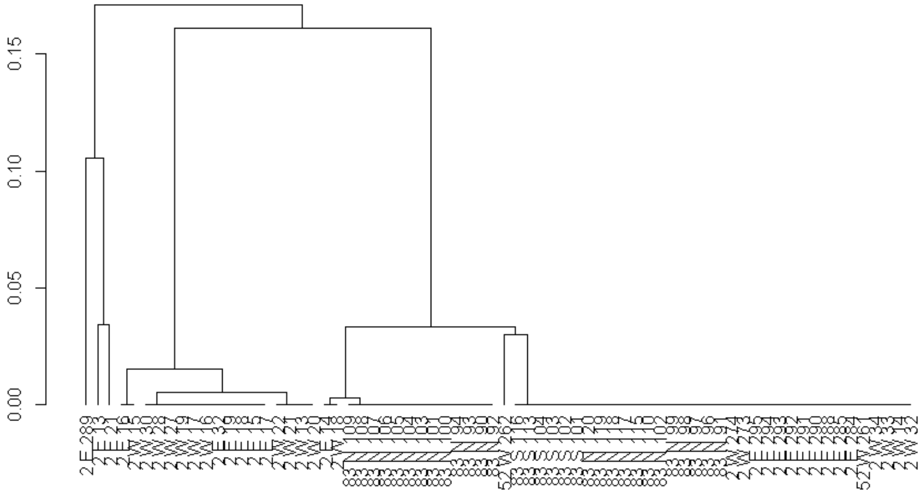


Figure 6.2. Leftmost of Rightmost Branch of Dendrogram (Cosine Distance, Selected WIM Sites)

95 of the 161 segments were located in the two selected branches in Fig. 6.2 and 6.3, with 73.7% segments located in the leftmost branch. This was due to the fact that much of the southbound Washburn segments were not in the selected branches due to 58% them having no deterioration increase. As a result, those segments did not appear in the two branches examined.

6.1.1.1. West vs. East

Table 6.3 is a contingency table showing the segment membership between the left and right selected branches shown by Fig. 6.2 and 6.3 and whether the segments were in the west (near Williston or Washburn) or in the east (near Buchanan or Devils Lake). While the west and east segments are represented in the left branch, the east segments are represented far more in the right. The Fisher's exact test results are displayed in Table 6.4.

Table 6.3. Contingency Table (Cosine Distance, Selected WIM Sites, West vs. East)

	Left	Right
East	39	22
West	31	3

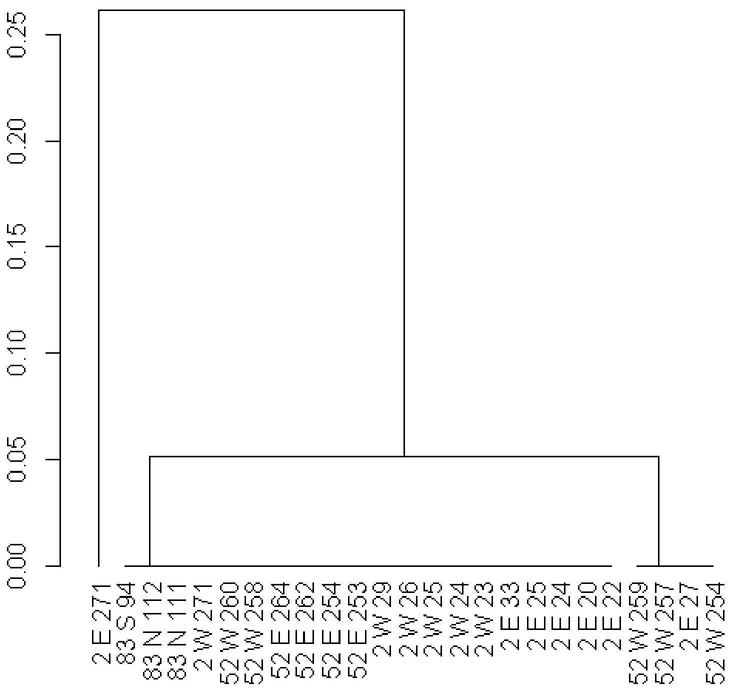


Figure 6.3. Rightmost of Rightmost Branch of Dendrogram (Cosine Distance, Selected WIM Sites)

Table 6.4. Fisher's Exact Test Results (Cosine Distance, Selected WIM Sites, West vs. East)

p-value	95 Percent Confidence Interval	Odds Ratio
0.00363	0.03061532 - 0.66069891	0.1743291

6.1.1.2. Williston vs. Others

Table 6.5 is a contingency table showing the segment membership between the left and right selected branches shown by figures 6.2 and 6.3 and whether the segments were near Williston or not.

Table 6.5. Contingency Table (Cosine Distance, Selected WIM Sites, Williston vs. Others)

	Left	Right
Others	33	12
Williston	37	13

52.6% segments were identified as being near Williston. The Fisher's exact test results for the selected branches are displayed in Table 6.6.

The west vs. east experiment had more statistically significant results when compared to the Williston vs. others experiment when using the cosine distance measure. The results indicate that

Table 6.6. Fisher's Exact Test Results (Cosine Distance, Selected WIM Sites, Williston vs. Others)

p-value	95 Percent Confidence Interval	Odds Ratio
1	0.3507533 - 2.6788390	0.9665521

the west vs. east category set related better to the top dendrogram branches than the Williston vs. others category set when using the cosine distance. The influential factors affected the west region differently than the east region. Due to the geographical distance between sections, more factors may have influenced the difference than overweight truck traffic.

6.1.2. Euclidean Distance

Fig. 6.4 and Fig. 6.5 are the rightmost left and far right dendrogram branches generated with Euclidean distances calculated with segments near Williston, Washburn, Buchanan, and Devils Lake. Fig. 6.4 contained segments largely from southbound Washburn, westbound Devils Lake, and Buchanan. Northbound Washburn, westbound Williston, and eastbound Devils Lake were located in Fig. 6.5.

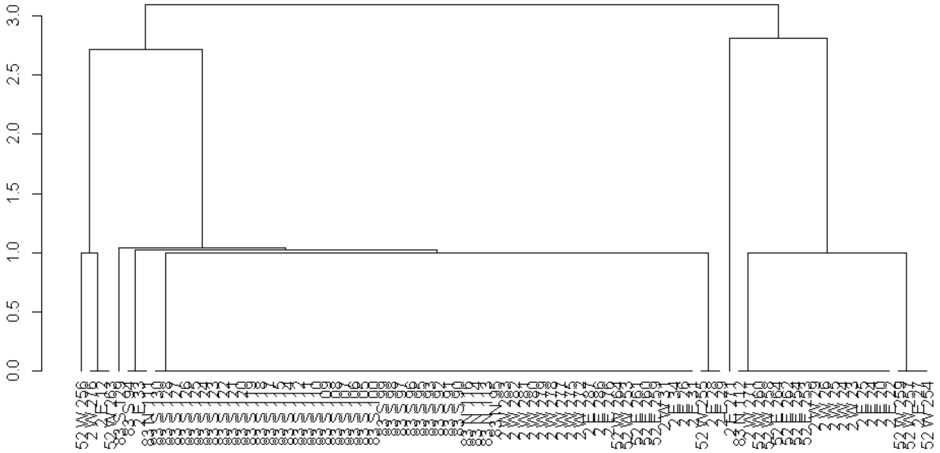


Figure 6.4. Leftmost of Rightmost Branch of Dendrogram (Euclidean Distance, Selected WIM Sites)

158 of the 161 segments are members of the selected branches that are counted in the contingency tables. As a result, over 98% of the segments was contributing to the Fisher's exact test analysis when using the Euclidean distance measure. The three excluded segments were two Williston segments and one Devils Lake segment. 57.5% of the selected segments are in the left branch while 42.4% of the selected segments reside in the right branch.

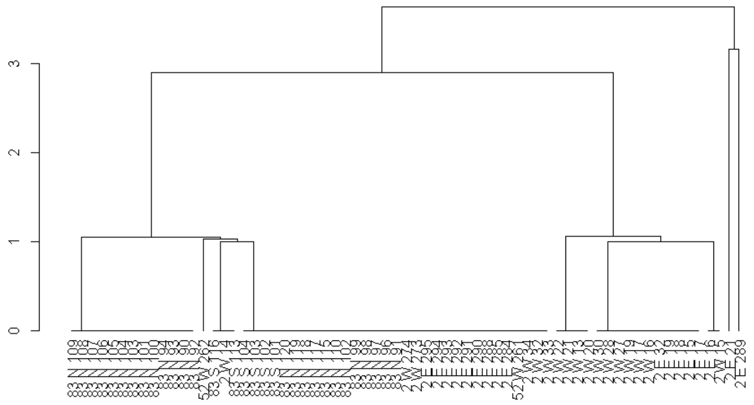


Figure 6.5. Far Right Branch of Dendrogram (Euclidean Distance, Selected WIM Sites)

6.1.2.1. West vs. East

The resulting contingency table when categorizing the left and right branches with the west and east labels is shown in Table 6.7.

Table 6.7. Contingency Table (Euclidean Distance, Selected WIM Sites, West vs. East)

	Left	Right
East	50	36
West	41	31

54.4% segments are categorized in the east while 45.6% are west segments. Table 6.8 displays the Fisher's exact test results for the contingency table represented in Table 6.7.

Table 6.8. Fisher's Exact Test Results (Euclidean Distance, Selected WIM Sites, West vs. East)

p-value	95 Percent Confidence Interval	Odds Ratio
1	0.5305761 - 2.0751284	1.04978

6.1.2.2. Williston vs. Others

As shown in tables 3.1 and 3.2, Williston had the highest increase in both class 9 and class 13 traffic. The eastbound Williston class 9 traffic volume increase was 25.6% higher than the next highest increase, which was eastbound Buchanan. The eastbound Williston class 13 traffic volume increase was 65.1% higher than the next highest increase, which was southbound Washburn. This led to the next category set selection, Williston vs. the other WIM sites (Washburn, Buchanan, and Devils Lake).

The contingency table utilizing this category set is represented by Table 6.9. 42.4% of the selected segments were identified as being near Williston. Compared to the west vs. east experiment, Table 6.10 illustrates that the Euclidean Williston vs. others experiment had a much lower p-value.

Table 6.9. Contingency Table (Euclidean Distance, Selected WIM Sites, Williston vs. Others)

	Left	Right
Others	58	33
Williston	33	34

Table 6.10. Fisher's Exact Test Results (Euclidean Distance, Selected WIM Sites, Williston vs. Others)

p-value	95 Percent Confidence Interval	Odds Ratio
0.07541	0.906787 - 3.616496	1.803932

6.2. Selected WIM Sites Discussion

When cosine distance of the selected WIM sites was used, the west vs. east experiment was more statistically significant than the Williston vs. others experiment. The situation was reversed with the experiments using Euclidean distances of the selected WIM sites.

The less significant of each situation was calculated with a p-value of one. [23] describes that such a p-value means other contingency tables with the same marginal totals are extreme and so all probabilities with the same marginal totals as was observed were included in the Fisher's exact test results. Another point drawn from [23] is that the probabilities calculated are only approximations. Thus, the exclusive p-value range can be inclusive.

The current results are not adequate for comparing the cosine distance and Euclidean distance with this pavement segment selection. The p-values cannot be compared to determine which is better. Using another statistical test may yield more usable results.

6.3. Northbound Washburn vs. Southbound Washburn Results

As distress differences were recognized in prior analysis, another experiment was completed concerning only the northbound and southbound Washburn segments. The reason for this selection was to determine whether the northbound and southbound segments would separate on pavement distress score increases alone. This would reveal truck traffic's influence as the segments have the same weather and should be nearly identical in construction and rehabilitation.

Fig. 6.6 is a parallel coordinates plot that illustrates the pavement distress signatures discovered among for the Washburn segments. Red represents the northbound segments. The variables displayed are the flexible pavement distress types, previously described in Subsection 3.2.2.

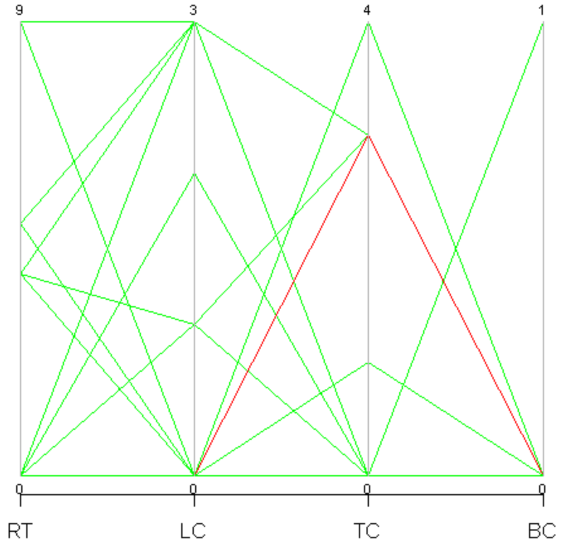


Figure 6.6. Parallel Coordinates Plot of the Northbound vs Southbound Washburn Pavement Distress Signatures

Table 6.11. Distinct Signatures for Northbound and Southbound Washburn

RT	LC	TC	BC	CNT	MEM
0	0	0	1	1	SB
0	0	3	0	1	NB
5	0	0	0	2	SB
0	0	1	0	2	BOTH, 50%
4	0	1	0	13	NB
4	0	0	0	16	BOTH, NB (75%)
0	0	0	0	37	BOTH, SB(89%)
				72	

While 89% of segments with no distress increase were southbound, northbound segments were the majority with the highest rutting as can be seen in Table 6.11. Nearly 74% of the pavement distress signatures identified as the two main groups recognized in the all selected WIM site experiments.

6.3.1. Cosine Distance

Fig. 6.7 depicts the dendrogram generated with Washburn segments by using the cosine distance. Two distinct branches were created though the rightmost branch does have two branches that merge just before the final merge.

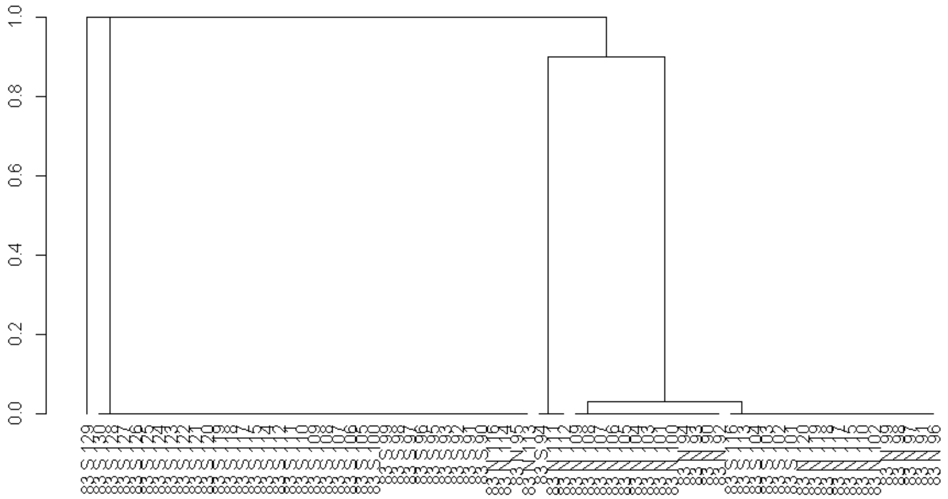


Figure 6.7. Full Dendrogram (Cosine Distance, Northbound Washburn vs. Southbound Washburn)

The balanced distribution of northbound to the rightmost branch and southbound to the leftmost branch is illustrated by the Table 6.12. Table 6.13 shows the Fisher's exact test results which show significance between the northbound and southbound segments.

Table 6.12. Contingency Table (Cosine Distance, Northbound Washburn vs. Southbound Washburn)

	Left	Right
Northbound	4	27
Southbound	33	7

Table 6.13. Fisher's Exact Test Results (Cosine Distance, Northbound Washburn vs. Southbound Washburn)

p-value	95 Percent Confidence Interval	Odds Ratio
3.06e-09	0.006413212 - 0.135925923	0.03387017

6.3.2. Euclidean Distance

Fig. 6.8 depicts the dendrogram generated with Washburn segments by using the Euclidean distance. While two distinct branches were created, there were additional groups discovered. While the leftmost branch has two noticeable clusters, the rightmost branch has one primary cluster. There does appear to be some variance in distress types due to that.

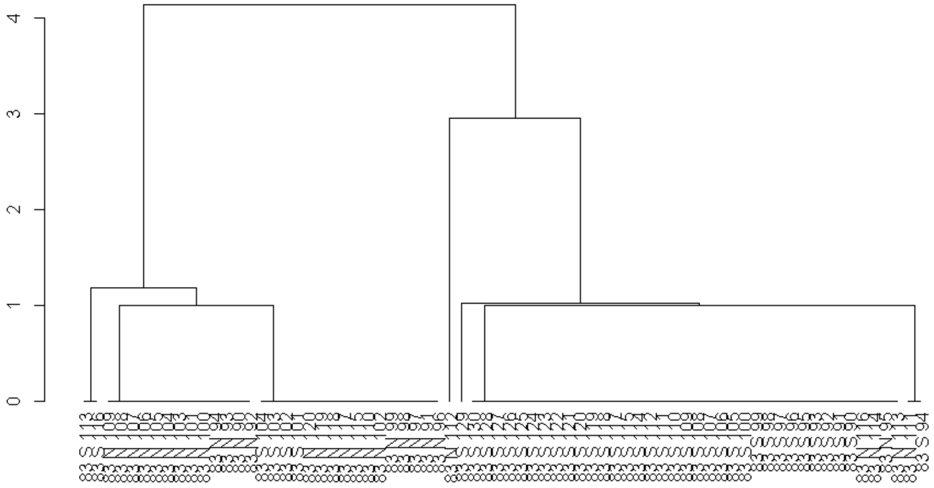


Figure 6.8. Full Dendrogram (Euclidean Distance, Northbound Washburn vs. Southbound Washburn)

The balanced distribution of northbound to the rightmost branch and southbound to the leftmost branch is illustrated by the Table 6.14. Table 6.15 shows the Fisher's exact test results which show significance between the northbound and southbound segments with the p-value. Both the 95% confidence interval and odds ratio consist of relatively large values.

Table 6.14. Contingency Table (Euclidean Distance, Northbound Washburn vs. Southbound Washburn)

	Left	Right
Northbound	25	6
Southbound	6	35

Table 6.15. Fisher's Exact Test Results (Euclidean Distance, Northbound Washburn vs. Southbound Washburn)

p-value	95 Percent Confidence Interval	Odds Ratio
1.59e-08	6.135429 - 102.315666	22.74258

6.4. Northbound Washburn vs. Southbound Washburn Discussion

The test results for the cosine distance implementation output a slightly smaller p-value than the Euclidean distance results. As previously stated in Subsection 2.4.1, the smaller the odds ratio, the closer the clusters discovered are to representing the expected classification. Thus, the cosine distance provides a slightly more significant and expected clustering than the Euclidean distance in this experiment.

7. CONCLUSIONS AND FUTURE WORK RECOMMENDATIONS

7.1. Conclusions

The Washburn segment subset experiment showed the most influence due to truck traffic on pavement distress. With the two lane directions experiencing the same precipitation and freeze-thaw cycles on the same soil being constructed and rehabilitated in similar ways, the main variables remaining are the truck volume and weight.

While eastbound Williston did have the largest volume, northbound Washburn had the highest overweight percentage both both class 9 and class 13 trucks between 2010 and 2011. This coincides with Rys etal in their assertion that overweight vehicles increased pavement distress significantly.

[22] had found that rutting occurred more with more axle configurations with three or more axles in a group while cracking was impacted more by single and tandem axles.

According to Tables 3.1 and 3.2, Buchanan, Devils Lake, and Washburn all have over a third more overweight class 13 truck traffic than Williston. Buchanan and Devils Lake have the highest percentage of class 9 trucks, which consist of a front steering axle and two tandem axles.

Both Williston and Washburn each had approximately 20% of the pavement segments with a rutting score increase. While Washburn's part coincides with what was determined by [22], the Williston part does not. The initial expectation in this analysis was to see more pavement distress increases near Williston due to the nearby oil production. The increased truck volume near Williston may be partially explained by the oil production.

It could also explain Washburn's volume and overweight increases. Washburn is near the southern edge of the oil-producing region. It may have an increase in overweight volumes due to less enforcement. As Williston is near the epicenter of oil production, they may be much more weight limit enforcement.

While Buchanan and Devils Lake had the overweight class 9 truck traffic increases, they were not represented well in the distress signatures with surface cracking increases. If they were, that would have agreed with the work demonstrated in [22].

One reason this could be the case is that there was an imbalance in the segment counts for each WIM site. In particular, Buchanan and Devils Lake was less represented as many of the originally selected miles were filtered out due to distress score decreases. This may be due to recent rehabilitation that occurred between the 2010 and 2011 collection times. This has not been verified with the NDDOT but it could explain the decrease.

While the northbound and southbound Washburn segments have been compared, the Devils Lake segments also showed a distinct separation between the eastbound and westbound segments. Additional analysis is needed to better determine what specific overweight truck traffic characteristics are influencing pavement distress increases and how it is occurring. Additional work concerning the inclusion and exclusion of the environmental, construction-related, and traffic-related factors could also be done.

7.2. Future Work Recommendations

7.2.1. Additional Experiment Comparisons

The international roughness index (IRI) and bituminous patching could be included in the distress signatures to compare those results with the ones described in this analysis. Another comparison that could be made would be with only rutting score increases. As noted by [8] and [22], NMDOT and MDOT have composite indices that could be compared with this analysis's results. Such a comparison could validate the claim made in Chapter 1 that composite indices may lack accuracy due to their weighting factors becoming out-of-date.

In this analysis, some partial work was done in comparing the seven distress type signature with the four distress types used by the NDDOT to trigger rehabilitation in their cost-benefit analysis system [25]. The four distress types include rutting, alligator cracking, transverse cracking, and bituminous patching. It would be recommended to select another time period and pavement section set as this set had very little change in both alligator cracking and bituminous patching. Another comparison was partially done comparing the seven distress type signature with the three distress types deemed significant according to an ANOVA analysis that was done with the selected data. The significant distress types discovered were longitudinal cracking, rutting, and IRI.

7.2.2. Additional Experiment Considerations

As with [8], GIS could be utilized with the hierarchical clustering results to increase spatial understanding. In the generated dendrogram, it was recognized that nearby pavement segments tended to cluster together as their distress increased at the same rate. [8] recognized independence among the eight typical flexible pavement distress types. It was acknowledged that DOTs need to continue evaluating all eight pavement distress type scores. It would be interested to determine whether the independence noted in New Mexico would be seen in North Dakota. Additional knowledge about the relationship between pavement distress and weather may be discovered.

Including the significant pavement distress signature with no distress score increase would be an option. Though there is no increase, it may explain the difference between pavement segments with score increases and those without such increases. Prior rehabilitation of the segments could also be considered in future work as in [8]. [8] also claimed that pavement deterioration could be modeled as a Markov process as it is most influenced by the prior year's deterioration levels. It would be fascinating to validate that assertion with existing North Dakota data.

As for the significance testing to validate the clustering results, an alternative to Fisher's exact test would be Barnard's test. [13] does describe improvements that may overcome the inconclusive results noted in the selected WIM sites experiments. [22] considered other vehicle classes such as classes 6, 7, 8, 10, and 11. Please see Fig. 3.1 for vehicle class configurations. NDDOT does collect passenger vehicle traffic data and it may interesting to consider its impact. Other characteristics that could be included are average daily truck traffic (ADTT) and cumulative truck traffic (CTT), both of which were considered by [22].

REFERENCES

[1] Truck characteristics analysis. Technical report, Federal Highway Administration, Washington, D.C., 1999.

[2] kip. http://sizes.com/units/kip.htm/, 2011.

[3] Alan Agresti. A survey of exact inference for contingency tables. *Statistical Science*, pages 131–153, 1992.

[4] Per Ahlgren, Bo Jarneving, and Ronald Rousseau. Requirements for a cocitation similarity measure, with special reference to pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology*, 54(6):550–560, 2003.

[5] Sue Ahn, Srivatsav Kandala, J Uzan, and Mohamed El-Basyouny. Impact of traffic data on the pavement distress predictions using the mechanistic empirical pavement design guide. *Road Materials and Pavement Design*, 12(1):195–216, 2011.

[6] Morven Leese Daniel Stahl Brian S Everitt, Sabine Landau. *Cluster Analysis*. Wiley, 2011.

[7] Mecit Cetin and Christopher M Monsere. Exploratory methods for truck re-identification in a statewide network based on axle weight and axle spacing data to enhance freight metrics: Phase ii. 2012.

[8] Cong Chen, Su Zhang, Guohui Zhang, Susan M Bogus, and Vanessa Valentin. Discovering temporal and spatial patterns and characteristics of pavement distress condition data on major corridors in new mexico. *Journal of Transport Geography*, 38:148–158, 2014.

[9] Hinrich Schütze Christopher D. Manning, Prabhakar Raghavan. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

[10] Gerard V Dallal et al. *The little handbook of statistical practice*. Gerard V. Dallal, 1999.

[11] NDDOT Communication Division. Pathways van, 2012.

[12] Graziano Fiorillo and Michel Ghosn. Procedure for statistical categorization of overweight vehicles in a wim database. *Journal of Transportation Engineering*, 2014.

[13] John Ludbrook. Analysis of 2× 2 tables of frequencies: matching test to experimental design. *International journal of epidemiology*, 37(6):1430–1435, 2008.

[14] David Meyer and Christian Buchta. *proxy: Distance and Similarity Measures*, 2015. R package version 0.4-15.

[15] John S Miller and William Y Bellinger. Distress identification manual for the long-term pavement performance program. Technical report, 2014.

[16] North Dakota Department of Transportation. North dakota highway systems. https://www.dot.nd.gov/imgs/nd-highway-sys-lrg.png, 2007. File: nd-highway-sys-lrg.png.

[17] K. J. Preacher and N. E. Briggs. Calculation for fisher's exact test: An interactive calculation tool for fisher's exact probability test for 2 x 2 tables. http://quantpsy.org/fisher/fisher.htm, 2001. Accessed November 6th, 2015.

[18] Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc., 2013.

[19] Rich Quinley. Wim data analyst's manual. Technical report, 2010.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[21] Dawid Rys, Jozef Judycki, and Piotr Jaskula. Analysis of effect of overloaded vehicles on fatigue life of flexible pavements based on weigh in motion (wim) data. *International Journal of Pavement Engineering*, (ahead-of-print):1–11, 2015.

[22] Hassan K Salama, Karim Chatti, and Richard W Lyles. Effect of heavy multiple axle trucks on flexible pavement damage using in-service pavement performance data. *Journal of Transportation Engineering*, 132(10):763–770, 2006.

[23] Steve Simon. Can the p-value actually equal 1.0? http://www.pmean.com/06/PvalueEqualsOne.asp, 2010.

[24] Hao Wang, Jingnan Zhao, Zilong Wang, et al. Impact of overweight traffic on pavement life using weigh-in-motion data and mechanistic-empirical pavement analysis. In *9th International Conference on Managing Pavement Assets*, 2015.

[25] Stephanie Weigel. Private Communication, 2011–2015.

[26] Terry Woehl. Private Communication, 2011–2015.