Development of Cost-Effective Sensing Systems and Analytics (CeSSA) to Monitor Roadway Conditions and Mobility Safety

January 2021

A Research Report from the Pacific Southwest Region University Transportation Center

Chun-Hsing Ho, Northern Arizona University

Dada Zhang, Northern Arizona University

Jaiwei Gao, Northern Arizona University

Marco Gerosa, Northern Arizona University

Bertrand Cambou, Northern Arizona University

Pacific Southwest Region UTC
University Transportation Center

NORTHERN ARIZONA UNIVERSITY

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. PSR-19-12	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Development of Cost-Effective Sensing Systems and Analytics (CeSSA) to Monitor Roadway Conditions and Mobility Safety		**5. Report Date** January 2021 **6. Performing Organization Code** N/A
7. Author(s) Chun-Hsing Ho https://orcid.org/0000-0002-6690-4403 Dada Zhang Jaiwei Gao Marco Gerosa Bertrand Cambou		**8. Performing Organization Report No.** PSR-19-12
9. Performing Organization Name and Address METRANS Transportation Consortium University of Southern California 650 Childs Way, RGL 216 Los Angeles, CA 90089-0626		**10. Work Unit No.** N/A **11. Contract or Grant No.** USDOT Grant 69A3551747109
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final report (Jan. 2020 – Jan. 2021) **14. Sponsoring Agency Code** USDOT OST-R
15. Supplementary Notes Project webpage: https://www.metrans.org/research/-development-of-cost-effective-sensing-systems-and-analytics-cessa-to-monitor-roadway-conditions-and-mobility-safety		
16. Abstract The project presents a pavement sensing system along with a list of promising computing models that can be used to predict pavement conditions using a vehicle-based sensing technology. The project started with data acquisition obtained from the previous field data collection followed by a series of data computing using machine learning methods to determine a promising computing algorithm. Subsequently, statistical analyses were performed to evaluate the effect of sensor placements/locations within a vehicle on the accuracy of pavement condition assessments. Based on analysis results, random forest algorithm is the best fitting machine learning algorithm than other three algorithms (Linear Regression, Support Vector Machine, and Neural Network) for the pavement condition assessment. It is also found that the pavement temperature significantly influences the number of significant points (pavement distress) provided the fact that the number of significant points decrease during cold weather condition while the number of significant points increase as the pavement temperature is getting warmer. The Time-Series analysis indicates the number of the significant points will increase quickly in the following two years, which indicate that the pavements will be deteriorated if the maintenance and rehabilitation will not be scheduled.		
17. Key Words Pavement conditions, sensor technology, machine learning, road maintenance		**18. Distribution Statement** No restrictions.

19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 48	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Pacific Southwest Region UTC
University Transportation Center

Table of Contents

About the Pacific Southwest Region University Transportation Center

The Pacific Southwest Region University Transportation Center (UTC) is the Region 9 University Transportation Center funded under the US Department of Transportation's University Transportation Centers Program. Established in 2016, the Pacific Southwest Region UTC (PSR) is led by the University of Southern California and includes seven partners: Long Beach State University; University of California, Davis; University of California, Irvine; University of California, Los Angeles; University of Hawaii; Northern Arizona University; Pima Community College.

The Pacific Southwest Region UTC conducts an integrated, multidisciplinary program of research, education and technology transfer aimed at improving the mobility of people and goods throughout the region. Our program is organized around four themes: 1) technology to address transportation problems and improve mobility; 2) improving mobility for vulnerable populations; 3) Improving resilience and protecting the environment; and 4) managing mobility in high growth areas.

U.S. Department of Transportation (USDOT) Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Disclosure

Principal Investigator, Co-Principal Investigators, others, conducted this research titled, "Development of Cost-Effective Sensing Systems and Analytics (CeSSA) to Monitor Roadway Conditions and Mobility Safety" at Department of Civil Engineering, Construction Management, and Environmental Engineering, School of Informatics, Computing, and Cyber Systems, Northern Arizona University. The research took place from 01/06/2020 to 01/05/2021 and was funded by a grant from the US Department of Transportation in the amount of $91,538. The research was conducted as part of the Pacific Southwest Region University Transportation Center research program.

Acknowledgements

The authors would like to express their gratitude to Dr. Yongqi Li, Manager, Pavement Management Section at Arizona Department of Transportation for his opinions and advice on the data collection and results.

Abstract

The project presents a pavement sensing system along with a list of promising computing models that can be used to predict pavement conditions using a vehicle-based sensing technology. The project started with data acquisition obtained from the previous field data collection followed by a series of data computing using machine learning methods to determine a promising computing algorithm. Subsequently, statistical analyses were performed to evaluate the effect of sensor placements/locations within a vehicle on the accuracy of pavement condition assessments. Based on analysis results, random forest algorithm is the best fitting machine learning algorithm than other three algorithms (Linear Regression, Support Vector Machine, and Neural Network) for the pavement condition assessment. It is also found that the pavement temperature significantly influences the number of significant points (pavement distress) provided the fact that the number of significant points decrease during cold weather condition while the number of significant points increase as the pavement temperature is getting warmer. The Time-Series analysis indicates the number of the significant points will increase quickly in the following two years, which indicate that the pavements will be deteriorated if the maintenance and rehabilitation will not be scheduled.

Development of Cost-Effective Sensing Systems and Analytics (CeSSA) to Monitor Roadway Conditions and Mobility Safety

Executive Summary

Pavement condition surveys normally involve data acquisition, interpretation, and documentation. Automated pavement condition survey is considered one of most commonly used methods for pavement condition assessments. However, as of today, the automated pavement condition surveys have not yet been well adopted by highway agencies as a promising system due to the fact that a well-established dynamic vibration and sensor models that can accurately capture the signatures of pavement surface condition have not been proposed and widely adopted by transportation authorities due to its costly expenses on the equipment and software that an agency has to invest at front. In recognition of the immediate need, the project is presented to develop cost effective sensing technology and advances knowledge in the field by creating computing algorithms using pavement vibration responses as inputs to analyze and filter raw data to (1) determine promising computing models for prediction of pavement distresses and (2) evaluate the promising/optimum placements/locations of sensor loggers within a vehicle, and (3) provide an affordable method that will benefit state, city, county governments, as well as local communities who have an immediate need but with limited budgets to evaluate the road quality and prioritize repair needs.

Due to the pandemic, the implementation of the research was modified to best reflect the recommendations provided by local health organizations (Coconino Health Department, Northern Arizona University Health Services). All scheduled field trips for data collections in the summer of 2020 were cancelled. However, we were able to acquire vibration data collected between 2017-2018 on the I-10 corridors in Phoenix as input for computing purposes. The report started with introduction of pavement sensing setup and field data collecting activities (Chapter 1). After data were retrieved, all data were analyzed against their validity to be used for condition assessments using the time series function as cluster analysis to exclude data that was not statically relevant (Chapter 2). Four computing algorithms were used including random forest (RF), linear regression (LR), support vector machine (SVM), and neural network (NN). Among the four computing methods, the RF is considered the best fitting machine learning algorithm than other three algorithms.

Subsequently, we performed a series of statistical analyses including ANOVA, probability-probability (P-P) Plot, quantile-quantile (Q-Q) Plot, and Cumulative Distribution Function (CDF) plot to help improve the fitting process (Chapters 3-4). The objectives of these analyses are to evaluate the effect of sensor placements and speed of a vehicle on the accuracy of pavement distress identifications (good, fair, and poor) and determine threshold to identify a level of

pavement distress. The results indicate that the thresholds vary based on statistical analysis. The threshold obtained from the sensor inside the vehicle (M5) exhibited lower values than the other four sensors. This is due to the placement of M5 being inside the vehicle as compared with four sensors being placed on top of control arm. Thus, it is recommended all five sensors should be used simultaneously to be accurately predict pavement conditions. Additionally, the pavement temperature significantly influences the number of significant points (pavement distress) provided the fact that the number of significant points decrease during cold weather condition while the number of significant points increase as the pavement temperature is getting warmer. We also used The Time-Series analysis to fit the model for predicting the number of significant points for the next two years. The analysis result showed that the predicted number of the significant points will increase from the forecast plot in the following two years, which means that the pavements will be deteriorated if the maintenance and rehabilitation will not be scheduled.

The cost-effective sensing systems and analytics (CeSSA) presented in the report indicate that the CeSSA is cable of capturing pavement vibration patterns and determining a level of pavement distress.

Introduction

Pavement condition surveys normally involve data acquisition, interpretation, and documentation. These activities characterize surface condition, such as surface cracking, deformation, and other surface defects for both flexible and rigid pavements. Currently, there are three key major pavement distress detecting techniques: 1) manual inspection, 2) imaging process detection, and 3) vibration-based detection. In 2008, Erikson et al. (1) and Mohan et al. (2) further expand the vibration mode to mobile sensing system and GPS in smartphones to detect and report the surface conditions of roads. This system uses the inherent mobility of a vehicle to gather data from vibration and GPS sensors, and then process the data to assess road surface conditions. To build a complete automated imaging system, highway agencies have to commit to a significant amount of up-front investment, in addition to equipment upgrade expenses afterward. Another pavement detection system is performed through the measurement of vibration data using accelerometers attached on a vehicle (3-5). As of today, the automated pavement condition surveys have not yet been well adopted by highway agencies as a promising system due to the fact that a well-established dynamic vibration and sensor models that can accurately capture the signatures of pavement surface condition have not been proposed and widely adopted by transportation authorities. Thus, there is a need to further advance the development of pavement sensing technology and methodology.

Sensor technology has been used by highway agencies in pavement condition surveys. Current methods of automated sensing detection system in support of pavement condition monitoring are limited and outdated, particularly with respect to predictive accuracy, reliable dynamic range in vibration, and calibration control synthesis to achieve promising performance-based results. This project supports fundamental research to the vibration modeling of pavement conditions impacted by varying vehicle parameters and signature extraction based on intelligent sensing algorithms to estimate real-time pavement conditions in support of rapid decision-making. When traveling on highways, a vehicle equipped with sensors on board should integrate dynamic vibration effect associated with sensors taking into account for a whole system that consists of (i) a vibration model with 3 dimensional components in x, y, and z directions along with varying vehicle speeds and weights that collect vibration responses based on road roughness condition generated by an exogenous dynamical contact between tires of the vehicle and the road surface, and (ii) a multi-domain signal processing model that effectively transfers dynamic vibration data in time, spectral and time-frequency domains and extract signatures of pavement condition after adaptive filtering and machine-learning process. At present, such integrated dynamic vibration and sensor systems have not yet established as a result of lacking reliable theory support and actual practice. It should be noted that due to the pandemic, all filed data collection scheduled for the summer of 2020 was cancelled. Thus, all vibration data used for the project was from previous field work collected on the I-10 corridors in Phoenix between 2017 to 2018.

The project advances knowledge in the field by creating computing algorithms using pavement vibration responses as inputs to analyze and filter raw data to (1) determine promising computing models for prediction of pavement distresses and (2) evaluate the promising/optimum placements/locations of sensor loggers within a vehicle, and (3) provide an affordable method that will benefit state, city, county governments, as well as local communities who have an immediate need but with limited budgets to evaluate the road quality and prioritize repair needs.

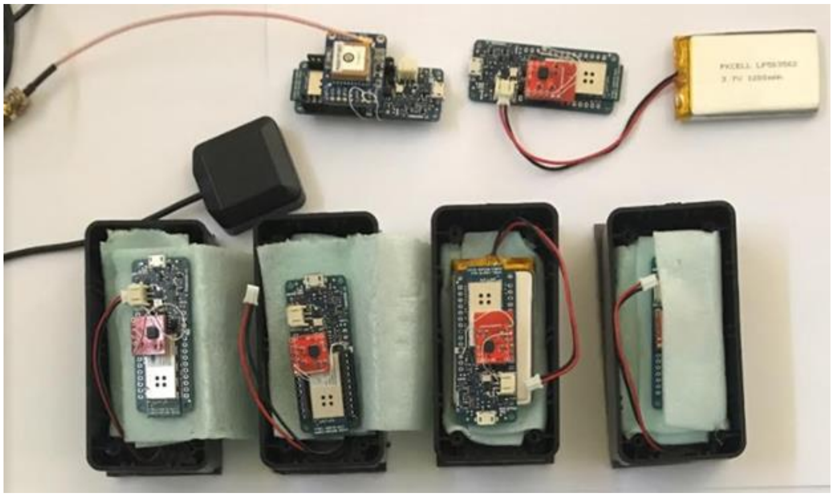
Chapter 1 Data Acquisition

This chapter explains how vibration data was collected through vehicle-based sensing system developed by the research team.

Introduction to pavement sensing system

A sensor logger consisting of ADXL 335 triple-axis accelerometers, Arduino MKR1000 computer boards, GPS, and a battery was designed (Figure 1.1) for field data collection. The sensors communicate three-dimensional vibration data to a laptop via a WiFi router. A 2016 Honda Accord was used through the entire experiment because it was newly purchased by the Northern Arizona University (NAU), so all mechanical system still remained in an excellent condition. Furthermore, using the same vehicle for all road testing could keep the suspension and the body mass of the vehicle in a constant way so that the effect of the damping system and body mass of a vehicle on the data collection can be neglected. After reviewing the vehicle, it is determined that all sensor loggers will be placed on top of the control arms. The decision is based on the fact that the control arms are responsible for connecting a vehicles suspension to its frame and provide a flat surface to mount sensors. The configuration of sensor loggers is believed to make the vibration data collected from road testing more direct, reliable and accurate than the ones obtained by a smartphone.

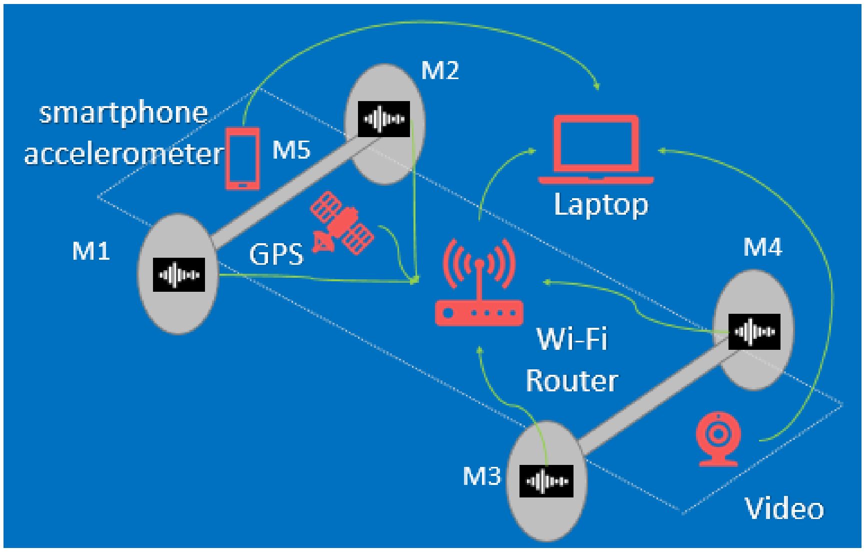
Figure 1.1. Components of sensor logger



Road testing

A year-round road test on I-10 corridors in Phoenix, Arizona was conducted between February 2017 to February 2018 to monitor the resilience of pavement conditions. Prior to travelling, vehicle-based accelerometers were mounted to the vehicle (Figure 1.2), one on each wheel's control arm (M1-M4), one inside cab of the vehicle (M5), and a sixth iPhone sensor mounted inside the cab of the vehicle. The iPhone sensor was used to compare the accuracy of vibration data with the vehicle based sensors. The conceptual framework of the system is shown in Figure 1.2. Accelerometers located by each wheel transmits data wirelessly through a WiFi router to an on-board laptop where all data will be staved and stored in the computer. GPS is linked to each accelerometer for data to be georeferenced. Smartphone data is transmitted separately from the accelerometer sensors so allowing the team to compare their effectiveness on road conditions and monitor the resilience of pavement roughness.

Figure 1.2. Layout of sensor placements

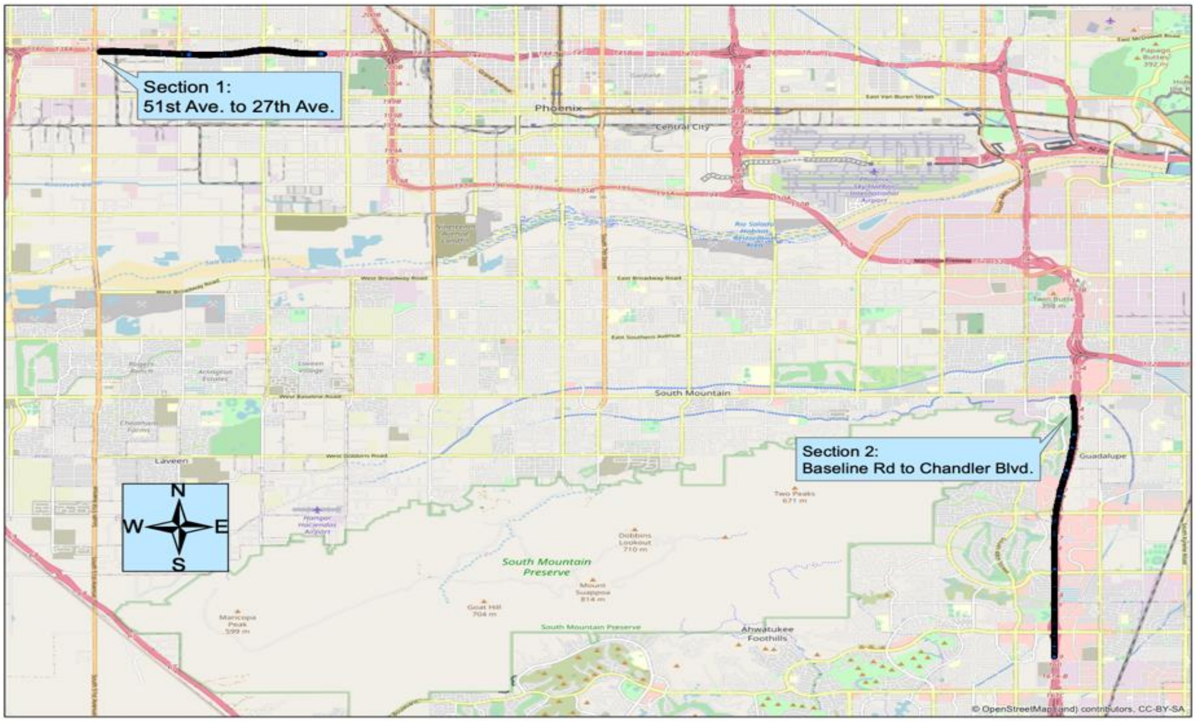


Two road testing sections on Interstate 10 corridors were chosen in Phoenix, Arizona (Figure 1.3). The two sites on Interstate 10 were also chosen based on heavy traffic volume, differing pavement roughness, and relatively straight passages. The first corridor (Section 1), 27th Avenue through 51st Avenue, and the second section, Baseline Road through Chandler Boulevard. In 2016, the average annual daily traffic (AADT) including both directions were approximately 186,000 for the Baseline Ave. – Chandler Blvd. section and 230,000 for the 27th Ave. – 51st Ave. section.

The 1st and 2nd right lanes were surveyed going east and west bound directions for a total of four times per study corridor. The target testing speed was 95-kph (60-mph) as to remain within a safe speed while testing. Because traffic congestion was an issue in these areas, tests were mostly conducted around midnight to avoid traffic distraction. During data collecting, a pavement temperature was measured by an infrared thermometer and recorded for each road test section. The testing period covered four seasons to ensure a year-round pavement temperatures ranging from 4°C - 66°C (40°F - 150 °F) were recorded.

After completion of each road testing, vibration data was extracted from the laptop for analysis. The data output is in acceleration of gravity in the x, y, and z-directions, and GPS information. The z-direction corresponds to the vertical motion of the wheel caused by a bump or change in slope of the road, the x-direction corresponds to the vehicle accelerating and braking, and the y-direction corresponds to the vehicle turning left or right. However, when a car passes over rough pavement the sensor on the wheel's control arm can shake in all directions, not just in the z-direction.

Figure 1.3. Testing sections on the I-10 corridors in Phoenix, AZ



Chapter 2 Development of Computing Algorithms

The focus of this chapter is on how an advanced machine learning algorithm was developed to predict road conditions. Our work comprised five parts: establishing a database, hierarchical clustering, investigation of a resampling method, classifier construction, and cross-validation.

Establishing a Database

The machine learning algorithm requires training from a validated database. We built this database based on vibration data collected from March of 2017 through February of 2018 in the I-10 corridors in the Phoenix region as mentioned in Chapter 1. Three accelerator values (x-axis, y-axis, and z-axis) for each accelerator sensor and two values (latitude and longitude) for GPS were collected from each trip. All vibration signatures were analyzed accordingly using the algorithms to be discussed later in the following sections. Based on results, the degrees of pavement conditions (good, fair, and poor) were determined as indicated by Ho et al. (6).

Hierarchical Clustering

To exclude the influence of high correlation time series on the model results, we conducted a cluster analysis using the time series function for 15 features. According to Afyouni et al. (7), if the time series contains autocorrelation, the standard error of the sample correlation coefficient will be biased. Thus, we calculated the autocorrelation coefficient (ACF) for each feature before applying the hierarchical clustering analysis. Table 2.1 indicates that the ACF of each time series is close to zero, which means that the features are not interdependent.

The hierarchical clustering analysis is an unsupervised machine learning method where it can automatically generate clusters according to the dataset characteristics and it does not need to pre-define the number of clusters. Figure 2.1 shows the results of the hierarchical clustering analysis that indicates that there are some clusters' correlation coefficients greater than 0.7. Thus, the result states that only one of these time series with a high coefficient is needed to represent the data characteristics in these clusters (8). Therefore, we exclude some duplicated features, including Z1, Z2, Z3, X4, and Y4 from the original dataset. Table 2.2 shows the change after the hierarchical clustering analysis is performed.

Table 2.1. The Autocorrelation Coefficient of Features

Variable Name	Autocorrelation Coefficient (ACF)
X1	0.0293
Y1	0.0259
Z1	0.0200
X2	0.0278
Y2	0.0446
Z2	0.0219
X3	0.0248
Y3	0.0330
Z3	0.0216
X4	0.0285
Y4	0.0354
Z4	0.0173
X5	0.0913
Y5	0.1491
Z5	0.0419
Average	0.0408

Figure 2.1: Hierarchical Clustering Analysis

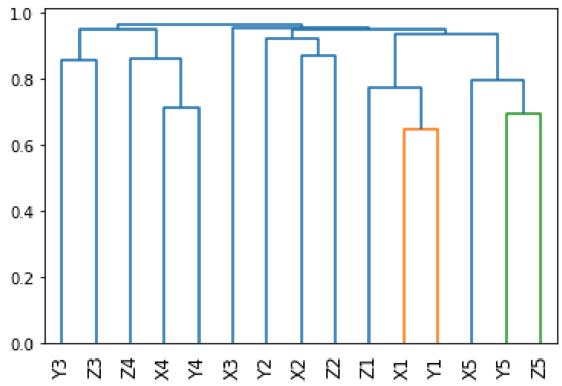


Table 2.2: Chosen Time Series

	The Input Features
The Original Dataset	X1 Y1 Z1 X2 Y2 Z2 X3 Y3 Z3 X4 Y4 Z4 X5 Y5 Z5
After Hierarchical Clustering	X1 Y1 X2 Y2 X3 Y3 Z4 X5 Y5 Z5

Resampling Method

We calculated the number of data points for each class in the dataset and the result are in Table 2.3. As we can see, the dataset is significantly unbalanced. If the predictive model used an unbalanced dataset, the accuracy would be an impropriate measure because the model will be biased towards the majority class (9). To overcome this problem, we used the resampling method as an unsupervised machine learning method to pre-process the dataset. The resampling method is a way to increase or decrease data points of a class according to the pattern of the dataset. There are three kinds of resampling methods, including the oversampling method, the undersampling method, and the combination method (10). To compare the performance of the three different resampling methods, the dataset was split in 70% for training and 30% for testing and a neural network model was used to perform the computations. In the combination method, we doubled the number of poor-type data points and half-cut the number of good-type data points. The results as shown in Table 2.4 were obtained using the Python libraries scikit-learn (0.23.1) and imbalanced-learn (0.7.0).

Table 2.3. Summary of the original dataset

	Poor	**Fair**	**Good**	**Total**
Num.	6	95	4834	4935

Table 2.4. The Performance of the Resampling Methods

		Poor	**Fair**	**Good**
Oversampling Method	**Num.**	4834	4834	4834
	Precision	0.99	0.97	1.00
	Recall	0.97	0.99	1.00
Undersampling Method	**Num.**	6	6	6
	Precision	0.00	1.00	0.67
	Recall	0.00	0.33	1.00
Combination Method	**Num.**	12	95	2417
	Precision	0.98	0.75	0.63
	Recall	0.99	0.60	0.46

According to the results in Table 2.4, the combination method could achieve the best performance. However, if we focus on the precision and recall scores, the oversampling method has the highest score. Thus, both scores are close to one in the oversampling method, which means there is a concern on over-fitting. As for the undersampling method, it directly decreases the sample number of each class into six. As we cannot train the model based on less data points, we selected the combination method for the subsequent computations.

Classifier Construction

To select the most suitable machine learning method for the determination of pavement conditions, we implemented four commonly used machine learning models: Logistic Regression

(LR), Support Vector Machine (SVM), Random Forest (RF), and Neural Network (NN). In this section, we introduce the principles of these machine learning methods.

Logistic Regression (LR)

Logistic Regression uses a linear prediction function to predict the probability that the result of observation data i is k. The formula is as follow:

$$f(k,i) = \beta_{(0,k)} + \beta_{(1,k)}x(1,i) + \beta_{(2,k)}x_{(2,i)} + \cdots + \beta_{(M,k)}x_{(M,i)} \qquad (2.1)$$

Where $\beta_{(m,k)}$ is a regression coefficient, which indicates how much the m-th observation variable affects the k-th result.

When we regard $\beta_{(0,k)}$ as the term of $x_{(0,i)} = 1$, we can yield the vectorized form of Eq. 2.1 as shown in Eq. 2.:

$$f(k,i) = \beta_{(k)} \cdot x_{(i)} \qquad (2.2)$$

where $\beta_{(k)}$ is a regression coefficient vector, which represents the influence degree or importance of the observation values represented by the observation vector $x_{(i)}$ on the result k.

After training the model using the training dataset, regression coefficients were defined. Thus, the model can predict the class (k) of the new input according to the input values (xi).

Support Vector Machine (SVM)

The second method we investigated was SVM, which is a linear model to create a line or a hyperplane to be able to separate the data into classes. To improve SVM from a binary classifier to a multi-class classifier, we adopted the one-vs-one strategy (OVO). In this strategy, we replaced the m-class classifier with m(m-1)/2 binary classifiers. To accomplish the OVO strategy, the training group dataset was divided into *m* data groups according to classes at first. Then, SVM was performed for every two data groups to obtain m(m-1)/2 binary classifiers. When we wanted to predict the class of a new sample, the input data of the new sample was calculated by classifiers to form a matrix as shown in Eq. 2.3 (11).

$$R = \begin{pmatrix} - & r_{12} & \cdots & r_{1m} \\ r_{21} & - & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & - \end{pmatrix}_{m \times m} \qquad (2.3)$$

Eq. 3.3 is an m by m matrix. Each r is assigned to be either 0 or 1, representing the result of a binary classifier. The prediction result can be obtained according to the matrix, such as the voting OVO (V-OVO) strategy (12).

Random Forest (RF)

The third method was Random Forest, which is an ensemble learning algorithm. It generates many individual learners and obtains results from them. Random Forest uses the bagging approach as a computing method, meaning that each classifier is built individually according to a bootstrap sample of the input data. Comparing to the regular decision tree classifier, the parameter at each node is a randomly selected number of features instead of using all of the features. This action helps reduce the correlation between the feature attributes (12).

Neural Network (NN)

The last method is NN. There are many kinds of Neural Networks that can be used for analysis. We employed the multilayer perceptron neural networks (MLPNNs). MLPNN is one of the most commonly used neural network. It has fast operations and small training set requirements (14). There are three layers including the input layer, the hidden layer, and the output layer. The hidden layer processes and transmits the input to the output layer. Eq. 2.4 expresses the basic principle of the neural network (14).

$$y_j = f\left(\sum w_{ji}\, x_i\right) \quad (2.4)$$

In Eq. 2.4, each neuron j in the hidden layer multiplies a corresponding input parameter x_i by a weight w_{ji} and then sum each product. Finally, f is a function to calculate the final result according to the sum.

Eq. 2.5 expresses the error of the neural network algorithm, i.e., the difference between the actual output of the training dataset and the predicted output.

$$E = \frac{1}{2}\sum_j \left(y_{pj} - y_j\right)^2 \quad (2.5)$$

During the construction of the neural network model, an important process is to reduce the overall error to an acceptable range or reach the preset maximum number of iterations by modifying the weight w_{ji}. We use the L-BFGS algorithm (15). This algorithm has the advantages of fast and high accuracy for datasets with small data volume.

Cross-validation

We employed the stratified 5-fold cross-validation method to assess the models. Stratified K-fold cross-validation is a criterion to choose and evaluate the prediction model (16). It effectively reduced bias and improved predictive accuracy to the relatively small size dataset (17). To be more specific, the stratified k-fold cross-validation method randomly divides all data samples into k groups and performs k rounds of training and testing. In each round, one of the groups is used as the test group and the other k-1 groups are used as training sets. It is worth to mention that the folds are made by the same percentage of samples for each class (7) to fit to an unbalanced dataset situation. Finally, the evaluation results of k rounds are combined as the final evaluation of a model.

Analysis and comparison of computing models- Classifier Performance

Based on the models developed previously, a series of analyses on the different algorithms were performed and the results are discussed in the following sections including classifier performance and evaluation. The classifier performance introduces the performance matrix used to evaluate and compare the four different models while the classified evaluation states the result of the performance matrixes and highlights the important points.

We calculated the Matthews Correlation Coefficient (MCC) and the evaluation matrix representing the performance of each classifier. Comparing to other statistical measures on imbalanced datasets, MCC is a more reliable statistical rate representing the whole situation (18). We also built an evaluation matrix to enable a more detailed comparison, including metrics such as precision, recall, f1-score, and accuracy. They are calculated based on True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) values (19). The relationship between these four terms is shown in Table 2.5. Accuracy represents the average correctness of the model. A high precision indicates that the algorithm is predicting the right condition most of the time. A high recall designates that the majority of the specific road type was correctly predicted. F1-score indicates the average of precision and recall. Equations 2.6 to 2.9 explain the formulas to obtain these metrics.

Table 2.5. Summary of the original dataset

The Prediction Situation / The Actual Situation	Positive	Negative
Positive	True Positives (TP)	False Negatives (FN)
Negative	False Positives (FP)	True Negatives (TN)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.6)$$

$$Precision = \frac{TP}{TP+FP} \qquad (2.7)$$

$$Recall = \frac{TP}{TP+FN} \qquad (2.8)$$

$$F1\ Score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \qquad (2.9)$$

Analysis and comparison of computing models- Classifier Evaluation

After training the four different models, the performance results were yielded and shown in Tables 2.6. to 2.9. According to the MCC, the following results were obtained: RF(0.7654) > NN(0.5826) > SVM(0.5161) > LR(0.2122). RF (Random Forest) had the highest MCC. Table 2.6 shows the performance of the LR model, for which the overall accuracy is 0.96, an acceptable number. But for the "fair" label, the precision and recall are zero, which means the LR model cannot figure out the fair type in our dataset. Apart from that, the F1-score of the poor label is 0.5, which is relatively low. Table 2.7 indicates that the SVM model has a high precision and a low recall for the "fair" label. It means the samples determined by the SVM model as fair type are all fair type, but there are still many fair samples that have not been determined right. As can be seen in Table 2.8, in the RF model, all the evaluation parameters are higher than 0.75, which means that the accuracies of prediction are acceptable for each class. There are some issues for the other three machine learning algorithms. Table 2.9. shows that the NN model has a similar situation to the SVM model. It cannot predict the data points labeled as "Fair" as well. Based on the results from Tables 2.6. through Table 2.9., it is determined that RF has the highest accuracy and the best performance overall than the other three algorithms.

Table 2.6. The Performance of the Logistic Regression Model

MCC: 0.2122		**Accuracy:** 0.96	
	Precision	**Recall**	**F1-score**
Good	0.96	1.00	0.98
Fair	0.00	0.00	0.00
Poor	1.00	0.33	0.50
Average	0.65	0.44	0.49

Table 2.7. The Performance of the Support Vector Machine Model

MCC: 0.5161		**Accuracy:** 0.97	
	Precision	**Recall**	**F1-score**
Good	0.97	1.00	0.98
Fair	1.0	0.21	0.35
Poor	1.0	0.67	0.8
Average	0.99	0.63	0.71

Table 2.8. The Performance of the Random Forest Model

MCC: 0.7654		**Accuracy:** 0.98	
	Precision	**Recall**	**F1-score**
Good	0.99	1.00	0.99
Fair	0.86	0.63	0.73
Poor	1.00	0.67	0.80
Average	0.95	0.77	0.84

Table 2.9. The Performance of the Neural Network Model

MCC: 0.5826		**Accuracy:** 0.97	
	Precision	**Recall**	**F1-score**
Good	0.98	0.99	0.99
Fair	0.64	0.37	0.47
Poor	0.50	1.00	0.67
Average	0.71	0.79	0.71

Chapter 3 Distribution Fitting and ANOVA Tests to Analyze Pavement Sensing Patterns for Condition Assessments

Introduction

Pavement deterioration can be due to the traffic loading and/or environment factors. The periodic pavement maintenance is necessary to ensure a satisfactory pavement condition performance so that the operational safety of highways can be achieved. Pavement condition assessment is essential for pavement asset management and it involves data collecting, pattern detecting, and condition monitoring processes. Generally, the International Roughness Index (IRI) is the most commonly used assessment method to evaluate pavement deterioration through the values that calculated from each segment of highways. The higher IRI values represent the rougher road condition and are likely to have an immediate need for maintenance. However, the costs of IRI equipment including instruments and software equipped in a vehicle appear not to be affordable for all governments and organizations. Thus using acceleration vibrations to monitor and estimate pavement condition becomes one of common methods among public agencies and institutions who has an urgent need but with limited budget to measure road conditions.

The acceleration vibrations collected from smartphones or self-developed sensors are widely used in today's condition assessments (6, 20-22). One of challenges in this type of pavement sensing method is: how to properly determine the thresholds of pavement conditions using currently available methodologies including recurrent neural network, derivate and integral methods, fast Fourier transform, machine learning, and statistical methods. For example, the vertical acceleration can be used to predict the IRI values through statistical models such as multivariate liner regression (23), or identify pavement condition using recurrent neural network. The crack damage can also be determined via the frequency of fast Fourier transform and the numerical analysis based on the acceleration signals peaks (24). In addition to these numerical analysis methods, the paper presents an approach using regression analysis and probability distribution fitting to analyze pavement sensing patterns and signals collected on the I-10 corridors in Phoenix, Arizona. A vehicle is equipped with four sensors placed on the top of the control arms of the vehicle and one sensor is inside of the vehicle to gather the data for analysis.

The main objectives of Chapter 3 are 1) to optimize the use of vehicle based sensors for future pavement sensing work; and 2) to determine the thresholds of pavement conditions (i.e., good, fair, and poor) as a criterion to represent the pavement deterioration, so the maintenance and rehabilitation can be scheduled to ensure continuous quality improvement.

Literature Review

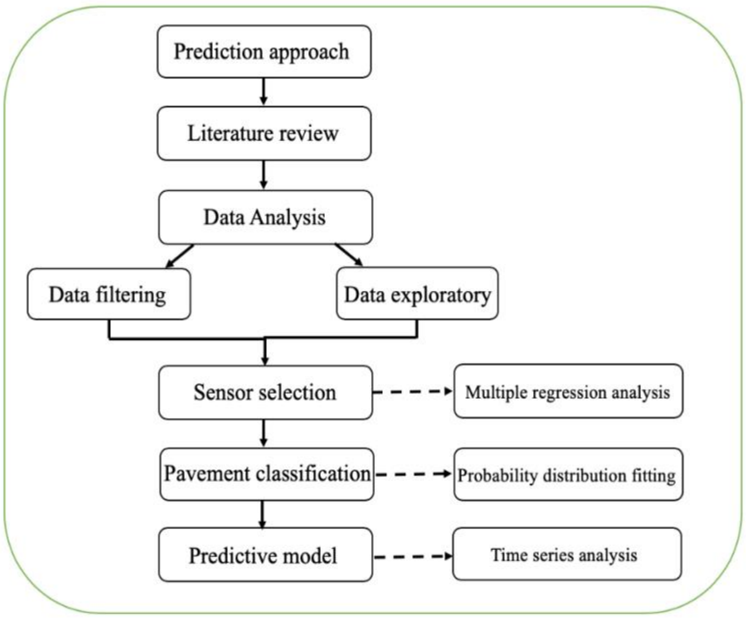
Road roughness is a commonly used parameter in evaluating pavement conditions and International Roughness Index (IRI) is a pavement condition index that is generally used to conduct pavement management assessments. The IRI values are obtained using equipment and computing algorithms for each roadway segment (20-25) with defined specifications in different countries (26). Due to the limited budget and the classifications of roadways, scholars and road agencies have searched for another approaches to monitor road conditions such as analyzing acceleration vibration of roads and using algorithms to predict IRI values. Thus, instead of computing IRI values, the application of acceleration vibrations in pavement condition assessments has become a popular and common way for pavement maintenance. Currently, a number of numerical analyses have been used to provide information on the levels of different pavement conditions such as machine learning, mathematics, and statistical models based on the data collected from smartphones or self-developed sensors (27-30). The mobile applications from smartphones are common devices that can be used to collect vibration data for road asset management with low costs (23, 31). The study of smartphones and pavement detection by Douangphachanh and Oneyama shows that the accelerometers have a linear relationship with road roughness (21). The self-developed sensors are another option for acquiring accurate acceleration vibration with features of speed, courses, and time. Chen et al. used multiple sensors to detect surface cracks based on accelerometers from smartphones and acceleration sensors that mounted on the vehicle (22). The method of using multiple sensors and smartphones for pavement detection are also introduced by Ho et al. and they found the vehicle-based sensors are more reliable than the iPhone sensor (6).

The mathematical algorithms of derivative and integral are general applied in analyzing accelerometers, velocity and displacement to evaluate pavement performance (32). Also, Ye et al. built numerical models to find the potential roadway conditions with acceleration extrema and frequency distribution based on veridical acceleration signals for analysis of pavement vibrations (33). Currently, machine learning and statistical models are widely used in condition assessments to detect and predict the condition of pavements (34). Padarthy et al. developed an algorithm from machine learning techniques for detecting potholes and road asset management by processing lateral acceleration and speed (35). Combing statistical models with machine learning is another way to perform pavement condition assessments. The methods of analytic hierarchy process and fuzzy logic theory are applied in project prioritization and condition assessments where an evaluation of pavement condition can be achieved (36). The methods of Bayesian, Time Series Analysis and Monte Carlo Markov Chain (MCMC) are considered as part of numerical approaches to simulate pavement performance and improve roughness prediction (37, 38). However, as mentioned previously, the methodologies used to determine the thresholds of pavement distress conditions are still not well studied. In addition, to the extent of the authors' previous work (6), the number of sensors in a vehicle needed for pavement sensing also requires further investigation. Therefore, the project presents a methodology using regression analysis and distribution fitting method that can be used to analyze the acceleration response of multiple vehicle-based sensors and then to determine the levels of pavement deterioration for condition assessments.

Methodology

We take multiple ways to analyze the pavement sensing patterns and signals for condition assessments using statistical models. Figure 3.1 illustrates the steps for road condition assessment based on the data collected from previous research. As shown in Figure 3.1, the multiple regression analysis is used for sensor selection in order to find the most significant relative sensor in pavement distress detections among sensors mounted on a vehicle. Then, the fitted distribution model is used to find significant points that would be considered as the pavement determination to classify the pavement condition. Finally, the time-series models are fitted to predict the number of significant points for the future years.

Figure 3.1. Methodology of chapter



Data Collection and Specifications

As previously mentioned in Chapter 1, all pavement sensing patterns and signals were collected by a 2016 Honda Accord that was purchased by Northern Arizona University. Four sensors (named as M1, M2, M3, and M4) were placed on the top of the control arm with another one (named as M5) being placed inside of the vehicle to gather vibration data simultaneously. A GoPro was attached on either the front or the rear of vehicle to validate the occurrence of cracks, construction joints, detection loop, or road reflectors from sensing patterns and signals. The pavement temperature was recorded by infrared thermometer and the driving speed was maintained at 60 miles per hour except the traffic congestion occurred in road testing. M1 and M3 are sensors placed on the left side of vehicle while M2 and M4 are sensors placed on the right side. Sensor M5 was placed on top of cap inside of the vehicle. The details about vehicle-based sensors and the data collection process are explained in reference (6).

Data Analysis

The acceleration vibrations from pavement sensing signals are used for pavement condition assessment. The vibration magnitudes of x-axis and y-axis represent the acceleration response when a vehicle made a turn and acceleration-stop (39), and of z-axis is used to detect a bump or a crack, which is generally used to detect pavement roughness in research projects (33). In the past, many research papers primarily analyze vertical accelerations for detecting the pavement conditions, but the paper further explores an approach using a resultant from all three directions and defined as "M" for the road condition assessments. The "M" is called total magnitude by combining all acceleration values together due to the sensors are placed on the control arm were affected by multiple direction in motion of a vehicle moving, which is based on the author's previous work (6). The total magnitude, M can be computed as follow (6,29):

$$\boldsymbol{M} = \sqrt{\boldsymbol{x}^2 + \boldsymbol{y}^2 + \boldsymbol{z}^2} \qquad (3.1)$$

where x, y, and z represent the acceleration vibration along three axes of x, y and z. The higher magnitude reflected on the corresponding patterns indicates the pavement distress is significant and needs an immediate attention.

Data Filtering

The filtering process is performed using geographic information systems (GIS) software to ensure the pavement sensing patterns and signals are properly displayed and matched on I-10 corridors in a GIS map to avoid any errors. It should be noted that the data was collected from March 2017 to February 2018 excluding July of 2017 due to unexpected technical issue during the sensing process (6). Before analyzing the collected data, the pavement sensing patterns and signals should be displayed on the arterial map as shown in Figure 1.3.

Data Exploratory Analysis

Using visualization and transformation to explore all magnitude values from data appeared to be an appropriate method and model for further analysis. As mentioned above, sensors M1 to M4 were placed on the control arms and M5 was placed on the cap of the vehicle. Figure 3.2 and Figure 3.3 show the boxplot of magnitude values from all five sensors in two sections. The data appear right-skewed (positively skewed) distributions based on the kernel density plots as well as contain high outliers as shown in Figure 3.4 and Figure 3.5. To meet the assumptions of normality in statistical models, data transformation is needed. In statistics, the data shows the positively skewed distribution or the negatively skewed (or left-skewed) distribution, the general data transformation includes logarithm, square root, cubic root, and Box-Cox Power (40). Any zero magnitude values were eliminated to avoid the error when the data is transformed in a logarithmic scale.

Figure 3.2. Boxplots of Magnitude values from 51st Ave. through 27th Ave.

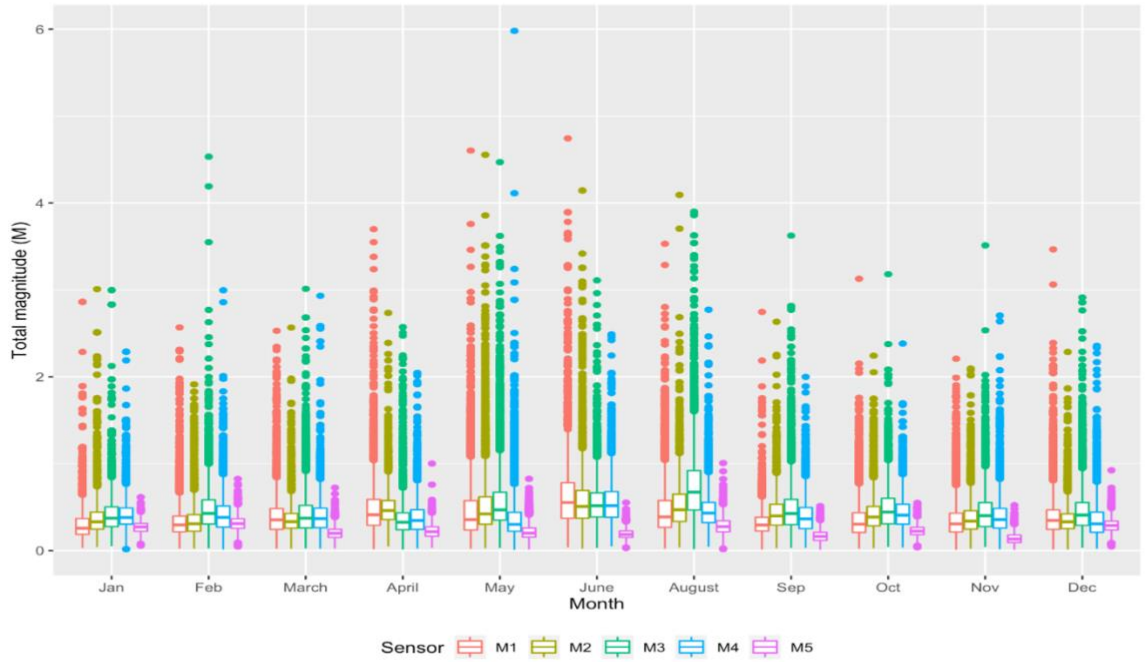


Figure 3.3. Boxplots of Magnitude values from Baseline Rd. through Chandler Blvd.

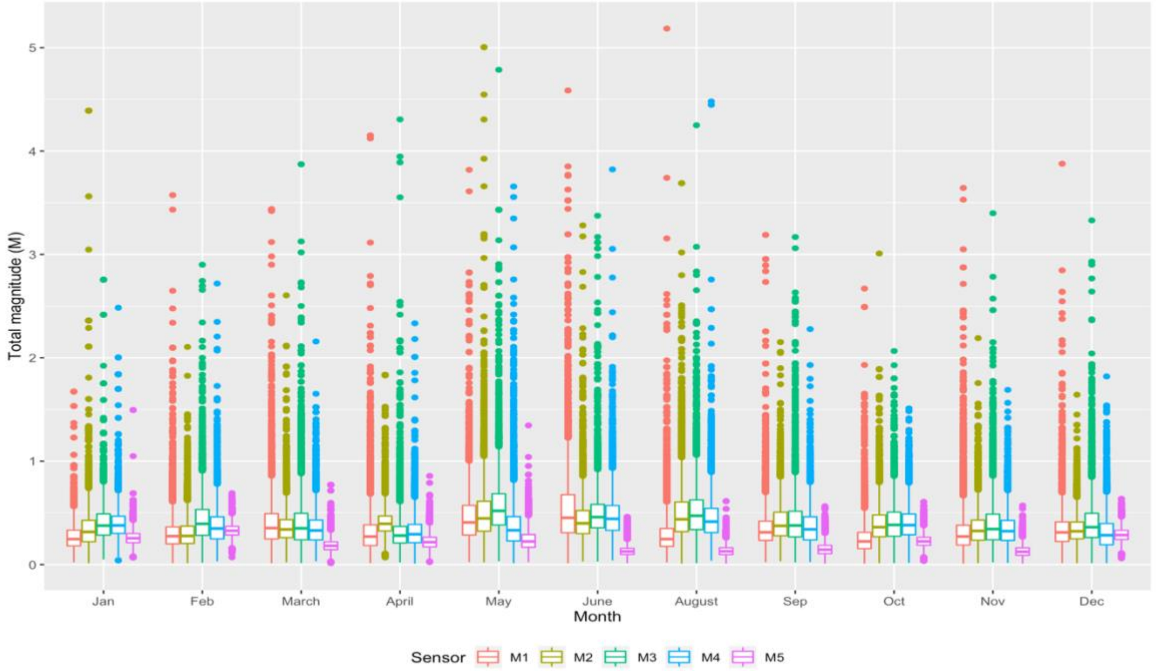


Figure 3.4. Kernel Density plot for section 1

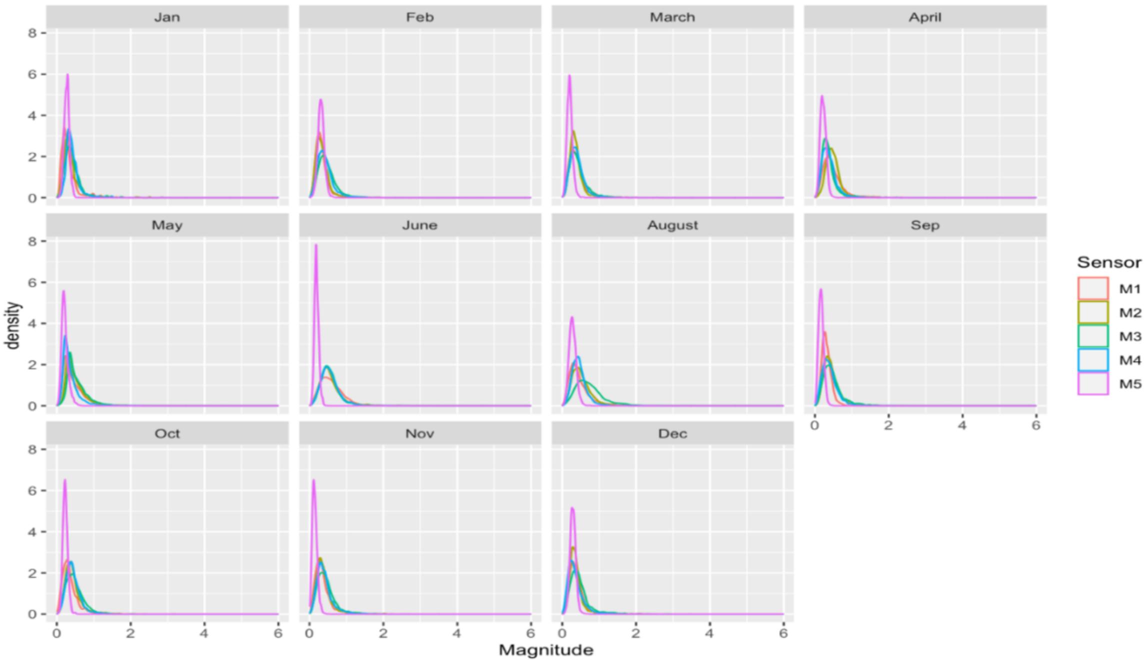
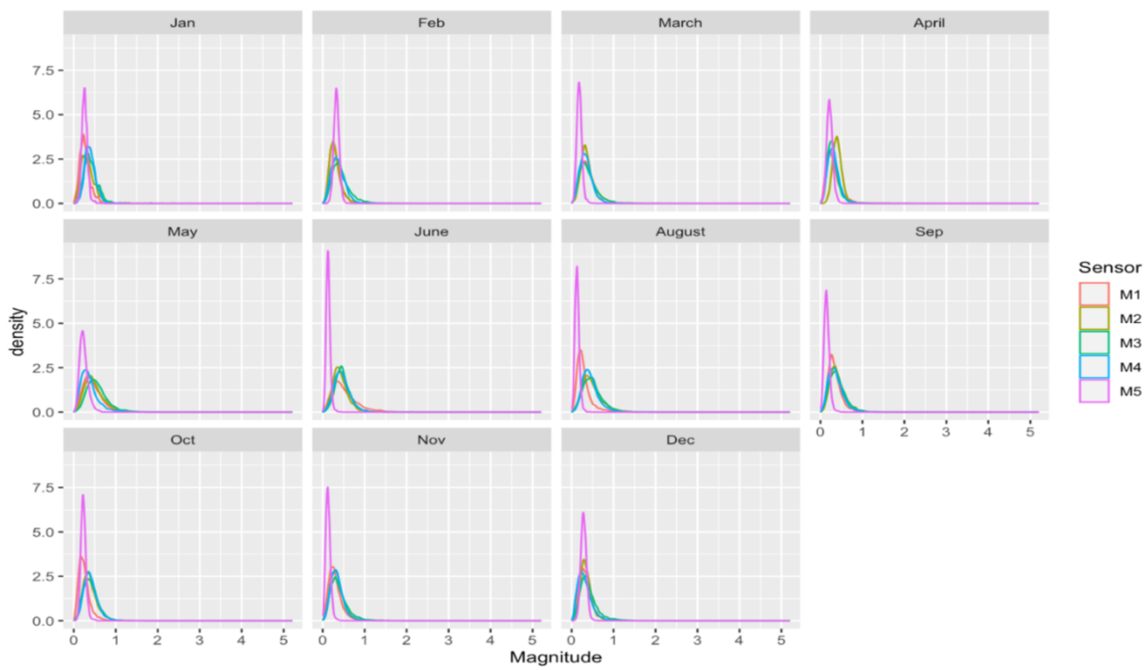


Figure 3.5. Kernel Density plot for section 2



ANOVA Tests

In order to test the equality of treatments (sensors), the Analysis of Variance (ANOVA) tests are constructed by building cell means model as shown follow (41):

$$\boldsymbol{y_{ij} = \mu_i + e_{ij}}, \boldsymbol{i = 1, 2, \dots, t} \text{ and } \boldsymbol{j = 1, 2, \dots, r} \quad (3.2)$$

where y_{ij} is the response of j^{th} months from the i^{th} sensors, μ_i is the mean for all acceleration from the i^{th} sensor, e_{ij} is random error, which is independent identically distributed to a normal distribution with zero mean and a constant variance.

The hypothesis test was performed as expressed below:

$$\boldsymbol{H_0}: \boldsymbol{\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5} \text{ vs. } \mathbf{H_a}: \textbf{not all the } \boldsymbol{\mu_i} \textbf{ are equal} \quad (3.3)$$

Test statistics: $\mathbf{F} = \frac{\mathbf{MST}}{\mathbf{MSE}}$ (3.4)

where MST is the mean square of sensors and MSE is the mean square of error.

If the p-value is smaller than the significance level 0.05 (or F statistic is larger), we reject the null hypothesis and conclude that the means of sensors differ, which indicates that all sensors should be analyzed to determine pavement deterioration. The Tukey test will be applied for pairwise comparison to find which group causes the significantly difference. On the contrary, if the p-value is greater than the significance level 0.05 (or F statistics is small), we fail to reject null hypothesis and there is no sufficient evidence to conclude that the significantly difference among all five sensors. Thus, it would suggest that any one of five sensors can be exclusively in pavement sensing and the result would have statistically been used for pavement condition assessments.

However, it is essential to check assumptions of normality, linearity, independence, and homoscedasticity to validate this model for the accuracy of statistical inference and precise. There are multiple functions can be used in R, such as aov, lm, or glm for ANOVA analysis. The paper used lm function to conduct ANOVA test so that the Global Validation of Linear Models Assumptions (gvlma) function can be applied to evaluate the assumptions. The summary of gvlma will indicate which assumption is violated.

Distribution Fitting and Percentile Analysis

The concept of defining significant points and classifying pavement conditions is based on probability distribution, which is a function to describe the likelihoods of a variable occurrence. There are many distributions such as normal, beta, chi-square distribution, etc. in statistics. Given a distribution, the probability of an event occurrence can be estimated and found using a percentile number. Similarly, if all magnitude values of vibration data can be fitted in a probability distribution, then a specified percentile of all magnitudes can be estimated and therefore a critical value would be defined as a threshold to further classify a pavement condition.

The distribution fitting approach is to fit an appropriate probability distribution from a series of the data and predict the probability of an event occurrence in an interval. The fitting process involves the statistical methods to estimate the distribution parameters based on the sample data. The paper used the package, fitdistrplus, in R to fit several parametric distributions and find the best fitted model by comparing Akaike Information Criterion (AIC) scores. Moreover, the probability-probability (P-P) Plot, quantile-quantile (Q-Q) Plot, and Cumulative Distribution

Function (CDF) plot are created from the package to help improve the fitting process by modifying the distribution parameters.

Since raw data shows the positively skewed distribution, Gamma distribution and Lognormal distribution are considered as the best fitted models for analysis. After computing 99th percent from the fitted model, the corresponding magnitude value can be defined as threshold in the pavement condition classification. On the other hand, the remaining 1 percent of the magnitude values would indicate that pavement deterioration is determined, which is also called significant points in the paper.

Time Series Analysis

Time-Series Analysis is a method for analyzing time-series data to obtain significant statistics under considerations of stochastic, stationary, autocorrelation, and autoregressive. It is a common and important method to predict the values using a forecasting model referring to the original observed values. In the paper, the number of significant points is represented as a time series and the assumption for the model is non-stationary. Then the model capable of this series data is called Auto Regressive Integrated Moving Average (ARIMA) model. The general ARIMA process has the following form (42):

$$\boldsymbol{W_t} = \boldsymbol{\alpha_1 W_{t-1}} + \cdots + \boldsymbol{\alpha_p W_{t-p}} + \boldsymbol{Z_t} + \cdots + \boldsymbol{\beta_q Z_{t-q}} \qquad (3.5)$$

where $\{Z_t\}$ is a purely random process with mean zero and variance constant.

Results and Discussion

This section presents all computing processes and results from the previous section. The results of sensor selection, the determination of threshold values for road condition classification, and predictable model for the number of significant points are shown in Tables 3.1-3.4, Figure 3.7 through Figure 3.11.

Equality of Sensors

Before applying ANOVA test, the data was transformed into a logarithmic scale as referred in data exploratory analysis. Construct cell means models (Equation 3.2) and the results of ANOVA tests (Table 3.1 and Table 3.2) show that p-values are less than the significance level 0.05 in both sections, which indicate that the means of five sensors differ in log scale as expressed in Equation 3.3. The result states that all sensors should be included for any further analysis in order for the results to be statistically valid. Table 3.3 also shows the results of checking assumptions where the large p-values indicate that the assumptions are not violated. Additionally, the Tukey's Test (Table 3.4). Even though the differences among the other four pairs (M1 to M4) are not significant at the significance level 0.05, it concludes all sensors are going to be used for further analysis.

Table 3.1. Results of ANOVA tests Section 1

Analysis of Variance Table: Section 1 - 51st Ave. through 27th Ave.					
	df	**Sum Square**	**Mean Square**	**F value**	**P-value**
Sensor	5	68.838	13.7676	352.32	$< 2.2 \times 10^{-16}$
Residuals	50	1.954	0.0391		

Table 3.2. Results of ANOVA tests Section 2

Analysis of Variance Table: Section 2 – Baseline Rd. through Chandler Blvd.					
	df	**Sum Square**	**Mean Square**	**F value**	**P-value**
Sensor	5	83.417	16.683	335.64	$< 2.2 \times 10^{-16}$
Residuals	50	2.485	0.050		

Table 3.3. Results of Global Validation Test

	Road Section 1		**Road Section 2**	
	p-value	**Decision**	**p-value**	**Decision**
Global Stat	0.827	Assumptions acceptable	0.175	Assumptions acceptable
Skewness	0.560	Assumptions acceptable	0.768	Assumptions acceptable
Kurtosis	0.842	Assumptions acceptable	0.987	Assumptions acceptable
Link Function	1.000	Assumptions acceptable	1.000	Assumptions acceptable
Heteroscedasticity	0.291	Assumptions acceptable	0.288	Assumptions acceptable

Note: the Global Sta tests for the null hypothesis, the Skewness and Kurtosis check normality, and Heteroscedasticity for the constant variance of the residuals.

Table 3.4. Tukey's Test for Multiple Comparisons

Road Section 1: 51st Ave. to 27th Ave.				
Mixing comparison	**Difference between means**	**Simultaneous 95% Confidence Limits**		**p-value**
M2-M1	0.110	-0.128	0.349	0.689
M3-M1	0.216	-0.023	0.454	0.094
M4-M1	0.072	-0.167	0.310	0.913
M5-M1	-0.490	-0.729	-0.252	0.000
M3-M2	0.106	-0.133	0.344	0.720
M4-M2	-0.038	-0.277	0.200	0.991
M5-M2	-0.600	-0.839	-0.362	0.000
M4-M3	-0.144	-0.382	0.095	0.439
M5-M3	-0.706	-0.944	-0.467	0.000
M5-M4	-0.562	-0.800	-0.323	0.000
Road Section 2: Baseline Rd. to Chandler Blvd.				
Mixing comparison	**Difference between means**	**Simultaneous 95% Confidence Limits**		**p-value**
M2-M1	0.181	-0.088	0.450	0.033
M3-M1	0.250	-0.019	0.519	0.080
M4-M1	0.135	-0.134	0.404	0.615
M5-M1	-0.454	-0.723	-0.185	0.000
M3-M2	0.069	-0.200	0.338	0.948
M4-M2	-0.045	-0.314	0.224	0.989
M5-M2	-0.635	-0.904	-0.366	0.000
M4-M3	-0.115	-0.384	0.154	0.748
M5-M3	-0.704	-0.973	-0.435	0.000
M5-M4	-0.590	-0.859	-0.321	0.000

Classifying Pavement Condition

The results of estimated parameters from Lognormal distribution and Gamma distribution using magnitude values from M3 collected in March 2017 as a computing example are shown in Table 3.5. Comparing AIC scores, the better fitted model is the fitted lognormal distribution since it has lower AIC score than the fitted gamma distribution. The plots of CDFs, Q-Q and P-P also indicate that the Lognormal distribution is better than Gamma distribution to fit the data from M3 as displayed in Figure 3.6. Once the fitted distribution is determined (e.g. Lognormal distribution), the estimated parameters of mean and standard deviations can be computed. Subsequently, a new distribution model with specified parameters was built. It is found that the fitted distribution matched the histogram of raw data as shown in Figure 3.6 in Histogram and theoretical densities. After computing the 99th percentile from the fitted distribution model, the corresponding magnitude (critical) value is determined as the threshold for road condition classification. Figure 3.7 shows all threshold values that are used to determine the pavement deterioration in two road sections. It is noticed that the determination of threshold values varies based on statistical analysis. In addition, the threshold obtained from M5 exhibit lower values than the other four sensors. This is due to the placement of M5 being inside the vehicle as compared with four sensors being placed on top of control arm.

Table 3.5. Parameter estimation results from fitted distribution models

	Fitting of the distribution "Lognormal" by maximum likelihood		
Parameters:	Log likelihood: 74.09	AIC: -144.193	BIC: -131.523
	Fitting of the distribution "Gamma" by maximum likelihood		
Parameters:	Log likelihood: 17.792	AIC: -31.584	BIC: -18.914

Figure 3.6. Plot of fitted distribution model for M3 in March

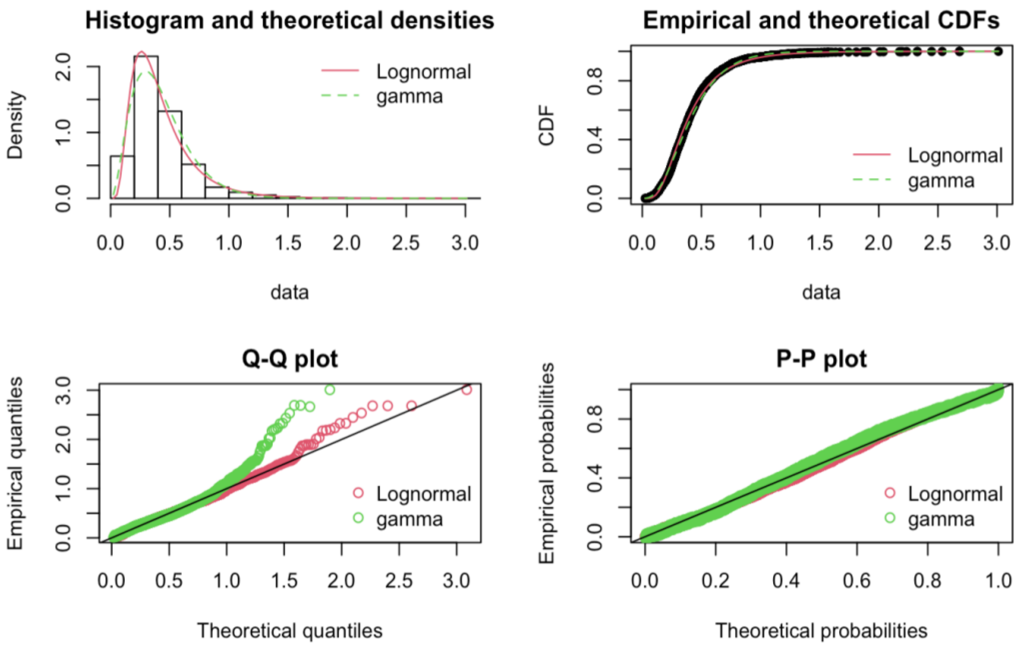
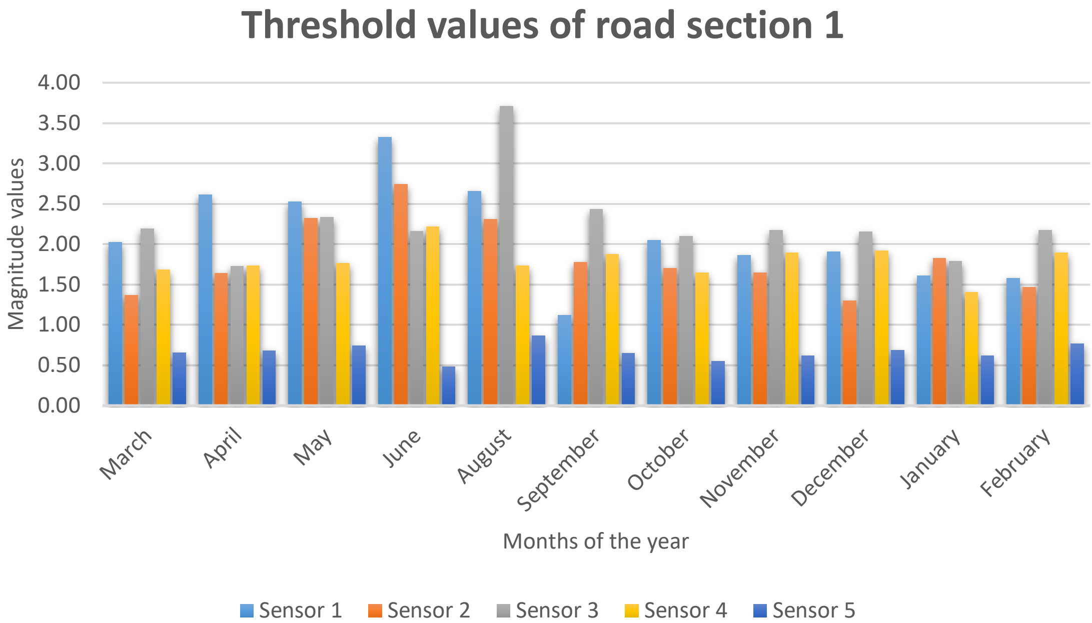
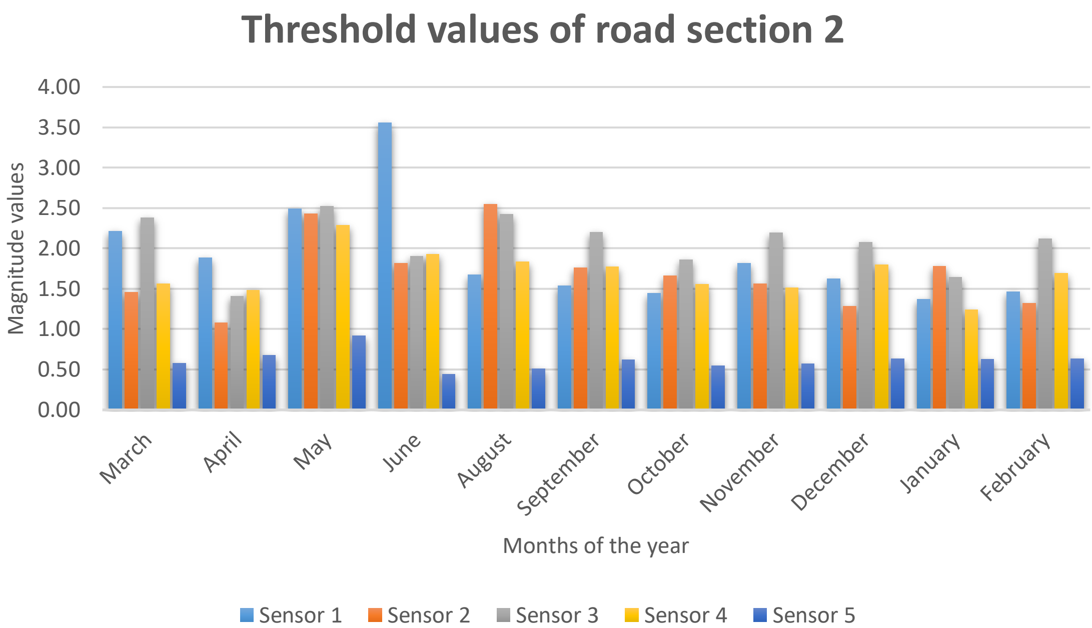


Figure 3.7. Threshold values of road condition classification





In addition to the study of sensor placements and the determination of threshold values in pavement condition assessments, the project also intends to evaluate how extreme heat events could influence the number of significant (pavement deterioration) based on a yearlong data collected in Phoenix. Using the threshold values obtained from Figure 3.7, the total number of significant points from M1 to M5 against 11 months (except for July 2017) is shown Figure 3.8. Obviously, the pavement sensing patterns and signals vary in each month as well as to the significant points. The ANOVA test were conducted to evaluate if the pavement temperature has a significant impact on the number of significant points. The results (Table 3.6) show that there is an association between two variables at the significance level 0.05 in both sections which indicates the pavement temperature does play an important role in controlling pavement conditions. The selected significant points in the two testing sections from April, August, November of 2017 and January of 2018 were imported in GIS software and are graphically illustrated in Figure 3.9 and Figure 3.10. Given the results from Figures 3.9-3.11 and Table 3.6, it is evident that the pavement temperature significantly influences the number of significant points (pavement distress) provided the fact that the number of significant points decrease during cold weather condition while the number of significant points increase as the pavement temperature is getting warmer.

To further validate the accuracy of statistical and GIS mapping findings, a number of images as shown in Figure 3.11 were taken from recorded videos to verify the results. The pavement distress points as displayed in Figure 3.11 accurately correspond to the locations as circled in blue in Figure 3.9. Thus, it is confident to say that the selected significant points generated by a series of statistical analyses could appropriately represent the locations of pavement distress on highways.

Figure 3.8. Pavement temperature and selected significant points

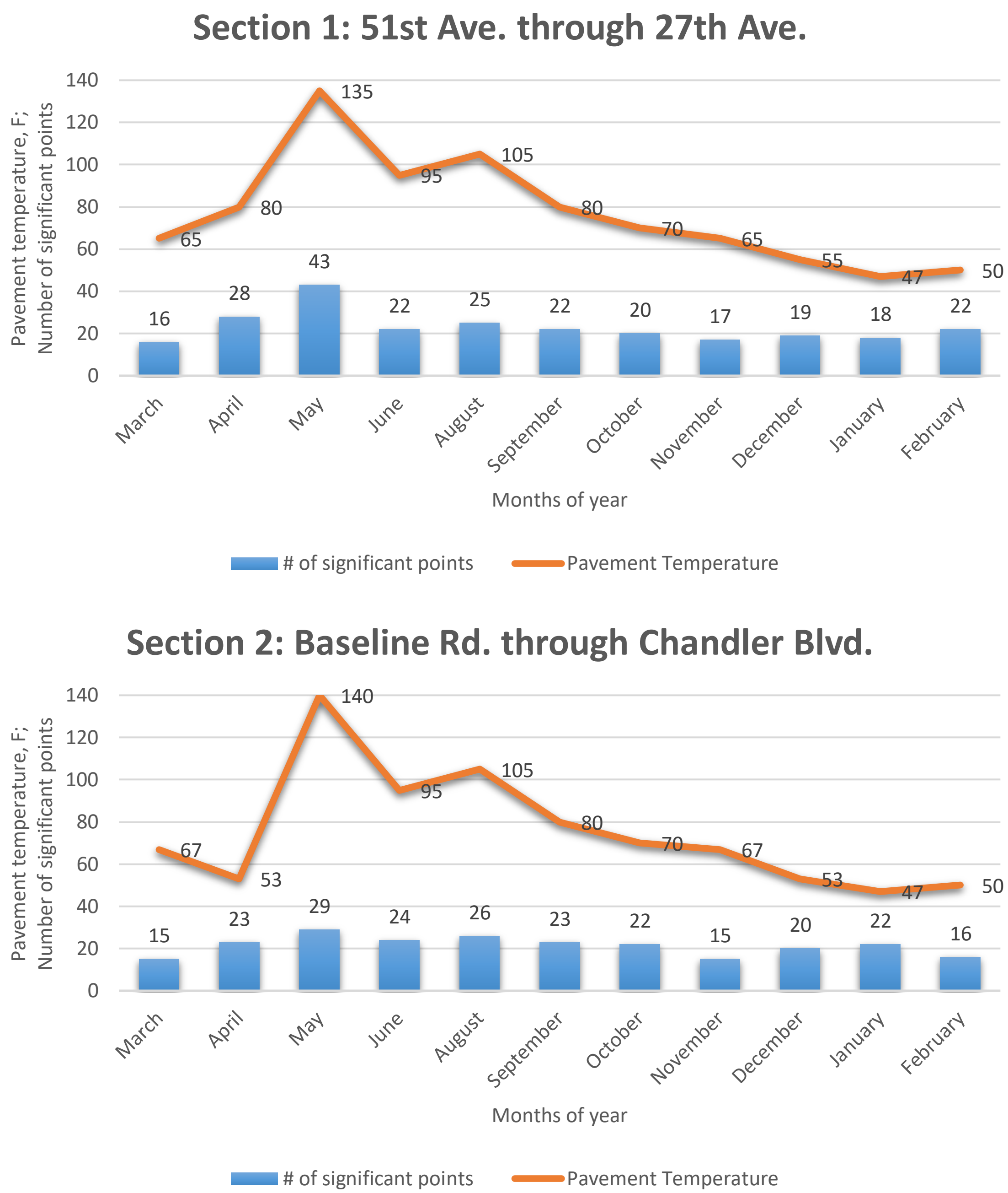


Table 3.6. ANOVA Test Results of Pavement Temperature

Analysis of Variance Table: Section 1 - 51st Ave. through 27th Ave.					
	df	**Sum Square**	**Mean Square**	**F value**	**P-value**
Sensor	5	68.838	13.7676	345.05	$< 2.2 \times 10^{-16}$
Temperature	1	0.218	0.2182	5.575	0.0165
Residuals	49	1.915	0.0391		

Analysis of Variance Table: Section 2 – Baseline Rd. through Chandler Blvd.					
	df	**Sum Square**	**Mean Square**	**F value**	**P-value**
Sensor	5	83.417	16.683	335.64	$< 2.2 \times 10^{-16}$
Temperature	1	0.112	0.112	2.286	0.0331
Residuals	49	2.436	0.049		

Figure 3.9. Pavement Condition of Road Section 1

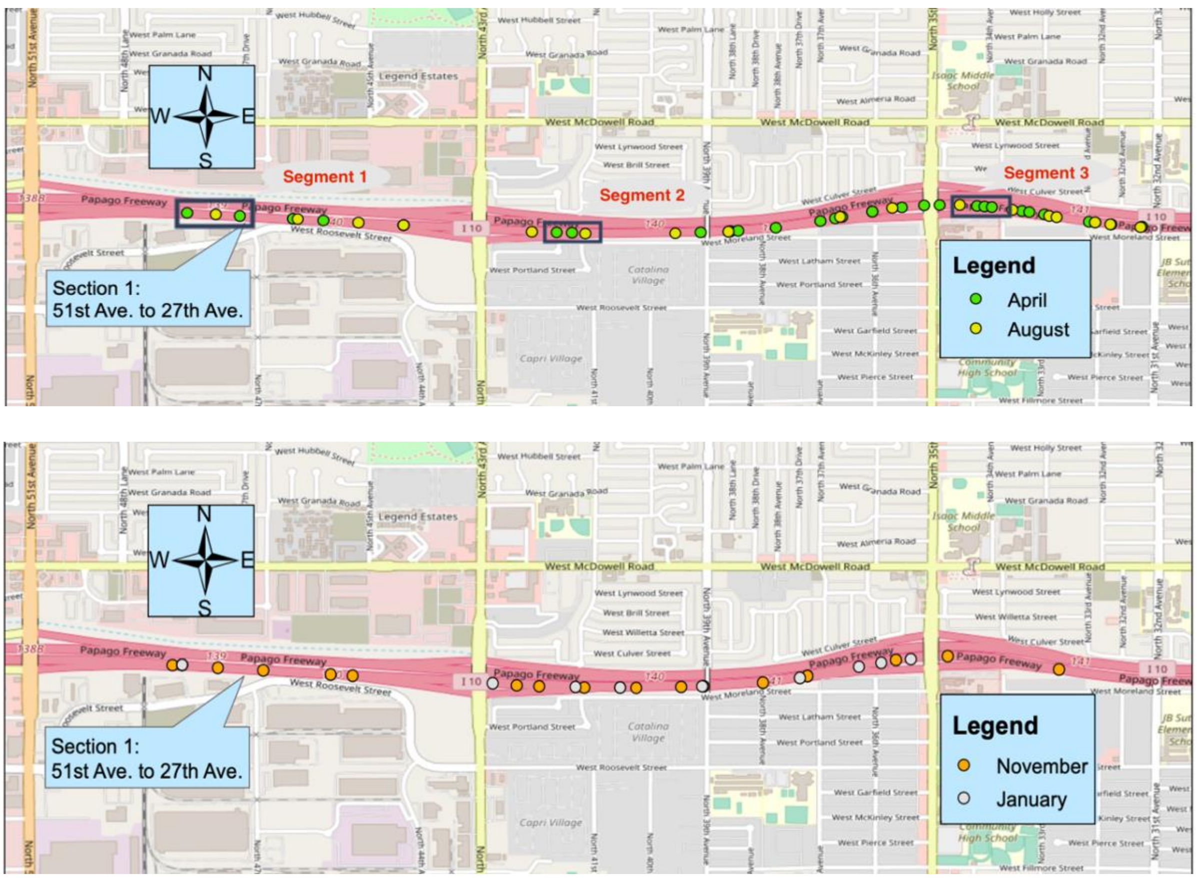
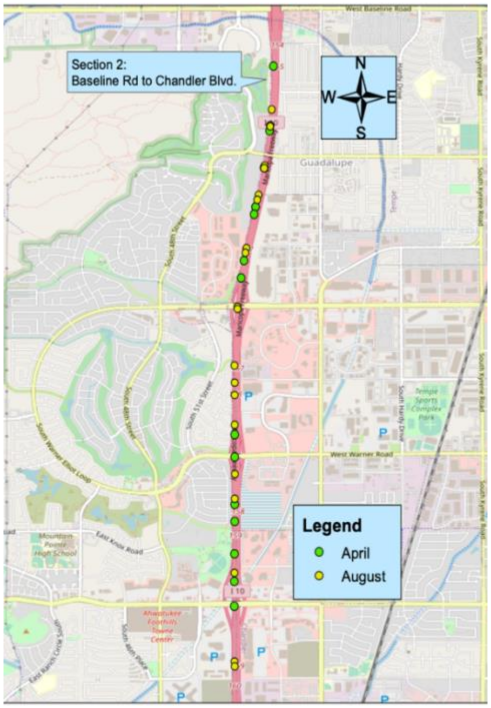


Figure 3.10. Pavement Condition of Road Section 2



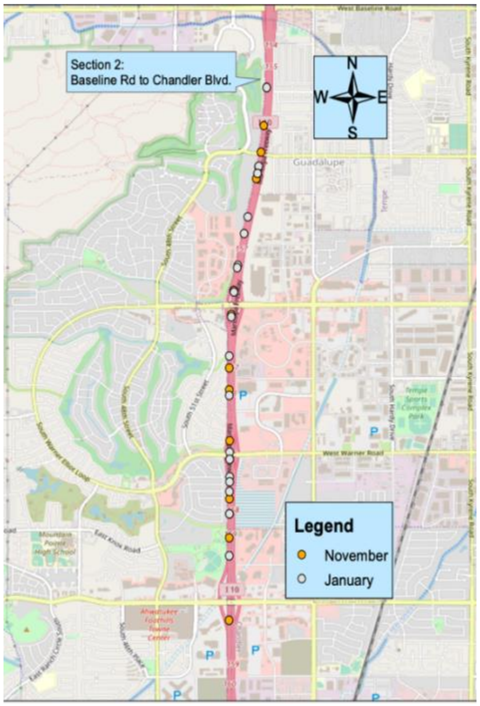


Figure 3.11. The Significant Points Correspond to the Road Condition in Road Section 1.

(a) 1st segment

(b) 2nd segment

(c) 3rd segment

The number of significant points can also be predicated by using Time-Series Analysis and the forecast plots are shown in Figure 3.12. Since both two sections have small sample size and does not roughly uniformly distribute, the ARIMA models were chosen based on lower AIC and BIC values. The fitted models were developed to fit the data as expressed below in Equation 3.5 and Equation 3.6 for two road sections:

section 1: $\mathbf{x_t - x_{t-1} = 0.6 + \{z_t\}}$ (3.6)

section 2: $\boldsymbol{x_t - x_{t-1} = 0.1 + \{z_t\}}$ (3.7)

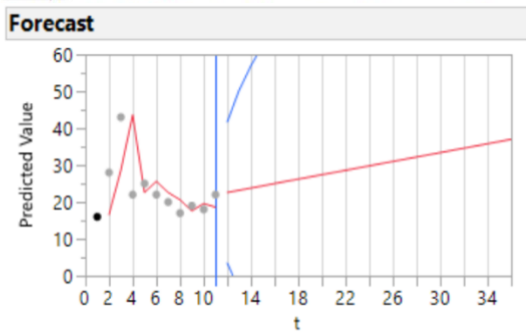
where t is the variable of month.

The predicted number of the significant points will increase from the forecast plot in the following two years, which indicate that the pavements will be deteriorated if the maintenance and rehabilitation will not be scheduled.

Figure 3.12. The Forecast Plots of Number of Significant Points

Parameter Estimates

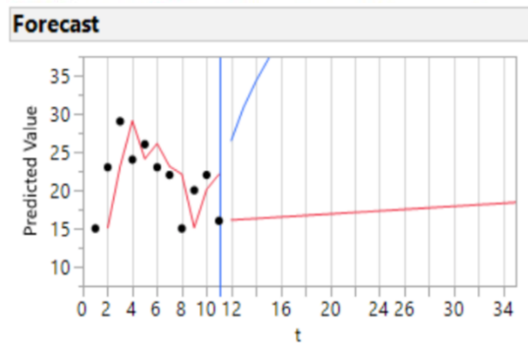
Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
Intercept	0	0.60000000	2.927678	0.20	0.8422	0.6	0.6



(a) Road section 1

Parameter Estimates

Term	Lag	Estimate	Std Error	t Ratio	Prob>|t|	Constant Estimate	Mu
Intercept	0	0.10000000	1.586627	0.06	0.9511	0.1	0.1



(b) Road section 2

Note: x-axis shows the variable of month (t) and y-axis shows the predicted number of significant points.

Chapter 4 Using Multiple Sensors to Detect Pavement Deterioration Through Frequentist and Bayesian Methods

Introduction

The road roughness is critical in pavement management and maintenance for road asset decision-making. How to measure and make prediction of pavement performance in a cost-effective way are widely discussed in many research projects. This chapter presents a new idea for detecting pavement deterioration through multiple vehicle-based sensors and accelerometer app from smartphone.

Generally, the pavement condition assessment uses the international roughness index (IRI) to monitor roughness condition of a target road segment and it becomes a key measurement in pavement management system (43). IRI is obtained from the longitudinal profile with accumulated output that divided by the distance based on the rectified slope with either unit of meter per kilometer or inch per mile (44). The higher IRI values shows the increase roughness of the pavement and the lower IRI values indicates the smooth pavement condition (45). The IRI is a standard index that is applied widely to quantify the road condition and becomes a valuable information in transportation. However, the higher cost and constraints are challenge to road agencies. To reduce the cost and testing time, more researchers use acceleration vibration to determine the pavement deterioration through multiple methods including numerical analysis, machine learning and mathematical models (46-48). Many studies show that the significant relationship between acceleration vibration and IRI values or road surface roughness (49-51), which is an evidence to support that the acceleration responses are useful in detecting road conditions.

Chapter 4 is to continue the evaluation of the effect of multiple sensors on pavement condition assessment. This chapter is focused on two types of data based on the vibration response to analyze the pavement deterioration, which are vertical acceleration and total magnitudes. The total magnitude is a combination expression of the vibration values along axis of x, y and z. Based on the previous research and the current analysis, the threshold values can be determined to classify the pavement condition based on the fitted probability models. Moreover, the Bayesian and frequentist methods of regression analysis are applied to predict the probability of future pavement deterioration, along with Markov Chain Monte Carlo (MCMC).

Pavement Detection Test

The pavement detection test was accomplished in one year that tests two road segments on I-10 in Phoenix, Arizona. During the tests, a 2016 Honda Accord vehicle equipped with four sensors were placed on the top of the control arm and one sensor was placed inside of the vehicle (6). Also, a smartphone was attached on the dashboard of the vehicle to collect vibration responses. The data includes values of acceleration along axis of X, Y, and Z, speed, longitude, latitude, altitude, and time. The details about data collection process are explained in reference (6) and Chapter 3.

Data Analysis

Data cleaning is necessary before performing methodology through ArcGIS software. Using the software, the sensing patterns and signals can be displayed and matched on I-10 corridors in the arterial map. The data collected in July was deficient due to unexpected technical issue in the collection. Since four sensors were placed in different locations of the vehicle, the time difference occurred in data collection. In the pavement detection test, sensor 3 and sensor 4 placed on the rear wheels should be matched the vibration data from front wheels (sensor 1 and sensor 2). Referring to the vehicle specification and features, the wheelbase of Honda Accord is 108.3 inches and the time gap can be estimated based on the driving speed. Sensor 5 was placed inside of the vehicle, which has similar vertical displacement as the sensors placed on front wheel. Thus, there is no need in data modification for sensor 5.

The project analyzes two types of data to identify the pavement crack. One data series is just vertical acceleration and the pavement distress corresponds to the acceleration peaks. The other data series is defined as total magnitude (M), see Equation 3.1. In this case, the larger the magnitude values, the worse pavement distress. In the paper, M1 and M2 represent the magnitude values from front-wheel, and M3 and M4 represent magnitudes for the rear-wheel.

When the time gap is reduced from the data, the vertical acceleration from all sensors indicates the same location in the pavement detection tests. Figure 4.1 shows an example of vertical acceleration responses from road segment 1 (51st Ave. to 27th Ave.) in March versus the incremental time within 9 seconds. The acceleration peak represents crack, pothole or bump and all five sensors play roles in the detection test. Figure 4.2 shows the magnitude values versus incremental time from the same sample data. In this case, the higher point (M3 = 2.0) from sensor 3 indicates pavement deterioration when the incremental time around 8.5 seconds. Additionally, sensor 5 has the lowest total magnitude values in the example, then the study will examine the relation between sensor 5 and other sensors for further analysis.

Figure 4.1 Vertical acceleration vs. Incremental Time

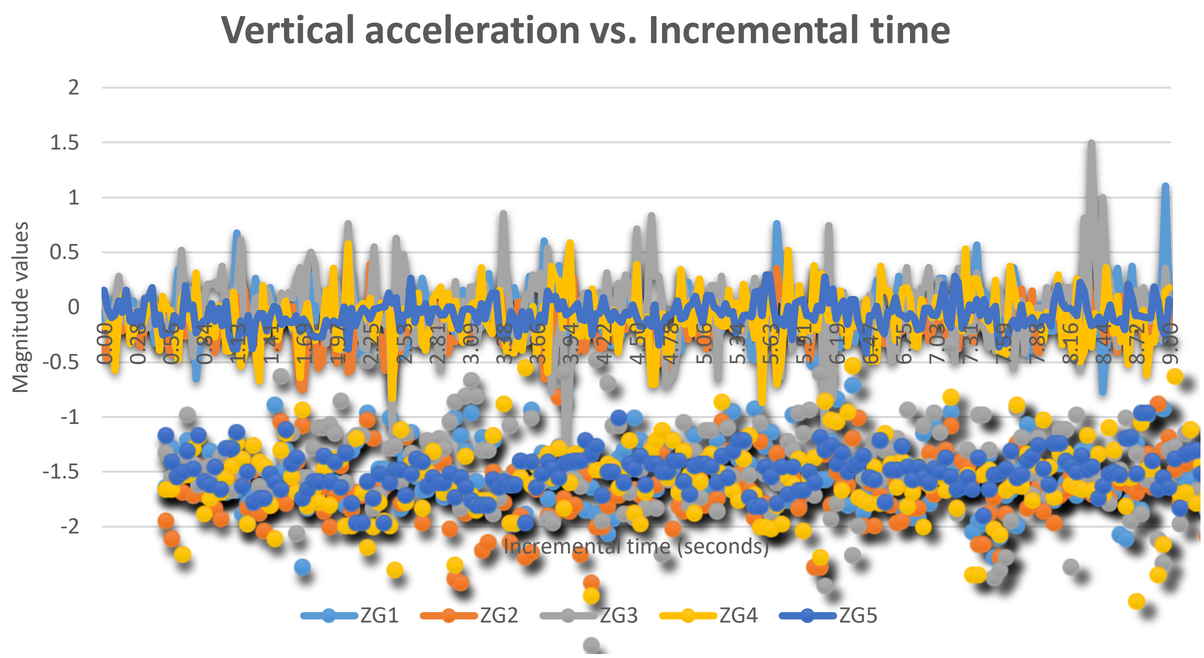
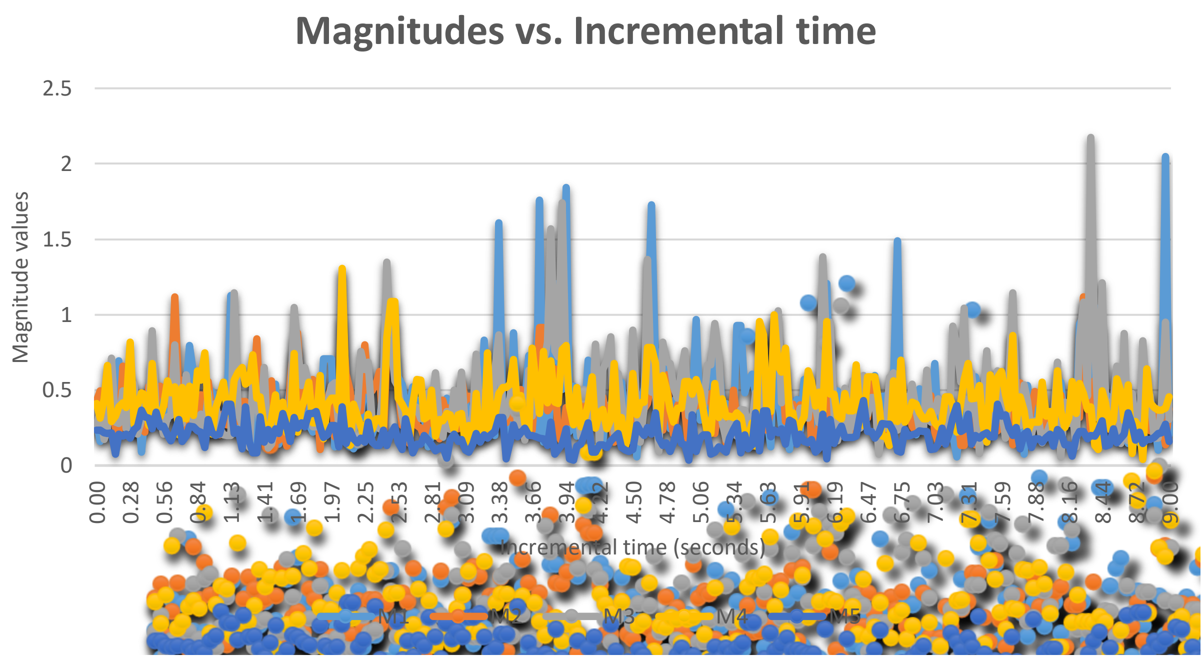


Figure 4.2. Magnitudes vs. Incremental Time



Factorial Experiment and Analysis of Covariance (ANCOVA)

The factorial treatment design is used to investigate the relationships among several types of treatments and the comparisons can be affected by vary conditions (53). From the pavement detection test, since the simultaneous vibration data collected on all five sensors with different speed, then the factorial design would be an appropriate method to be performed. The advantages of the design include simultaneous investigation of several factors, measure interaction between factors, and provide precision analysis on individual factor when there are no interactions. However, when the number of factors increases, the interpretation of analysis becomes more difficult and the precision is reduced.

Analysis of covariance (ANCOVA) is an extension of ANOVA, which tests the mixture of quantitative and qualitative (e.g. categorical variables) predictors. ANCOVA is usually used for comparing variables in two or more groups as well as the interaction of each variable (54). The main effects and interactions both are available to access from the results. In the project, to determine which sensor has a significant effect on the acceleration response, and whether the speed is taken into account in the test, the factorial design and ANCOVA needs to be applied. If five sensors detect the same vibration response, then either sensor could be used to evaluate pavement condition for either speed values. If the results are dependent on the driving speed, then the sensors and speed should be considered for analyzing vibration responses.

Results

The models include the simple model along with both main effects and the interaction ANCOVA model. The models are implemented in R with lm functions, which are shown below:

Model 1: $\boldsymbol{lm}(\boldsymbol{Vibration} \sim \boldsymbol{Sensor})$ (4.1)

Model 2: $\boldsymbol{lm}(\boldsymbol{Vibration} \sim \boldsymbol{Sensor} + \boldsymbol{Speed})$ (4.2)

Model 3: $\boldsymbol{lm}(\boldsymbol{Vibration} \sim \boldsymbol{Sensor} \times \boldsymbol{Speed})$ (4.3)

From the ANCOVA Table (see Table 4.1), the second row compares the simple model (Eq. 4.1) to the main effect additive model (Eq. 4.2), and the third row compares the main effect additive model to the interaction model (Eq. 4.3). P-values is very small that indicates each comparison is significant and the speed does matter in the test. Furthermore, the $\text{sensors} \times \text{speed}$ interaction effects exist. There is no need to test for main effects since the interaction effects are significant.

Table 4.1. Analysis of Variance Table

Res. Df	Residual of Sum Squares	Degree of Freedom	Sum of Square	F	P-value
249130	14833				
249129	14825	1	7.664	129.5	5.359×10^{-30}
249125	14812	4	4.314	18.22	5.597×10^{-15}

Referring to model 3 (Eq. 4.3), the sensor that has significantly an effect on the speed can be determined. From Table 4.2, all sensors and speed have effect on the vibration response. However, there is no sufficient evidence to conclude that the interaction of sensors placed on the rear-wheel of the vehicle and the speed exists. Moreover, understanding the relation between the vibration and pavement temperature is necessary before classifying the pavement condition. Table 4.3 shows ANOVA of the test whether the vibration is affected by the vertical acceleration values. In this case, the pavement temperature has a significantly effect on vertical acceleration at the significance level 0.05.

Table 4.2. Analysis of Variance Table (Model 3)

	Estimated Coefficients	Standard Error	t-value	P-value
(Intercept)	-9.24×10^{-2}	4.30×10^{-3}	-21.483	$< 2 \times 10^{-16}$
Sensor 2	-1.46×10^{-1}	6.08×10^{-3}	-23.934	$< 2 \times 10^{-16}$
Sensor 3	1.35×10^{-2}	6.08×10^{-3}	2.219	0.026476
Sensor 4	-3.45×10^{-2}	6.08×10^{-3}	-5.677	1.37×10^{-8}
Sensor 5	3.75×10^{-2}	6.08×10^{-3}	6.158	7.37×10^{-10}
Speed	4.54×10^{-4}	8.62×10^{-5}	5.263	1.42×10^{-7}
Sensor 2: Speed	4.51×10^{-4}	1.22×10^{-4}	3.696	0.000219
Sensor 3: Speed	6.71×10^{-5}	1.22×10^{-4}	0.551	0.581771
Sensor 4: Speed	-1.37×10^{-5}	1.22×10^{-4}	-0.113	0.910336
Sensor 5: Speed	-5.79×10^{-4}	1.22×10^{-4}	-4.748	2.06×10^{-6}

Table 4.3. ANOVA of Pavement temperature

	df	Sum Square	Mean Square	F value	P-value
Pavement Temperature	1	25.287	25.287	409.07	$< 2.2 \times 10^{-16}$
Vertical acceleration	250428	15479.4	0.0618		

Chapter 5 Conclusion and Future Recommendation

Through pavement sensing data collection, development of computing algorithms, determination of optimum sensor placements and thresholds, the project has the following conclusions:

1. The combination resampling method has the best performance to prevent an over-fitting occur due to an unbalanced dataset.
2. According to a Stratified K Fold cross-validation mechanism and an evaluation matrix (based on accuracy, MCC, precision, recall, and f1-score), the random forest algorithm is the best fitting machine learning algorithm than other three algorithms (Linear Regression, Support Vector Machine, and Neural Network) for the pavement condition assessment and prediction with its rated accuracy of 98%, a Matthews correlation coefficient of 76.54%, a precision of 95%, and a recall of 77%.
3. The determination of thresholds varies based on statistical analysis. The threshold obtained from M5 exhibit lower values than the other four sensors. This is due to the placement of M5 being inside the vehicle as compared with four sensors being placed on top of control arm. Thus, it is recommended all five sensors should be used simultaneously to be accurately predict pavement conditions.
4. The pavement temperature significantly influences the number of significant points (pavement distress) provided the fact that the number of significant points decrease during cold weather condition while the number of significant points increase as the pavement temperature is getting warmer.
5. The Time-Series analysis is applied to fit a model for predicting the number of significant points for the next two years. The predicted number of the significant points will increase from the forecast plot in the following two years, which indicate that the pavements will be deteriorated if the maintenance and rehabilitation will not be scheduled.
6. Time gap is reduced from different sensors so that all vibration responses indicate the same location in the pavement detection test, which improves precision. The interaction of sensor and the speed exist based on the statistical analysis. The speed should be taken into account in further analysis in pavement deterioration.
7. In general, the cost-effective sensing systems and analytics (CeSSA) presented in the report indicate that the CeSSA is cable of capturing pavement vibration patterns and determining a level of pavement distress

Recommendation for future work

Due to time limited, the authors recommend following works to be continued for future implementation:

- Analyze total magnitudes with same factorial design and analysis of covariance (ANCOVA) to test the interaction between speed and M values.
- Test the significant relation between pavement temperature and M. Even Chapter 3 shows pavement temperature has an effect on M values, but the data did not reduce time gap before performing the test. Thus, this step is important to improve the precision.
- Using previous research and current analysis to determine thresholds for classifying pavement condition, and comparing the results with IRI based on vertical acceleration and total magnitudes.
- Using frequentist and Bayesian methods to estimate the probability of pavement deterioration with specified threshold values. The priors are determined from the current sample data through Markov Chain Monte Carlo (MCMC) and Bayesian method.
- The high density intervals (HDI) will be constructed from the posterior distribution that gives the specified credible values for classifying the pavement condition.

Reference

1. Eriksson, J., Girod, L., Hull, B., Newton, R., Mad-den, S., and Balakrishnan, H. (2008), The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring, in Proc. of the 6th Intl. Conf. on Mobile systems, applications, and services (MobiSys), pp. 29–39
2. Mohan, P., Padmanabhan, V.N., and Ramjee, R. (2008). Nericell: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones. In the proceeding of SenSys'08, November 5–7, Raleigh, North Carolina, USA
3. Yu, B. and Yu, X. (2006). Vibration-based system for pavement condition evaluation. Applications of Advanced Technology in Transportation: pp. 183-189
4. The City of Tucson website. http://government.tucsonaz.gov/street-bonds. Accessed in January 17, 2015.
5. Transview Website. http://www.transview.org/aran/. Accessed in January 17, 2015
6. Ho, C.H., Snyder, M. and Zhang, D. (2020). "Application of Vehicle-Based Sensing Technology in. Monitoring Vibration Response of Pavement Conditions", Journal of Transportation Engineering, Part B: Pavements, doi:10.1061/JPEODX.0000205
7. Wang, Y. (2013). "Ordinal Logistic Regression Model for Predicting AC Overlay Cracking.", Journal of Performance of Constructed Facilities 27.3: 346-53. Web.
8. Thongtanunam, P., McIntosh, S., Hassan, A.E. and Iida, H. (2017). "Review. Participation in Modern Code Review: An Empirical Study of the Android, Qt, and OpenStack Projects". Empirical Software Engineering 22.2: 768-817. Web.
9. Yap, B.W., Rani, K.A., Rahman, H.A.A., Fong, S., Khairudin, Z. and Abdullah, N.N. (2013). An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. Proceedings of the First International Conference on Advanced Data and Information Engineering. Lecture Notes in Electrical Engineering, vol 285. Springer, Singapore
10. More, A. (2016). Survey of Resampling Techniques for Improving Classification Performance in. Unbalanced Datasets.
11. Liu, Y., Bi, J. and Fan, Z. (2017). A method for multi-class sentiment classification based on an. improved one-vs-one (OVO) strategy and the support vector machine (SVM) algorithm. Information Sciences, Volumes 394–395. 38-52. ISSN 0020-0255
12. Galar, M., Fernández, A., Barrenechea, E., Bustince, H. and Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognition. Volume 44. Issue 8. 1761-1776. ISSN 0031-3203
13. Alam, M. S. and Vuong, S.T. (2013). Random Forest Classification for Detecting Android Malware. IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, Beijing. 663-669. doi: 10.1109/GreenCom-iThings-CPSCom.2013.122.
14. Orhan, U., Hekim, M. and Ozer, M. (2011). EEG signals classification using the K-means clustering and a multilayer perceptron neural network model, Expert Systems with Applications. Volume 38. Issue 10. 13475-13481. ISSN 0957-4174
15. Moritz, P., Nishihara, R. and Jordan, M. (2016). A linearly-convergent stochastic L-BFGS algorithm. In Artificial Intelligence and Statistics. 249-258.

16. Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B, Methodological 36.2: 111-33.
17. Cui, G., Man, L.W., Zhang, G. and Li, L. (2008). Model Selection for Direct Marketing: Performance Criteria and Validation Methods. Marketing Intelligence & Planning 26.3: 275-92.
18. Chicco, D., Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1. score. and accuracy in binary classification evaluation. BMC Genomics 21, 6. https://doi.org/10.1186/s12864-019-6413-7
19. Hadjidemetriou, G. M., Christodoulou, S. E. and Vela P. A. (2016). "Automated detection of pavement. patches utilizing support vector machine classification", Mediterranean Electrotechnical Conference (MELECON), Lemesos, 1-5, doi: 10.1109/MELCON.2016.7495460
20. Du, Y., Liu, C., Wu, D., Jiang. S. Measurement of international roughness index by using-axis accelerometers and GPS. Mathematical Problems in Engineering. 2014 Mar.
21. Douangphachanh, V., Oneyama, H. Using smartphones to estimate road pavement condition. International Symposium for Next Generation Infrastructure, 2013, Wollongong, Australia.
22. Chen, C., Seo, H.S., Zhao, Y., Chen, B., Kim, J.W., Choi, Y., Bang, M. Pavement Damage Detection System Using Big Data Analysis of Multiple Sensor. International Conference on Smart Infrastructure and Construction 2019 (ICSIC) Driving data-informed decision-making 2019 (pp. 559-569). ICE Publishing.
23. Wang, G., Burrow, M., Ghataora, G. Study of the Factors Affecting Road Roughness Measurement Using Smartphones. Journal of Infrastructure Systems. 2020 10.1061/(ASCE) IS.1943-555X.0000558.
24. Yan, G., Wang, L., Ye, Z., Wang, W. Effects of Crack Damage on Acceleration Response of Asphalt Pavement via Numerical Analysis. Journal of Materials in Civil Engineering. 2020 Jul 1;32(7):04020163.
25. Arhin, S.A., Williams, L.N., Ribbiso, A., Anderson, M.F. Predicting pavement condition index using international roughness index in a dense urban area. Journal of Civil Engineering Research. 2015;5(1):10-17.
26. Múčka, P. International Roughness Index specifications around the world. Road Materials and Pavement Design. 2017 Jul 4; 18(4):929-65.
27. Li, C., Chen, S., Zhao, Y., Chen, Y. Road Pavement Identification based on Acceleration Signals of Off-road Vehicles Using the Batch Normalized Recurrent Neural Network. In2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA) 2019 Mar 29 (pp. 172-177). IEEE.
28. Yan, G.N., Ye, Z.J., Guo, M., Wang, L.B. Study on integral processing of pavement vibration acceleration. Proc., Advances in Materials and Pavement Prediction. 2018 Jul 16:83-7.
29. Zhao, B., Nagayama, T. IRI estimation by the frequency domain analysis of vehicle dynamic responses. Procedia Engineering. 2017 Jan 1; 188:9-16.
30. Campillo, J.R. *A simplified pavement condition assessment and its integration to a pavement management system* (Doctoral dissertation, Arizona State University).

31. Bhatt, U., Mani, S., Xi, E., Kolter, J.Z. Intelligent pothole detection and road condition assessment. arXiv preprint arXiv:1710.02595. 2017 Oct 6.
32. Bridgelall, R. Connected vehicle approach for pavement roughness evaluation. Journal of Infrastructure Systems. 2014 Mar 1;20(1):04013001.
33. Ye, Z., Lu, Y., Wang, L. Investigating the pavement vibration response for roadway service condition evaluation. Advances in Civil Engineering. 2018 Jul 8;2018.
34. Inkoom, S., Sobanjo, J., Barbu, A., Niu, X. Pavement Crack Rating Using Machine Learning Frameworks: Partitioning, Bootstrap Forest, Boosted Trees, Naïve Bayes, and K-Nearest Neighbors. Journal of Transportation Engineering, Part B: Pavements. 2019 Sep 1;145(3):04019031.
35. Padarthy, M., Sami, M., Heyns, E. Investigation on identifying road anomalies using in-vehicle sensors for cooperative applications and road asset management. Transportation research record. 2020 Jul;2674(7):585-95.
36. Sun, L., Gu, W. Pavement condition assessment using fuzzy logic theory and analytic hierarchy process. Journal of Transportation Engineering. 2011 Sep 1;137(9):648-55.
37. Hong, F., Prozzi, J.A. Estimation of pavement performance deterioration using Bayesian approach. Journal of infrastructure systems. 2006 Jun;12(2):77-86.
38. Hunt, P., Bunker, J. Time series analysis of pavement roughness condition data for use in asset management. In Proceedings of the 3rd Asia-Pacific Conference on Systems Integrity and Maintenance 2002 (pp. 148-153). Queensland University of Technology.
39. Hsiao, WC., Horng, M.F., Tsai, Y.J., Chen, T.Y., Liao, B.Y. A driving behavior detection based on a zigbee network for moving vehicles. In2012 Conference on Technologies and Applications of Artificial Intelligence 2012 Nov 16 (pp. 91-96). IEEE.
40. Olivier, J., Norberg, M.M. Positively skewed data: revisiting the box-cox power transformation. International Journal of Psychological Research. 2010;3(1):68-95.
41. Kuehl, R.O. Design of Experiments: Statistical Principles of Research Design and Analysis. Cengage Learning, United States, 2000.
42. Chatfield, C., Xing, H. The analysis of time series: an introduction with R. CRC press; 2019 Apr 25.
43. Dong, Y., Shao, Y., Li, X., Li, S., Quan, L., Zhang, W., Du, J. Forecasting pavement performance with a feature fusion LSTM-BPNN model. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management 2019 Nov 3 (pp. 1953-1962).
44. Zhou, G., Wang, L., Lu, Y. International roughness index model enhancement for flexible pavement design using LTPP data. 2008.
45. Nguyen, H.L., Pham, B.T., Son, L.H., Thang, N.T., Ly, H.B., Le, T.T., Ho, L.S., Le, T.H., Tien Bui, D. Adaptive network based fuzzy inference system with meta-heuristic optimizations for international roughness index prediction. Applied Sciences. 2019 Jan;9(21):4715.
46. Yan, G., Wang, L., Ye, Z., Wang, W. Effects of Crack Damage on Acceleration Response of Asphalt Pavement via Numerical Analysis. Journal of Materials in Civil Engineering. 2020 Jul 1; 32(7):04020163.

47. Prażnowski, K., Mamala, J., Śmieja, M., Kupina, M. Assessment of the road surface condition with longitudinal acceleration signal of the car body. Sensors. 2020 Jan;20(21):5987.
48. Thiandee, P., Witchayangkoon, B., Sirimontree, S., Lertworawanich, P. An Experiment on Measurement of Pavement Roughness via Android-Based Smartphones. arXiv preprint arXiv:1907.13120. 2019 Jul 28.
49. Vemuri, V., Ren, Y., Gao, L., Lu, P., Song, L. Pavement Condition Index Estimation Using Smartphone Based Accelerometers for City of Houston. In Construction Research Congress 2020: Infrastructure Systems and Sustainability 2020 Nov 9 (pp. 522-531). Reston, VA: American Society of Civil Engineers.
50. Zhang, J., Wang, L., Jing, P., Wu, Y., Li, H. IRI Threshold Values Based on Riding Comfort. Journal of Transportation Engineering, Part B: Pavements. 2020 Mar 1;146(1):04020001.
51. Du, R., Qiu, G., Gao, K., Hu, L., Liu, L. Abnormal road surface recognition based on smartphone acceleration sensor. Sensors. 2020 Jan;20(2):451.
52. Zeng, H., Park, H., Fontaine, M.D., Smith, B.L., McGhee, K.K. Identifying deficient pavement sections by means of an improved acceleration-based metric. Transportation Research Record. 2015 Jan;2523(1):133-42.
53. Kuehl, R.O. Design of Experiments: Statistical Principles of Research Design and Analysis. Cengage Learning, United States, 2000.
54. Faraway, J.J. Linear models with R. CRC press; 2014 Jul 1.

Data Management Plan

Products of Research
All field data collected for this research was vibration responses from five sensors mounted on a vehicle that travelled on the I-10 corridors in Phoenix. All vibration data was used to analyze and predict pavement conditions using Matlab, R, Excel, and ArcGIS.

Data Format and Content
The format of all vibration data is in a csv file that was further converted to a .xls format. For computing purposes, all vibration data were analyzed against their accuracy for prediction of pavement conditions using Matlab (in a m. format) and python (in a .ipynb format). For statistical analysis purposes, all vibration were analyzed using R software and the output files are in .rmd format. When pavement conditions were identified, maps were created using ArcGIS software and its format is in a .mxd format.

Data Access and Sharing
All vibration and analysis data are saved and stored at Northern Arizona University's secure drive. General public can make a request to the PI who will create a path directory using Google drive, Dropbox, or Onenote to allow general public access the requested data.

Project data are also stored on Dataverse, and can be accessed here:
https://doi.org/10.7910/DVN/MGGWCN

Reuse and Redistribution
Part of data requested by general public may be reused or redistributed for research purposes with permission of the PI.