PREDICTION OF DEFECT HOTSPOTS FOR HIGHWAY MAINTENANCE MANAGEMENT: A MULTI-ASSET MACHINE LEARNING APPROACH

by

Arash Karimzadeh

A dissertation proposal submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Infrastructure and Environmental Systems

Charlotte

2020

Approved by:

____
Dr. Omidreza Shoghli

____
Dr. Hamed Tabkhivayghan

____
Dr. Glenda Mayo

____
Dr. Srinivas Pulugurtha

____
Dr. Sandra Clinton

ABSTRACT

ARASH KARIMZADEH. Prediction of defect hotspots for highway maintenance management: A multi-asset machine learning approach. (Under the direction of DR. OMIDREZA SHOGHLI)

Given multiple budget and revenue constraints that the transportation sector encounters, predictive analytics enables maintenance agencies to make effective decisions, prioritize maintenance tasks, and provide efficient life-cycle planning. To this end, risk-based predictive models have provided promising results in representing the susceptibility of assets to future defects. Hence, the main objective of this study is to provide an integrated framework for predicting the occurrence probability of multiple defects on different highway asset types. Several gaps in previous models were identified, including limitations in predictive frameworks given the inadequate scope of available inspection data, expert-based selection of contributing factors, and ignoring the interrelationships between neighboring assets. Therefore, this study proposes a risk-based method that combines a risk score generator and a Machine Learning (ML) algorithm to predict the hotspots of multiple defects in a given roadway. To find the best fit, the model is chosen from a pool of ML algorithms containing linear and non-linear methods. To measure the efficiency of the proposed model, its performance is investigated on a selected case study. The proposed framework produced results with significant accuracy within the extent of available data in the case study for calculating risk scores of erosion, obstruction, and cracking on paved ditches given historical weather, traffic, maintenance, and inspection data of five selected neighboring assets (flexible pavements, unpaved ditches, slopes, small pipes and box culverts, and under drain pipes and edge drains). Additionally, the contribution of the considered factors was investigated to further study the importance of individual contributors. The framework offers decision-makers a holistic view of degradation risks of multiple assets, which could enable them to prepare an integrated asset management program. Additionally, a similar framework can be applied to other linear infrastructure systems such as sanitary sewers, water networks, and railroads.

ACKNOWLEDGEMENTS

My doctoral journey could have been impossible without the support, assistance, and guidance of a number of individuals. First and foremost, I would like to express my utmost gratitude to my advisor Dr. Shoghli for his guidance and encouragement during my Ph.D. studies. I am extremely grateful to him for his support in the successful completion of my doctoral degree. I also want to offer my special thanks to my dissertation committee that contributed to my academic growth over the past years. Dr. Tabkhi, Dr. Pulugurtha, Dr. Clinton, and Dr. Mayo provided me with valuable insights into the different aspects of the research process through their guidance and advice. I am really lucky that I had the opportunity to work with and learn from them all.

I would like to thank the Virginia Department of Transportation (VDOT) for sponsoring and supporting this research project, and the Bristol district staff for providing relevant data for the modeling. I would also like to acknowledge the LEIDOS research teammate members for their significant contribution to this study. I graciously appreciate the help, support, and advice of Dr. Adrian Burde offered on the development of this study. I am also appreciative of all information and valuable feedback that Mr. Charles D. Gantt provided on the maintenance data utilized in this research study.

Additionally, I want to thank my parents and sisters for always supporting and encouraging me, especially during the entire journey of my doctoral studies. I would also like to thank my wife Elham. I am not sure how to put into words my appreciation and

gratitude for her support and sacrifice. Finally, to my dear son Arad, I want you to know that you are the greatest inspiration for me, and I hope this work makes you proud.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Overview

Most of the US interstate highways have passed or will exceed their design life in the next 20 years and require restorations and preservations (NASEM, 2019). Hence, Departments of Transportation (DOTs) strive to maintain roadways in a good state of repair by spending a major part of their budget on preservation works. However, tight budgetary constraints and continuous degradation of assets are the main challenges that DOTs encounter. Therefore, they are always looking for optimal solutions that make their investments effective and efficient (AASHTO, 2011; Shoghli et al., 2016). To this end, predictive analytics and data collection are the tools that pave the path toward more efficient decision making, maintenance prioritizing, and life-cycle planning. However, intensifying data collection considering the current practices in DOTs seems to be expensive and far-fetched, given the enormous US roadway system and the extent of highway assets. Therefore, researchers and practitioners are interested in enhancing predictive analytics utilized in Transportation Asset Management (TAM) programs to augment the potential of the current available data.

To this end, numerous models have been developed and applied in forecasting the deterioration of roadway assets. Among different available models, the techniques that provide information on the occurrence probability of defects can be utilized to estimate risks throughout roadways, since the risk is usually defined as the probability of undesired events multiplied by their impacts (Proctor et al., 2012; Renn, 2008). Therefore, such models are capable to provide an important aspect of risk (i.e. probabilities) in a risk-based asset management system. Additionally, the outcome of the models is utilized in preparing risk-maps that present hotspots of defects and their probabilities, which in turn shows the parts of roadways that are prone to a particular set of defects (Madaio et al., 2016). Later on, risk-maps could be used in managing risks, prioritizing inspections, and making maintenance decisions.

1.2 Problem statement

Given the important role of prediction models in TAM, their performance directly impacts the efficiency of maintenance decisions. Therefore, understanding the limitations of the models is a major step toward enhancing their performance and ultimately making smart and efficient decisions for maintenance works.

One of the challenges in developing data-driven prediction models is the inadequacy of the inspection data on all road segments due to the current implemented inspection methods in DOTs. Therefore, several predictive frameworks compatible with the inadequate scope of available inspection data were proposed in the literature. However, their limitations directly impact the performance of the developed models.

Although several factors contribute to the degradation of different asset types in a roadway system, most of the proposed prediction models in previous studies were developed based on a few subjectively selected contributing factors. However, due to the complexity of the degradation process, some of the contributors might be unknown to the experts. Moreover, there is not a complete agreement among practitioners and researchers on the factors considered in previous studies. Therefore, considering only a few factors might result in ignoring the major contribution of other factors to the degradation of assets and negatively impacts the performance of prediction models.

The degradation of assets might be influenced by the deterioration of nearby assets. Furthermore, neighboring assets might deteriorate similarly as they are in the same environment or constructed using the same materials. However, the interrelation was marginally considered in developing prediction models in previous studies and each asset was considered independently. Hence, the performance of the models might be negatively impacted by ignoring interrelations.

In summary, as identified in this study, the gaps and limitations in the current predictive models are: (i) limitations in predictive frameworks compatible with the inadequate scope of available inspection data, (ii) expert-based subjective selection of contributing factors, and (iii) marginal consideration of interrelationships between neighboring assets.

1.3 Objectives of the Study

To mitigate these challenges, this study is proposing a risk-based method that combines a risk score generator and a Machine Learning (ML) algorithm to predict the hotspots of multiple defects in a given roadway. The objectives of this study are fourfold:

1. To augment the limited extent of available inspection data by using density estimation of defects to meet the requirement of risk-based prediction models
2. To consider a wide range of factors with the potential contribution to the degradation of assets, including the interrelations between nearby assets

3. To reduce subjectivity in selecting contributing features by considering a wide range of potential contributors in developing prediction models with a data-driven approach for finding factors with significant contributions, instead of subjectively selecting them beforehand.

4. To predict risk scores of roadway segments susceptible to defects and spatially visualize the corresponding risk-based hotspots.

It is noteworthy mentioning that the scope of the present study is limited to roadway asset types, however, the proposed framework can be applied to any other linear infrastructures, such as railroads, sanitary sewers, and water networks.

To ensure the quality of the results, the proposed framework in this study is equipped with a comparative analysis among multiple ML algorithms that helps to choose the most accurate one for producing outcomes. The outcome of the framework is a set of risk maps that present the predicted hotspots of each defect for the considered asset types in a given road; wherein, hotspots refer to the parts of roadways with higher densities of defects. Moreover, another strength of the proposed framework is its multi-asset risk-based predictor essence, meaning that a single model could be applied to all road asset types.

1.4 Significance of the Study

By considering a wide range of contributing factors to the degradation of roadway assets, this study provides the capability of finding major factors in the occurrence of each defect. Indeed, the framework decides the most significant contributors among a wide range of potential candidates, instead of subjectively selecting them beforehand. In this way, the framework respects the difference among important contributors to the degradation of different highway assets. Furthermore, the deterioration of each asset is investigated considering its possible interrelation with other nearby assets. Hence, this study helps to improve the robustness of prediction models by considering potential factors that identify future deteriorations.

In addition, the proposed methodology lacks the limitation of the previous predictive frameworks and augments the scope of the available inspection data. Therefore, this study enables agencies to come up with prediction models for all road segments based on the annual inspection data on a selected subset of segments.

By forecasting the hotspots of different defects, the results of this study will enable agencies to include the probable location of defected assets in their life-cycle planning and enrich their budget management with the benefit of forecasting possible maintenance needs. Besides, the multi-asset attribute of the framework provides a holistic view of degradation risks of multiple assets and helps DOTs to prepare an integrated asset management program.

1.5 Organization

In the following chapters, first, a literature review is provided in Chapter 2 to ascertain the nature of methodologies utilized in previous studies, and explore the limitations and gaps in the body of knowledge. Then, in Chapter 3, the proposed methodology and the framework of the study are elaborated in greater detail. Next, the efficiency of the framework is measured by applying that to a case study chosen from the state of Virginia, and a comparative study on the selected algorithms is performed. To incorporate the effect of neighboring asset interrelations into the case study, a combination of assets was considered including one capital asset type (flexible pavement) and five adjacent roadside assets– paved ditches, unpaved ditches, slopes, small pipes and box culverts, and under drain pipes and edge drains. However, the proposed framework is similarly applicable to any other highway asset type. Later on, prediction models were developed to forecast the probabilities of observing three defects (erosion, obstruction, and cracking) on paved ditches. All accomplished results are presented and discussed in Chapter 4. Finally, the key findings are discussed and the conclusions, limitations of this study, and recommendations for future works are provided in Chapter 5.

CHAPTER 2: LITERATURE REVIEW

In this chapter, a literature review on the background, methodologies, and applications of prediction models is performed. In the first section, the approaches utilized in developing prediction models for roadway assets are reviewed. Then, based on previous studies, the application of prediction models in a risk-based management system is elaborated and the corresponding gaps in the transportation sector are highlighted. Next, the gaps and limitations in the previous predictive methods are elaborated.

2.1 Prediction Models

In an asset management program, the initial requirement is the knowledge about assets, their condition, and the level of service they provide. This knowledge helps to design, plan, and determine the appropriate maintenance tasks for different assets in a roadway system and optimize resources (Shoghli et al., 2017). Therefore, in predictive asset management programs forecasting the level of service and condition of assets is a major task that enables agencies to establish future maintenance plans (Radopoulou et al., 2016). To this end, prediction models are used to perform such forecasts by providing information about assets' future degradation, life expectancy, future conditions, initiation of defects, or the occurrence probability of defects.

2.1.1 Taxonomy of Approaches

Regardless of what the prediction models provide, they can be categorized into three classes in terms of the approach used in their development: deterministic, probabilistic, and machine learning (Karimzadeh & Shoghli, 2020). Deterministic models are in the form of mathematical functions and are widely used to predict future conditions or the remained life of roadway assets. In probabilistic approaches, a probability distribution such as Weibull is used to forecast the condition of an asset or the probability of a particular condition in the future (Anyala et al., 2014; Sanabria et al., 2017; H. Wang et al., 2017). Finally, in a Machine Learning (ML) approach, Artificial Intelligence (AI) is leveraged to explore relationships between contributing factors to the degradation of assets and their condition or the probability of observing a certain defect. ML prediction models are developed based on the learning process from historically recorded data.

Although deterministic models are the most common prediction tool in TAM programs they cannot effectively capture the stochastic nature of each asset's performance (Toole et al., 2007). Hence, to consider the uncertainty and variability of the deterioration process, an increasing interest in developing probabilistic models emerged from the early years of the twentieth century, especially for pavements (Choummanivong et al., 2013). The initiation and progression of different types of defects follow stochastic variations. For example, Pavement cracking is characterized as a stochastic process due to its random variations. Therefore, probabilistic models have gained more attention in forecasting cracking on pavements (Yang et al., 2004). However, due to assuming a fixed probability distribution in probabilistic models, sometimes the predictions are not consistent with the actual recorded data (Karimzadeh & Shoghli, 2020). Therefore, machine learning approaches were proposed to mitigate the limitations in deterministic and probabilistic approaches. Although machine learning deterioration models were developed several years ago, an increasing interest in using them emerged recently. Studies made efforts to develop deterioration models using machine learning and compared the outcomes with deterministic and probabilistic techniques. To this end, most research studies claimed that using machine learning resulted in more accurate outcomes (Karlaftis et al., 2015; C. Wang et al., 2016). Furthermore, they addressed the capabilities of machine learning models in adjusting to a changing environment in terms of the considered features. Hence, the application of these models proved to be more scalable and extensible compared to deterministic and probabilistic models (Chopra et al., 2018; Sanabria et al., 2017).

2.1.2 Risk-based Predictions

Risk management is vital for many organizations and decision makers to control and mitigate the risk of probable hazards (FTA, 2004). However, only a few state DOTs deployed a risk-based decision framework for their operation and maintenance tasks on an organization-wide scale. In addition, they implement risk-based management strategies for only a selected number of assets (Lin et al., 2015). In order to implement these strategies, data-driven risk maps are proven to be the most useful information that enables asset managers to prioritize inspections and make maintenance decisions (Madaio et al., 2016). In the context of asset management, risk maps are the tools for presenting risks of undesired events that might happen to an infrastructure. Risk includes two dimensions: the occurrence probability of an event and the impact of the occurrence of that event. In this concept, the term event refers to the occurrence of a problem, deficiency, or danger in elements of a system that jeopardizes the functionality of the system (Haimes, 2005).

Preparing risk maps is a common way to visualize the spatial distribution of the risk of undesirable events in different fields. For example, the recorded data of fire events in forest areas are used to show high-risk zones in terms of their susceptibility to wildfire (Kuter et al., 2018; Massada et al., 2009; Millington et al., 2008). As another example, earthquake magnitudes likely to be expected in different zones of a region are shown by earthquake risk maps (Gaull et al., 1990). In addition, in analyzing road traffic accidents, risk maps are widely used to find hotspots of the probable accidents (Erdogan, 2009; Rahman et al., 2018; J. Wang et al., 2011). However, a few studies provided frameworks for developing predictive risk maps of undesirable events in roadway infrastructure. Furthermore, the focus of previous studies on risks in highway systems was mostly on natural hazards, environmental events, flooding, and landslides (C. J. Anderson et al., 2015; Hunt, 1992; Lu, 2020; Sohn, 2006; Wright et al., 2012)

In a highway infrastructure, defects are major problems that negatively impact the condition of an asset and decrease its level of service. Hence, forecasting future hotspots of defects representing the locations in a roadway system with a higher likelihood of observing those defects is an essential task to provide the required information for the risk-based decision making framework in TAM.

2.2 Compatible Predictive Frameworks with Inadequate Inspection Data

The historical condition of road segments has been usually utilized as the foundation of developing predictive deterioration models. However, random inspection of roadways that is the current practice of several DOTs restricts the number of segments with adequate data. This usually results in noncontinuous historical conditions on the majority of road segments during all years of inspection. In other words, a unique condition record corresponding to each year of inspection is often unavailable for each segment. To overcome this limitation, most of the previous studies used the idea of grouping segments with similar deterioration characteristics (family groups) and estimating the average degradation of each group by utilizing a deterioration model (family deterioration model)(Mills et al., 2012; Saha et al., 2017).

Family deterioration models are widely used by many U.S. Departments of Transportation (DOTs) such as Pennsylvania DOT (Wolters et al., 2010), Delaware DOT (Mills et al., 2012), Colorado DOT (Saha et al., 2017), and North Carolina DOT (D. Chen et al., 2016). However, several challenges come with this approach. As an example, the condition of specific segments in a family might be different from the family's average condition. This is mainly attributed to the local variation of contributors to the degradation of assets such as traffic, weather, and maintenance (Pantuso et al., 2019).

Selecting the parameters for grouping similar road segments and creating families was mostly based on experts' opinions. Furthermore, the selected categories for each parameter were considered differently in previous studies. For example, Bannour et al. (2017) used Average Annual Daily Traffic (AADT) for grouping road segments into families. In their study, road segments were grouped based on three categories of AADT:

high (>4500), medium (1500-4500), and low (<1500). Another factor that was utilized in their creation of families was temperature. Then, roadways were grouped into three families located in tropical, subtropical-hot, and subtropical-cool zones in terms of temperature. However, D. Chen et al. (2014) utilized four categories of AADT for their family creation framework: 0-1000, 1000-5000, 5000-15000, and more than 5000. They also considered other factors in their study to build families of segments including pavement types (asphalt and Jointed Concrete Pavement (JCP)) and roadway types (Interstate, U.S., North Carolina, and Secondary Road).

It is worth mentioning that when the number of parameters increase, exploring the impacts of their interaction on the configuration of family groups would be complex. Consequently, the accuracy of the developed deterioration models for the created families depends on the expert's decision in choosing and categorizing the considered parameters. Hence, several studies attempted to minimize the subjectivity in family creator frameworks. For example, Tighe et al. (2001) utilized an unsupervised clustering algorithm (k-means) to create family groups but only 94 road segments were considered. As another example, Luo et al. (2006) proposed a cluster-wise regression method for predicting Pavement Condition Rating (PCR) but only based on one parameter: the age of pavements. Later on, Henning et al. (2014) considered more parameters in their clustering technique and created families based on drainage condition, road type (urban or rural), traffic loading, and climate region. Next, Saliminejad et al. (2016) clustered segments with respect to only their prior conditions and historical maintenance and rehabilitation tasks. In another study, Yacout et al. (2019) minimized the variability between performance indices in each cluster of road segments by utilizing Artificial Intelligence (AI). Then, Titus-Glover (2019) attempted to deploy an unsupervised clustering method to minimize subjectivity and improve exploring current and future performance patterns for each asset. Also, in a research study conducted by Chen et al. (2019), segments were clustered by leveraging a high-dimensional clustering approach. However, they considered the level of maintenance of assets as the only parameter in their clustering framework. Finally, Karimzadeh, Sabeti, and Shoghli (2020) used more features to cluster road segments compared to previous studies. However, they addressed identifying the optimal number of clusters as the main challenge in creating family groups. This challenge was also addressed by several studies in the literature (Jain, 2010; Sugar et al., 2003).

2.3 Interrelations of Roadway Assets

Based on the first law of geography, relations exist between everything, but neighboring elements are more related compare to the farthest ones (Zhu et al., 2018). Therefore, in a highway system that contains different asset types in a close distance, interrelations between nearby assets are expected. One of the reasons is the mutual impacts of the condition of nearby assets. For example, degradation of drain pipes under a pavement might cause the subsidence of the base and subbase layers and consequently the creation of potholes on the pavement. Another reason for the interrelation between neighboring assets in a highway system is the same environmental conditions in their vicinity or using similar materials in their construction. For example, a correlation exists between erosions on nearby unpaved shoulders, slopes, and unpaved ditches due to similar precipitations in their vicinity and their same constructive materials. Therefore, the mutual impacts of neighboring assets and the correlations between their deteriorations result in the interrelation between their conditions.

However, a few research studies have investigated such interrelations (Al-Mansour et al., 1994; Coffey et al., 2016; Forsyth et al., 1987; Ghabchi et al., 2013; Karimzadeh, Sabeti, Burde, et al., 2020; Karimzadeh, Sabeti, Tabkhi, et al., 2020). Moreover, most of the developed deterioration models in the literature did not consider such interrelationships and investigated each asset's condition independent from its neighbors. For example, Abaza (2017), Chopra et al. (2018), and Gao et al. (2012) forecasted the condition of pavements only based on the historical condition of pavement segments. This trend was also pursued for other asset types such as markings, signs, barriers, and culverts (Chimba et al., 2014; Halmen et al., 2008; Immaneni et al., 2009; Malyuta, 2015; Sitzabee et al., 2012).

2.4 Contributing Factors to Deterioration of Roadway Assets

In the literature review, several factors were identified that impact roadway assets' deterioration. For example, the role of material, traffic loading, weather condition, and historical maintenance on the degradation patterns of multiple assets was highlighted in several studies (Anyala et al., 2014; Bannour et al., 2017; Ford et al., 2012; Hong et al., 2010; Labi et al., 2003; Prozzi et al., 2017; Ré et al., 2011; C. Wang et al., 2016). Materials used in the construction or production of different asset types identify their resistance against degradations. Also, the structural characteristics of assets influence their deterioration trend. For example, the thickness of flexible pavements and the binder type are two main factors that impact the resistance of the pavement layer against degradation (Anyala et al., 2014). Moreover, traffic volume is a major factor that contributes to the deterioration of pavements and markings (Craig III et al., 2007). In addition, a variety of defects on different asset types are resulted from environmental factors. For instance, the deterioration trend of traffic signs and ditches are impacted by moisture, temperature, solar radiation, and precipitation (Markow, 2007).

However, most studies developed deterioration models where only a selected number of contributing factors were considered based on experts' opinions. Also, as a significant factor that improves the condition of highway assets, historical maintenance activities have received marginal attention when it comes to developing prediction models (Karimzadeh & Shoghli, 2020). For example, Chopra et al. (2018) studied pavements degradation by considering only traffic loadings data. As another example, Swargam (2004) proposed a prediction model to forecast the retroreflectivity of traffic signs utilizing only the age and characteristics of signs. Later on, Elwakil et al. (2014) investigated the effects of traffic, age, and the number of snow removals on the condition of pavement markings.

To summarize, three main gaps and limitations in the previous predictive models were identified: (i) limitations in predictive frameworks given the inadequate available inspection data, (ii) marginal investigation of possible interrelations between neighboring assets in the majority of studies, and (iii) subjective selection of contributing factors to degradation of assets.

CHAPTER 3: METHODOLOGY

3.1 Overview

After highlighting the main limitations and gaps in the literature, this section moves on to explain the methodology for filling the gaps and mitigating the challenges in previous studies. Figure 1 shows the proposed framework of this study. Within this framework, the required data were extracted from pertinent and available resources under four categories: weather, traffic, historical maintenance, and inspection. Then, the collected datasets were cleaned to ensure that they are free of missing information and errors. Finally, a preprocessing step was performed on the cleaned datasets to get them prepared for the analysis. After making the data ready, Kernel Density Estimation (KDE) Analysis was used to calculate the density of defects in the unit of area (defects per square mile) for all road segments. This parameter was named Risk Score (RS) as it corresponds to the probability occurrence of the considered defects estimated based on their densities in different parts of roadways. Then, in the machine learning component of the framework, prediction models were developed and validated for predicting RSs based on previous historical RSs and other considered contributors (i.e., weather, traffic, maintenance). To this end, multiple linear and nonlinear ML algorithms were used. Then, the performance of the utilized algorithms was evaluated and the most accurate RS prediction model was chosen through a comparative study. Ultimately, the results were finalized and visualized.

To measure the proposed framework's efficiency, 242 miles (389 kilometers) of Interstate I-81, I-77, and I-381 highways were used as the case study, as shown in Figure 2. The roadways that consisted of mainlines and ramps were firstly split into 2420 segments. To do so, road segments were defined as sections with one-tenth of a mile (161 meters) length covering fence-to-fence of the right of way.

A variety of asset classes can be found in each segment of the case study roadways. For example, pavements, signs, markings, ditches, and guardrails. In addition, each asset class has its corresponding defect types. For instance, pothole and patch are two common defects that happen on flexible pavements. Also, erosion, obstruction, and cracking are the common defects that might take place on paved ditches. In this study, six adjacent asset types were considered to incorporate the interrelation of neighboring assets into the case study investigations. Flexible pavements, paved ditches, unpaved ditches, slopes, small pipes and box culverts, and under drain pipes and edge drains were the considered asset types in this study. This set of asset classes was chosen because of the potential interrelations between their condition. These interrelations are mainly attributed to the fact that they all belong to a continuous drainage system and are located at a close distance (Karimzadeh, Sabeti, Burde, et al., 2020). For each of the six selected asset types, the considered corresponding defects were also taken into consideration.

Then, the proposed framework was applied to predict the Risk Scores (RSs) of three defects (erosion, obstruction, and cracking) on all paved ditches in each road segment. The condition of the nearby assets was also incorporated in the prediction model and their corresponding defects were considered in the analysis. Next, the data between 2015 and 2020 was obtained with the Fiscal Year (FY) being the unit of time. FY is the 12-month period that ends on June 30th of each year. For example, FY2016 refers to the 12-month period between July 1st, 2015, and June 30th, 2016. This time interval encompasses all maintenance activities that were performed during one year before the annual inspection in 2016, as well as the recorded weather and traffic attributes in this period. Accordingly, the considered timeframe was then split into five periods: FY2016, FY2017, FY2018, FY2019, and FY2020.

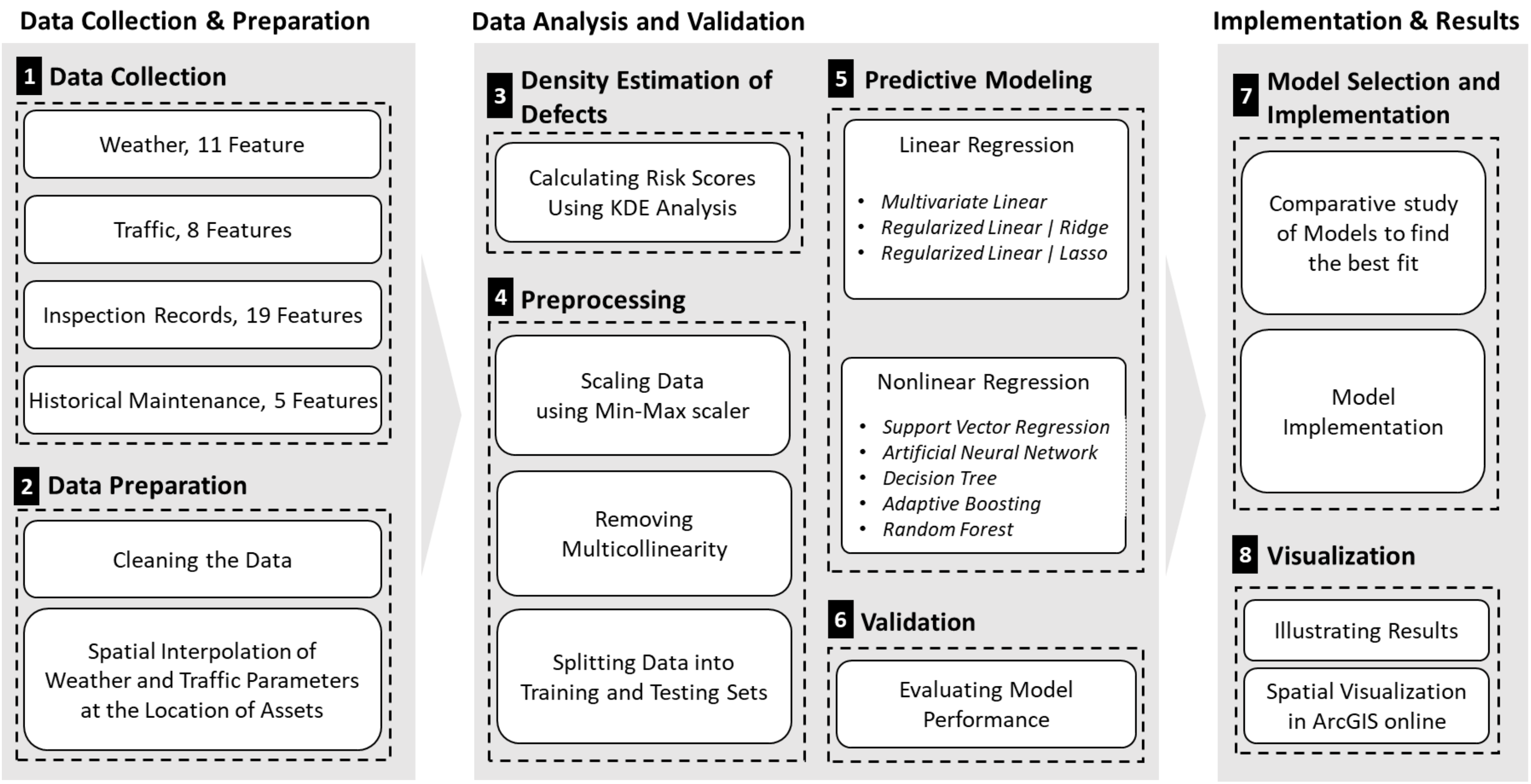

Data Collection & Preparation
1 Data Collection
Weather, 11 Feature
Traffic, 8 Features
Inspection Records, 19 Features
Historical Maintenance, 5 Features
2 Data Preparation
Cleaning the Data
Spatial Interpolation of Weather and Traffic Parameters at the Location of Assets
Data Analysis and Validation
3 Density Estimation of Defects
Calculating Risk Scores Using KDE Analysis
4 Preprocessing
Scaling Data using Min-Max scaler
Removing Multicollinearity
Splitting Data into Training and Testing Sets
5 Predictive Modeling
Linear Regression
• Multivariate Linear
• Regularized Linear | Ridge
• Regularized Linear | Lasso
Nonlinear Regression
• Support Vector Regression
• Artificial Neural Network
• Decision Tree
• Adaptive Boosting
• Random Forest
6 Validation
Evaluating Model Performance
Implementation & Results
7 Model Selection and Implementation
Comparative study of Models to find the best fit
Model Implementation
8 Visualization
Illustrating Results
Spatial Visualization in ArcGIS online

Figure 1 - Framework of the study

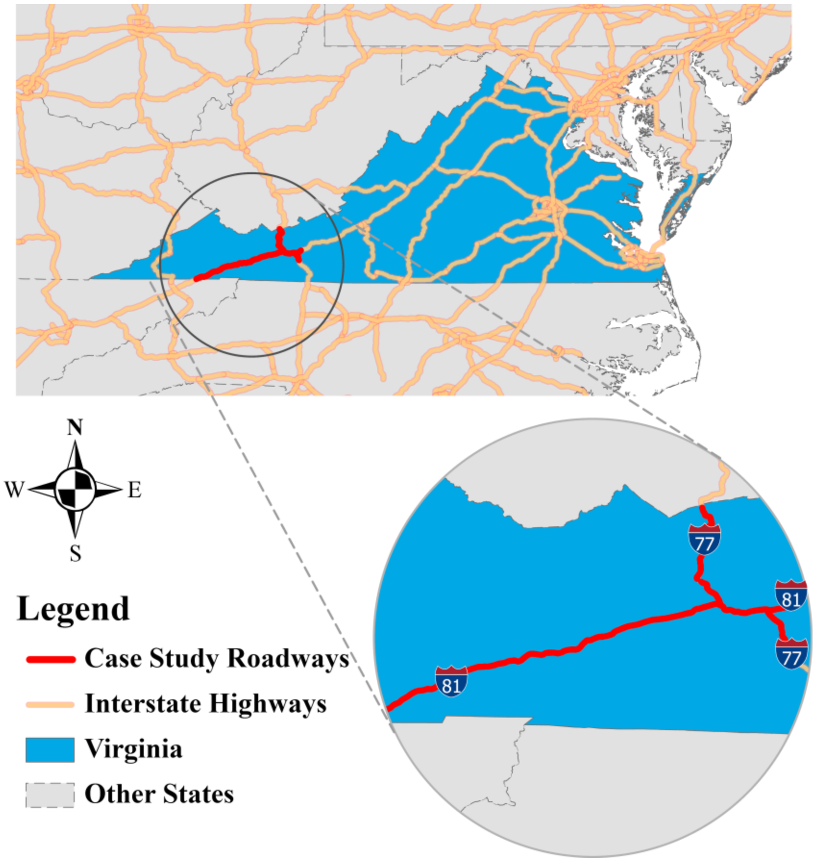


Figure 2 - Case study roadways

3.2 Data Collection and Preparation

The weather data was collected from the National Oceanographic and Atmospheric Administration (NOAA) database. Multiple features were used to include possible fluctuations of weather into the framework. Table 1 provides a full list of the considered weather features utilized in this study. In addition, the statistical description of the considered weather features is summarized in Table 2.

Table 1 - Weather features considered in the study

Index	Parameter	Definition
1	TMAX	Annual maximum daily temperature (° F)
2	TMIN	Annual minimum daily temperature (° F)
3	TMAXMIN	Annual average of daily max-min temperature difference (° F)
4	DWT32	Number of days with minimum temperature<32° F (0° C) in a year
5	DWT80	Number of days with maximum temperature>80° F (26.7° C) in a year
6	DWTMXN30	Number of days with Tmax-Tmin> 30° F (16.7° C) in a year
7	DSNW	Number of days with snow depth >1 inch (2.54 cm) in a year
8	EMSD	Maximum annual daily snow depth (inch)
9	EMXP	Maximum annual daily precipitation depth (inch)
10	PRCP	Total annual precipitation (inch)
11	SNOW	Total annual snow depth (inch)

Traffic data were extracted from the Virginia Department of Transportation (VDOT)'s public portal. The traffic data, in the form of a shapefile, contained the location of roadways in Virginia and their corresponding traffic attributes. Figure 3 illustrates the shapefile of Annual Average Daily Traffic (ADT). A variety of traffic features were used in the framework and the corresponding data were examined for the existence of missing information to ensure the considered dataset is error-free. Table 3 provides the full list of the considered traffic features in this study. Also, a descriptive statistical summary of the traffic features is provided in Table 4.

Table 2 – Summary of the statistical description of weather features

Parameter	Min	Max	Average	Median	Standard Deviation
TMAX	86.2	95.0	90.8	91.0	1.6
TMIN	-5.4	12.5	3.1	3.0	5.2
DWT32	70.9	127.5	100.8	100.8	10.7
DWT80	38.2	112.7	82.2	81.5	14.9
DSNW	0.0	7.0	3.7	4.5	1.8
EMSD	0.5	15.5	6.7	5.9	4.2
EMXP	1.8	4.9	2.8	2.7	0.6
PRCP	40.5	75.0	52.0	50.9	7.0
SNOW	0.8	41.4	17.4	16.4	10.8
DWTMXN30	30.6	72.0	53.4	57.8	11.2
TMAXMIN	20.4	24.0	22.1	22.3	0.9

Table 3 - Traffic features considered in the study

Index	Parameter	Definition
1	ADT	Average daily traffic
2	AAWDT	Average annual weekday traffic
3	ADT_4	Average daily traffic of 4-tire vehicles
4	ADT_BU	Average daily traffic of buses
5	ADT_TR	Average daily traffic of trucks with 1 trailer
6	ADT_1	Average daily traffic of trucks with 2 axles
7	ADT_2	Average daily traffic of trucks with 2 trailers
8	ADT_3	Average daily traffic of trucks with 3 axles

Table 4 – Summary of the statistical description of traffic features

Parameter	Min	Max	Average	Median	Standard Deviation
ADT	440	58000	30473	30000	11789
AAWDT	440	57000	29169	27000	11600
ADT_4	0	45775	23588	23359	9305
ADT_BU	0	375	175	175	77
ADT_TR	0	465	234	243	95
ADT_1	0	432	240	243	99
ADT_2	0	11793	5661	5404	2609
ADT_3	0	890	409	397	198

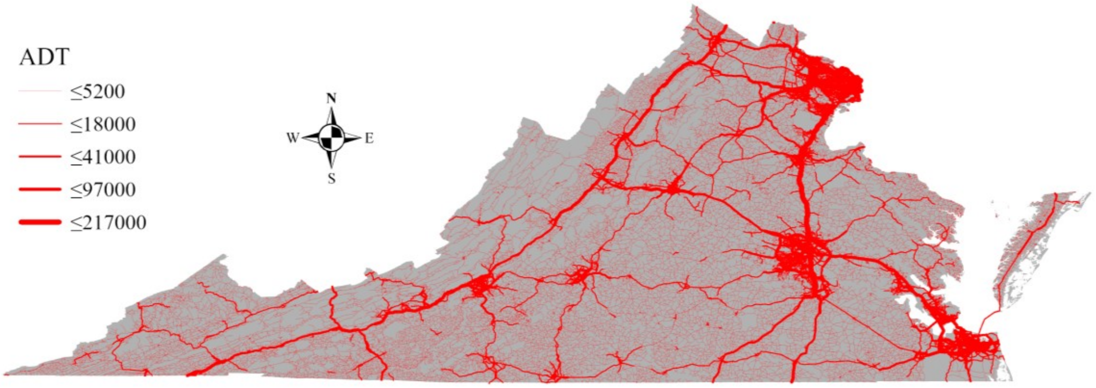


Figure 3 - Average Daily Traffic (ADT) of roadways in Virginia

The case study's historical maintenance information was extracted from a Maintenance Quality Assurance Program (MQAP) that recorded the history of maintenance tasks performed in FY2016, FY2017, FY2018, FY2019, and FY2020. Each record in the MQAP includes the time, type, and location of each maintenance task. The tasks that are relevant to the selected asset (i.e., paved ditch) were chosen. The selected maintenance tasks and their descriptions provided in Table 5 were obtained from Virginia DOT maintenance guidelines. This table was created in direct collaboration with a former VDOT maintenance crew with extensive highway maintenance experience. The maintenance expert validated all the obtained maintenance records from the VDOT task orders. After collecting the data, the records that contained missing information were removed. The historical maintenance was treated as a categorical feature with binary values: if a maintenance task was performed in a fiscal year, its corresponding feature would be 1, otherwise 0.

Finally, the inspection data were extracted from the same MQAP that was utilized to obtain maintenance records. In this resource, the conditions of the selected assets were accessible through the recorded data at the time of inspections. The conditions were reported in terms of the level of the corresponding defects of each asset class in 4 levels: very poor, poor, good, and very good. The definition of four classes of conditions for flexible pavements, paved ditches, unpaved ditches, slopes, small pipes and box culverts, and under drain pipes and edge drains are provided in Table 6, Table 7, Table 8, Table 9, Table 10, and Table 11, respectively.

In this study, very poor and poor conditions were aggregated as the observation of the corresponding defect and the need for fixing repair tasks. This classification that highlights the necessity of repairs is aligned with the trigger levels utilized in making maintenance decisions in VDOT. Therefore, very poor and poor conditions for each asset type were merged into the fail class while good and very good into the pass class. The fail condition under a specific defect means that the defect was observed in the considered asset item. Pass condition means that the considered asset item was defect-free under the specific defect type. Later on, the location of the failed asset items was used in presenting the spatial distribution of the defects. A summary of the selected assets, their corresponding defects, and the number of observed defects in different years of inspection are provided in Table 12.

Table 5 - Maintenance activities performed on paved ditches (VDOT, 2014)

Index	Code	Maintenance Name	Description
1	M70141	Hand Cleaning	Hand cleaning of drainage assets, traffic control devices, shoulders, tunnels, ferries, etc. Cleaning with manual tools (shovels, pickaxes, etc.). Cleaning without the use of machinery.
2	M70142	Machine Cleaning / Mechanical Sweeping	Machine cleaning or sweeping of drainage assets such as pipes, ditches, etc.; tunnels; roadside assets such as sidewalks, truck ramps, pedestrian trails, walls, etc.; traffic assets such as rumble strips; pavement assets including roads, and paved shoulders, etc. Also, to be used for cleaning when using pressurized water such as power washing.
3	M71152	Seeding, Fertilizing, Mulching (Serv)	Seeding, fertilizing, mulching, sodding, soiling, spreading lime. The cyclical and regular replacement and maintenance of vegetation to combat erosion.
4	M72223	Concrete Patching / Repair - Drainage (Serv)	Patching holes, blow-ups, and other irregularities on concrete surfaces for drainage assets. This activity includes cutting and removing damaged concrete and patching concrete areas.
5	M72224	Concrete Joint Repair - Drainage (Serv)	Removing and replacing joint filler, pouring joints, trimming joints, joint patching, and other maintenance of drainage concrete joints.

Since identifying defects' hotspots is an objective of this study, the total number of observed defects should be adequate to find the areas with the concentration of those defects. Therefore, when the number of defects for a certain asset type is zero, identifying hotspots makes no sense. As shown in Table 12, the number of recorded defects for all the considered asset types was not zero however some of the defects have been observed rarely.

Table 6 - Condition descriptions for flexible pavements (VDOT, 2014)

Defect	**Very Good**	**Good**	**Poor**	**Very Poor**
Pothole	No potholes or signs of distressed asphalt (i.e. troughing, rutting, heaving)	No potholes	One pothole present	More than one pothole present
Patch	No patches or all patches with smooth ride	No patches or distresses greater than or equal to ½ inch higher or lower than surrounding pavement	N/A	Patches or distresses greater than ½ inch higher or lower than surrounding pavement

Table 7 - Condition descriptions for paved ditches (VDOT, 2014)

Defect	**Very Good**	**Good**	**Poor**	**Very Poor**
Erosion	No undermining (no erosion present)	Undermining less than or equal to 3 inches deep	Undermining greater than 3 inches deep	Evidence of structural damage or collapse
Obstruction	No obstruction	Less than 25% of the cross section obstructed	Greater than or equal to 25% of the cross section is obstructed	Greater than or equal to 50% of the cross section is obstructed
Cracking	No cracking	Less than or equal to 10% of surface area showing cracking greater than ½ inch wide	Greater 10% of surface area showing cracking greater than ½ inch wide	Greater than or equal to 25% of surface area showing cracking greater than ½ inch wide

Table 8 - Condition descriptions for unpaved ditches (VDOT, 2014)

Defect	**Very Good**	**Good**	**Poor**	**Very Poor**
Erosion	No erosion.	Erosion less than or equal to 8 inches deep	Erosion greater than 8 inches deep	Erosion greater than 8 inches deep over more than 25% of the length
Obstruction	No obstruction.	Less than or equal to 25% of the cross section obstructed	More than 25% of the cross section obstructed	More than or equal to 50% of the cross section obstructed

Table 9 - Condition descriptions for slopes (VDOT, 2014)

Defect	**Very Good**	**Good**	**Poor**	**Very Poor**
Erosion	No slope erosion	Less than or equal to 8 inches deep erosion	Erosion along slope greater than 8 inches deep	Multiple erosion along slope greater than 8 inches deep
Erosion Pattern	N/A	No pattern of erosion that endangers the stability of the slope	Pattern of erosion that endangers the stability of less than 25% of the slope	Pattern of erosion that endangers the stability of greater than or equal to 25% of the slope
Lower Slope	Slope matches paved shoulder throughout segment	Less than or equal to 20% of slope length greater than 2 inches lower than paved shoulder	Greater than 20% of slope length greater than 2 inches lower than paved shoulder	Greater than or equal to 40% of slope length greater than 2 inches lower than paved shoulder
Higher Slope	Slope matches paved shoulder throughout segment	Less than or equal to 20% of slope length greater than 2 inches higher than paved shoulder	Greater than 20% of slope length greater than 2 inches higher than paved shoulder	Greater than or equal to 40% of slope length greater than 2 inches higher than paved shoulder

Table 10 - Condition descriptions for small pipes and box culverts (VDOT, 2014)

Defect	Very Good	Good	Poor	Very Poor
Pipe Obstruction	No obstructions that impede flow	Less than or equal to 25% diameter closed	Greater than 25% of diameter closed	Greater than or equal to 50% of diameter closed
Pipe Joint	Pipe inline and functioning as designed	No separated or damaged joints	Joint separation or mis-alignment is visible from the pipe opening	Joint separation with joint exposed
Pipe Erosion	No erosion at pipe end.	Less than or equal to 2 feet deep erosion within 1 foot of outfall	Greater than 2 feet deep erosion within 1 foot of outfall. No undermining of pipe end	Greater than 2 feet deep erosion and pipe end is undermining
Pipe Vegetation	N/A	No vegetation impacting flow	N/A	Vegetation is affecting the flow of water
End Wall	N/A	End walls and end section intact or free of damage	End walls or end section intact with minor separation or misalignment.	End walls or end section damaged, separated or missing

Table 11 - Condition descriptions for under drains and edge drains (VDOT, 2014)

Defect	Very Good	Good	Poor	Very Poor
Drain Outlet	No damage or deterioration to outlet pipe	Under drain pipes intact	Damage or deterioration to outlet pipe that effects flow	Outlet pipe damaged or deteriorated. Non-functioning
Drain Obstruction	No blockage	Less than or equal to 10% blockage of the diameter or end protection	More than 10% blockage of the diameter or end protection	More than or equal to 25% blockage of the diameter or end protection
End Protection	End protection intact	Damaged but functioning end protection	Wire mesh missing	Damaged non-functioning end protection

Table 12 - Recorded number of defects on selected asset types in different years

Asset Type	Defect 1						Defect 2						Defect 3						Defect 4						Defect 5					
	2015	2016	2017	2018	2019	2020	2015	2016	2017	2018	2019	2020	2015	2016	2017	2018	2019	2020	2015	2016	2017	2018	2019	2020	2015	2016	2017	2018	2019	2020
Flexible Pavement	Pothole						Patch																							
	55	82	106	51	79	102	9	25	5	4	6	10																		
Paved Ditch	Erosion						Obstruction						Cracking																	
	63	55	59	34	33	29	80	54	44	36	43	27	8	10	4	9	9	2												
Unpaved Ditch	Erosion						Obstruction																							
	8	9	8	2	10	2	18	20	12	4	7	1																		
Slope	Erosion						Erosion-pattern						Lower-slope						Higher-slope											
	14	6	15	3	8	2	7	6	5	2	8	2	33	31	15	7	16	7	35	12	7	1	10	4						
Small Pipes And Box Culverts	Pipe-Obstruction						Pipe-Joint						Pipe-Erosion						Pipe-Vegetation						End-Wall					
	11	18	11	24	9	8	2	6	6	1	1	0	3	5	1	1	3	1	10	11	5	10	6	2	17	15	13	12	9	8
Under Drains And Edge Drains	Drain-Outlet						Drain-Obstruction						End-Protection																	
	1	1	5	3	16	1	16	17	20	22	32	6	4	7	13	7	7	3												

Note: a few 🟥 🟧 🟧 🟨 🟩 🟩 many

3.2.1 Spatial Interpolation of Weather and Traffic Data

Since traffic data was in the form of a shapefile, ESRI's ArcGIS spatial join tool was utilized to extract traffic features at the location of the case study road segments. In addition, to extend known measurements of weather parameters (accessible at the location of weather stations) to roadway segments where no measurements were taken, a spatial interpolation technique was required. Global interpolation (trend surfaces and regression models), local interpolation (Thiessen polygons, inverse distance weighting, splines), geostatistical methods (simple kriging, ordinary kriging, block kriging, directional kriging, universal kriging, and co-kriging), and mixed methods (combined global, local and geostatistical methods) are common methods of spatial interpolation for weather parameters (Vicente-Serrano et al., 2003). Ordinary kriging was leveraged to interpolate the value of each weather feature on the considered road segments because of its proven performance in interpolating weather features (da Silva et al., 2019; Frazier et al., 2016; Plouffe et al., 2015).

Ordinary kriging uses the core idea that observations in neighboring locations are more related than those of farther ones. In this algorithm, based on the observed values of a certain parameter in several points, the parameter is interpolated and calculated at unobserved points. For this purpose, spatial autocorrelation is assessed and represented by semivariogram and statistical functions are utilized to fit the model autocorrelations and the semivariogram. Ordinary kriging, with minimum variance, is an unbiased interpolation method that considers statistical relationships and autocorrelations between the observed points (de Amorim Borges et al., 2016). In that way, this algorithm uses a linear combination of observed values for interpolation purposes, as shown in Equation 1.

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)\,; \quad \sum_{i=1}^{n} \lambda_i = 1 \tag{1}$$

Where $\hat{Z}(s_0)$ is the interpolated value at location s_0, λ_i; i=1,…,n are linear coefficients and $Z(s_i)$ i=1,…n are observations at locations s_i. Moreover, the estimator in ordinary kriging is unbiased, that means the predicted value at the location s_i is equal to the observed value at that location, as shown in Equation 2.

$$E\big[\hat{Z}(s_0)\big] = E[Z(s_0)] \tag{2}$$

Additionally, to interpolate the values at unobserved points, the variance of the prediction error, as shown in Equation 3, should be minimized.

$$Min\left\{var[\sum_{i=1}^{n} \lambda_i Z(s_i) - z(s_0)]\right\} \tag{3}$$

3.3 Density Estimation of Defects

Kernel Density Estimation (KDE) is a common tool in developing risk maps in different fields. For example, KDE has been widely used in transforming historical forest fire data into a smooth and continuous 2-D surface that shows high-risk areas to wildfires (Kuter et al., 2018). In addition, KDE is a common tool in analyzing road traffic accidents and providing associated risk-maps for the transportation management sector. Space-time plots additionally rely on KDE in finding hotspots of the probable accidents (Rahman et al., 2018).

DOTs annually inspect only a portion of roads as the representative of all parts of roadways. They usually divide roads into inspection units (i.e., segments) and randomly select a subset of segments for the annual inspection. Therefore, a complete set of historical data for all years of inspections is usually unavailable on each road segment. ‹To augment the available inspection data, the KDE was used to generate a continuous distribution of the density of defects for all segments of the case study in each year based on the sampled inspections, as shown in Figure 4. The location of the observed defects is shown in Figure 4(a), and the corresponding densities of defects are presented in Figure 4(b). The lowest density of defects is displayed with dark blue color, and the highest densities are colored in darker red. The KDE provides the distribution of the defect densities per unit area, which corresponds to the probability of occurrence of a particular defect. As a result, the term "Risk Score (RS)" was used for representing the outcome of the KDE in this study. By placing a kernel over each observation and summing all individual kernels over each point, the distribution of density is achievable (T. K. Anderson, 2009). Equation 4 shows the density estimation in a two-dimensional space using KDE.

$$f(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{d_i}{h}\right) \tag{4}$$

Where *f(x,y)* is the density estimation at the location *(x,y)*; *n* is the number of points or observations; *h* is the kernel bandwidth; *K* is the kernel weight function; and d_i is the distance between the location *(x,y)* and the i^{th} point or observation. In Equation 4, selecting kernel bandwidth is a subjective task (T. K. Anderson, 2009; Thakali et al., 2015).

However, several recommendations are available in the literature, such as Silverman's rule-of-thumb (Silverman, 1986), or selecting a bandwidth equal to 9 times the median of the nearest neighbor distances between the considered points (Chainey et al., 2013). Both methods were used in the case study, and the estimations for the bandwidth were 4.93 and 1.25 miles for the Silverman and nearest-neighbor-based methods, respectively. Finally, the average of these values was utilized in the analysis, and the kernel bandwidth was chosen as 3.1 Miles (5 Kilometers).

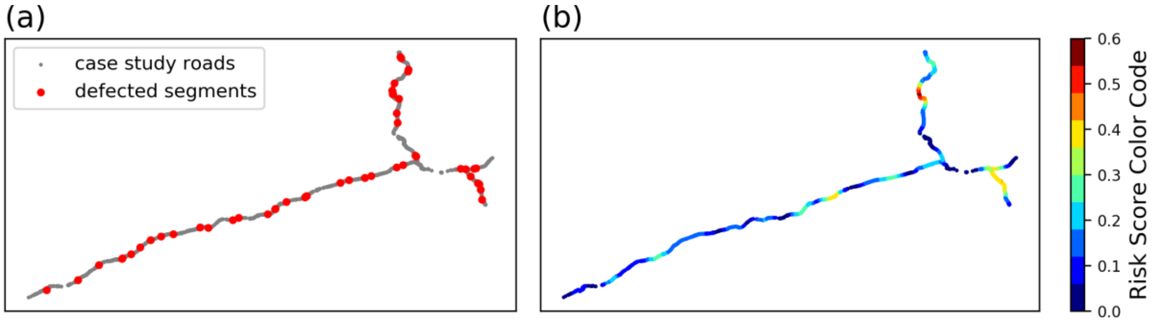


Figure 4 - (a) Spatial distribution of observed defects on paved ditches (b) Corresponding Risk Scores (RSs) of defects based on KDE analysis

After preparing the complete set of RSs of different defects for the six considered nearby assets an input dataset was created that contained all predictors (i.e. weather, traffic, maintenance, and RSs for all segments). Figure 5 presents the concept of the prediction proposed in this study. Figure 5 shows that the combination of weather, traffic, and maintenance for one year, as well as the prior year RSs of paved ditches and all other considered neighboring assets are used as the inputs to predict a particular defect's RS in the next year (*Year*$_2$).

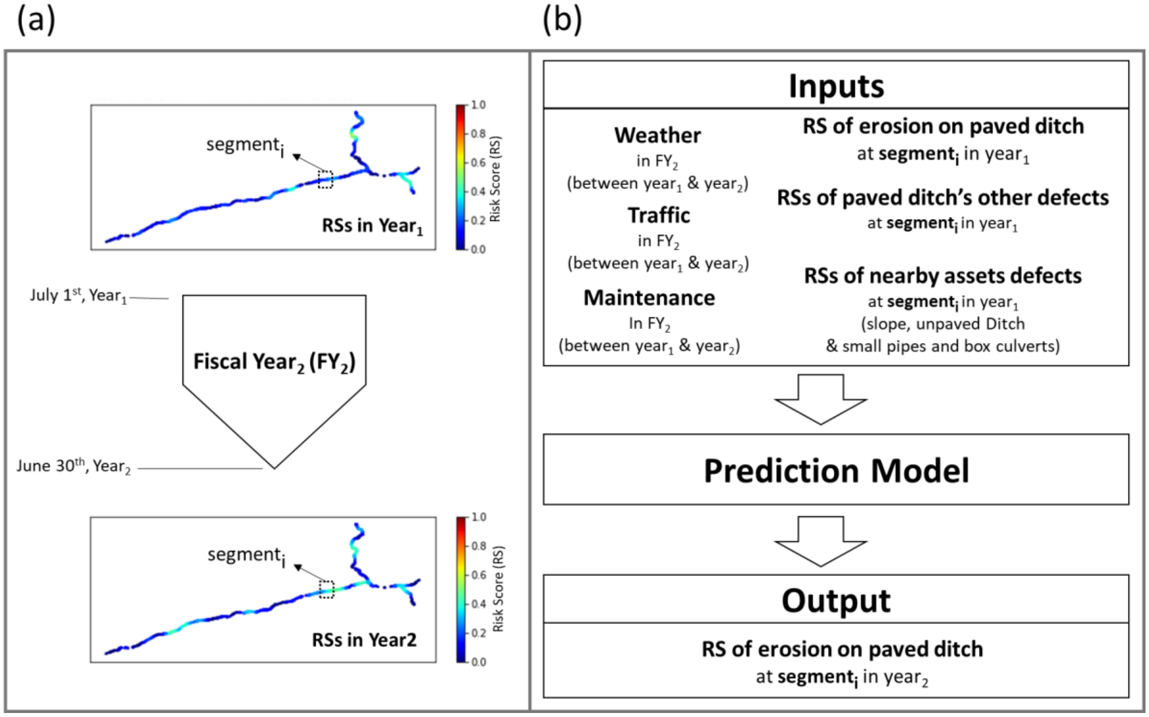

(a)
(b)
segment$_i$
RSs in Year$_1$
Risk Score (RS)
1.0
0.8
0.6
0.4
0.2
0.0
July 1st, Year$_1$
Fiscal Year$_2$ (FY$_2$)
June 30th, Year$_2$
segment$_i$
RSs in Year2
Inputs
Weather
in FY$_2$
(between year$_1$ & year$_2$)
Traffic
in FY$_2$
(between year$_1$ & year$_2$)
Maintenance
In FY$_2$
(between year$_1$ & year$_2$)
RS of erosion on paved ditch
at segment$_i$ in year$_1$
RSs of paved ditch's other defects
at segment$_i$ in year$_1$
RSs of nearby assets defects
at segment$_i$ in year$_1$
(slope, unpaved Ditch
& small pipes and box culverts)
Prediction Model
Output
RS of erosion on paved ditch
at segment$_i$ in year$_2$

Figure 5 - Schematic procedure of predicting Risk Scores (RSs) at segment$_i$ (a) Transformation of RSs in a fiscal year (b) Proposed prediction framework

For example, in this concept, to predict the RS of erosion on paved ditches at the end of FY2017, weather, traffic, and maintenance in FY2017, RSs of paved ditch's erosion, obstruction, and cracking at the end of FY2016, and RSs of neighboring asset types at the end of FY2016 would be the inputs. Accordingly, the series of inputs for all of the considered fiscal years were created, and then the machine learning component was used for predicting the RSs of paved ditches under the three selected defects.

Before using the input dataset in building prediction models, it needs to be cleaned from records containing non-logical transition of RSs during a fiscal year. In this study, a non-logical record refers to a segment that encountered a decreasing trend in its risk score in a fiscal year without performing any corrective maintenance on the segment in that year.

It means that based on the observed data, sometimes defects densities in some areas might decrease. However, a corrective maintenance might not be recorded in those areas, which is not logical. In order to select the corresponding corrective maintenance that fixes a certain defect, different maintenance types, as shown in Table 5, were considered. To this end, M_72223 and M_72224 were chosen as the corrective maintenance tasks performed to repair erosion and cracking on paved ditches. However, M_70141 and M_70142 were also taken into account as corrective maintenance tasks for fixing obstructions. Therefore, the filter was performed based on the selected corrective maintenance tasks and the decreasing trend of RSs on each segment. Finally, the non-logical records were filtered, and the remaining data were used in developing prediction models.

Figure 6 illustrates the spatial distribution of segments with logical data of erosion RSs in different fiscal years. According to this figure, the remained segments almost cover all parts of the case study roadways and only two segments were completely removed from the input dataset after performing the filter. Figure 7 shows the remained segments in the filtered dataset for developing prediction models for obstruction RS. As shown in this figure, remained segments covered all parts of the case study roadways and no segment was removed after applying the filter. Furthermore, Figure 8 presents the filtered segments for building the prediction model for cracking RSs. With respect to the figure, all segments were remained after conducting the considered filter.

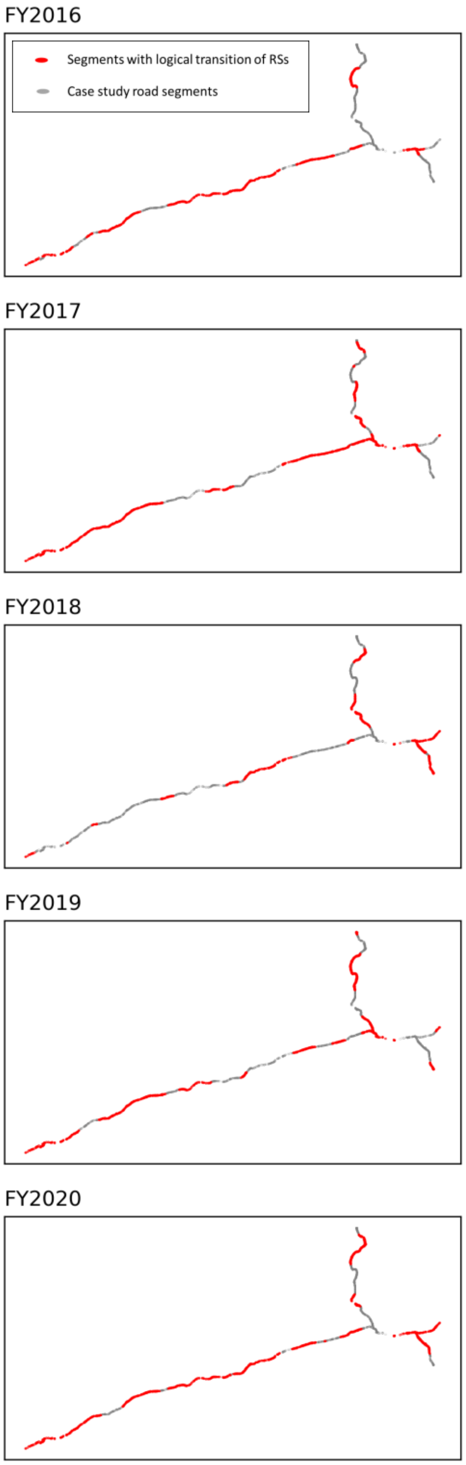


Figure 6 – Spatial distribution of remaining segments after filtering non-logical records for developing prediction model for erosion RS

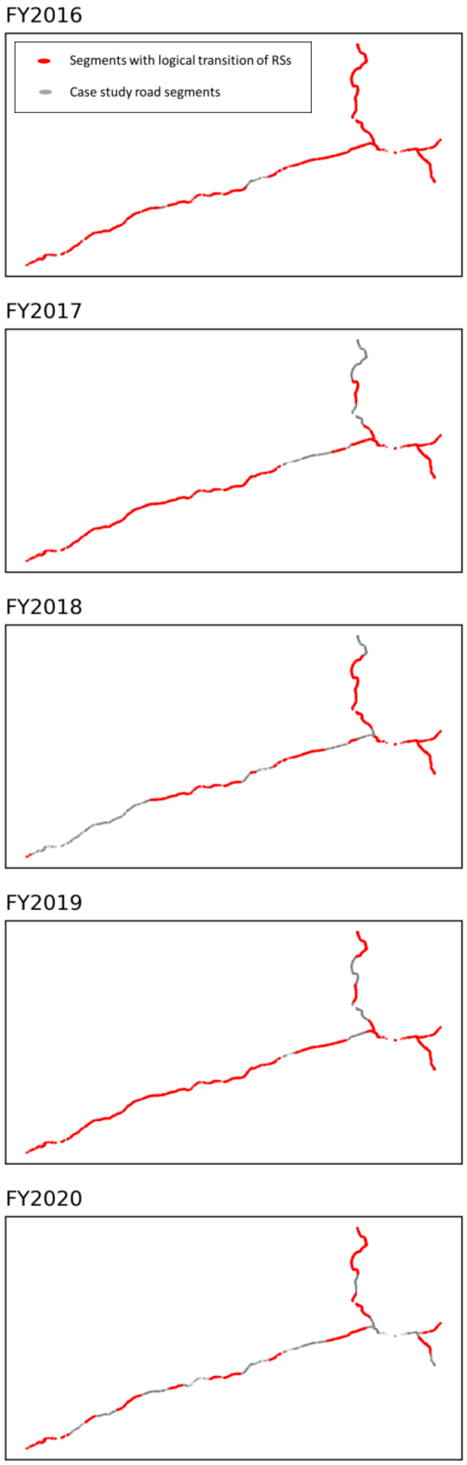


Figure 7 – Spatial distribution of segments after filtering non-logical records for developing prediction model for obstruction RS

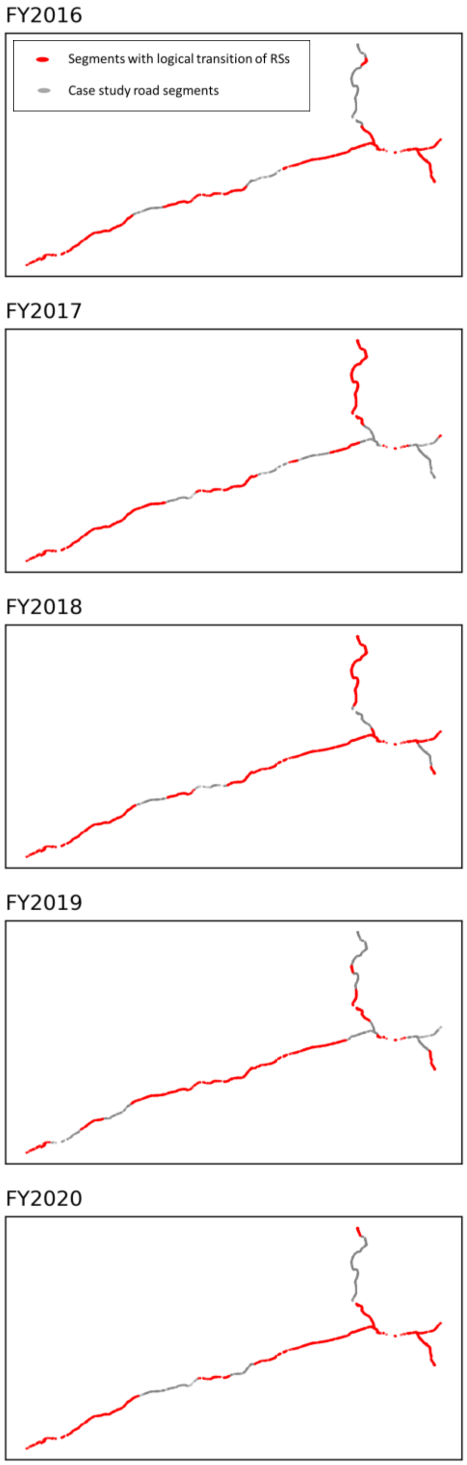


Figure 8 – Spatial distribution of segments after filtering non-logical records for developing prediction model for cracking RS

3.4 Preprocessing Data

Before developing machine learning-based prediction models, the data needs to be preprocessed and get ready for the next step. The preprocessing component of the framework contains three different modules. In the first module and as an attempt to remove the potential bias in the results, the input data were normalized using a min-max scaler (Aksoy et al., 2001). The utilized scaler linearly maps the continuous features of the input dataset (i.e., weather, traffic, and RSs) into a new continuous space between 0 and 1. This scaler was applied to each one of the continuous features separately.

In the next module, the multicollinearity in the input dataset was removed. Multicollinearity refers to a scenario in which high correlations among multiple features of a dataset are observed, which could potentially bias the outcomes (Yoo et al., 2014). In this framework, the multicollinearity was removed using a correlational investigation. The considered feature space is a mixed dataset consisting of both continuous and categorical (e.g., historical maintenance) features. Therefore, the feature reduction in this framework is performed in three steps. Firstly, the correlation between continuous features was measured. To do so, the features with absolute Pearson correlation coefficients greater than 0.9 were grouped and each group was represented with only one feature (Bujang et al., 2017; Yoo et al., 2014). Next, the Chi-square test was used to examine the correlation among categorical features. Any pair of features with a p-value larger than 0.05 was considered highly correlated and represented with one of them. Ultimately, the dependence between the reduced continuous and categorical spaces was investigated using the point-biserial correlation coefficient. The attributes with a correlation bigger than 0.9 were grouped and only one feature was considered as the representative of the groups.

In order to validate the prediction models, the data is split into training and testing sets. Next, prediction models are trained using the training set. Then, the testing set is utilized in measuring the performance of the model on unseen data. To this end, 60% of the data was used to develop the prediction model and the remaining 40% of the data was utilized to assess the performance of the developed model. Figure 9 further illustrates the training and testing process of model preparation.

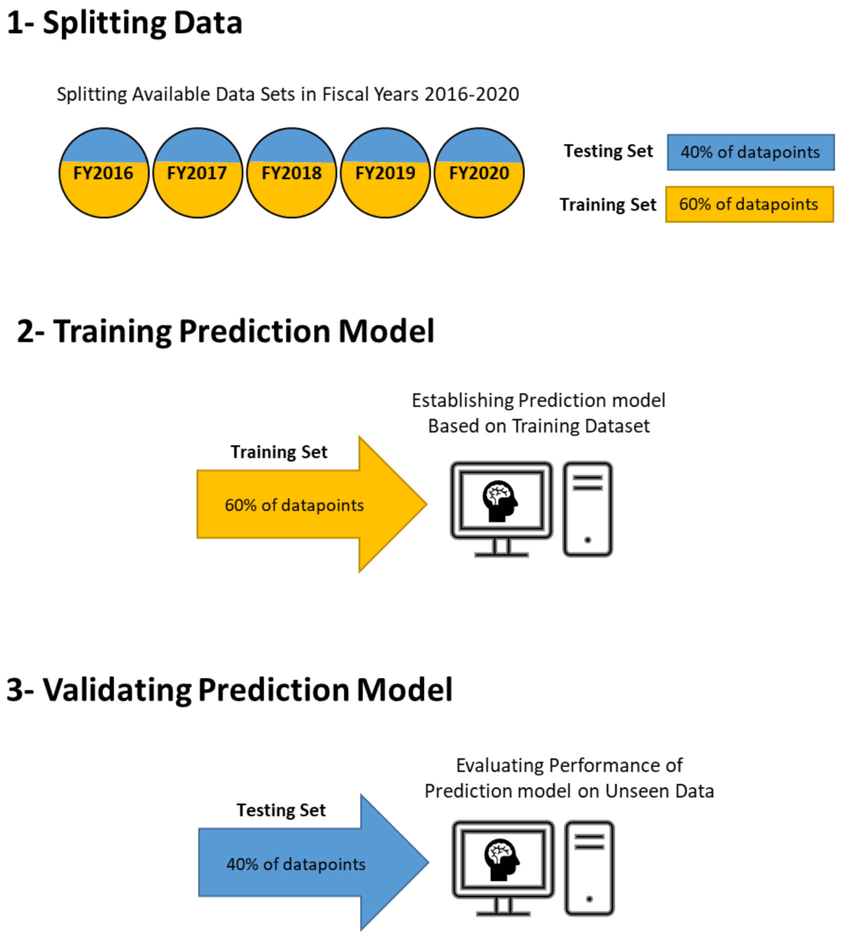

1- Splitting Data
Splitting Available Data Sets in Fiscal Years 2016-2020
FY2016
FY2017
FY2018
FY2019
FY2020
Testing Set
40% of datapoints
Training Set
60% of datapoints
2- Training Prediction Model
Establishing Prediction model Based on Training Dataset
Training Set
60% of datapoints
3- Validating Prediction Model
Evaluating Performance of Prediction model on Unseen Data
Testing Set
40% of datapoints

Figure 9 - Splitting data into training and testing sets, model training, and model validation processes

3.5 Predictive Modeling

A series of ML algorithms were used to predict risk scores, given the reduced feature inputs. Multiple ML-based linear and nonlinear models were used to find the best fit for the case study and also to run a comparative analysis. To this end, three linear models were selected: Multivariate Linear Regression (MLR), Regularized Regression using Ridge (RR), and Regularized Regression using Lasso (RL). In the nonlinear category, five models were considered: Support Vector Regression (SVR), Artificial Neural Network (ANN), and decision tree-based algorithms including Decision Tree (DT), Adaptive Boosting (ADB), and Random Forest Regression (RFR). A brief introduction over each one of the models is provided below.

3.5.1 Linear regression

3.5.1.1 Multivariate Linear Regression

Multivariate Linear Regression (MLR) is a supervised ML algorithm that models the relationship between one response variable and two or more explanatory variables. This technique fits a linear equation to the observed data points and provides information about correlations between dependent and independent variables. The first goal in most ML techniques is to develop a hypothesis (model) to predict a dependent variable (prediction) based on k independent variables (predictors). Therefore, a set of observations is used to develop the hypothesis of the MLR model that can be presented as Equation 5:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k \tag{5}$$

Where $h_\theta(x)$ is the prediction model, x_i's are the predictors, θ_0 is the intercept, and $\theta_i{}'s$ are regression coefficients. In the MLR model, the cost function defined in Equation 6 should be minimized, so that the coefficients can be found:

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \theta_0 - \sum_{j=1}^{k} x_{ij}\theta_j)^2 \tag{6}$$

3.5.1.2 Regularized Linear Regression

Given the interpretability and simplicity of the MLR method, it was widely used to build prediction models in different fields (Immaneni et al., 2009; Leggetter et al., 1994; Yuan et al., 2007). However, sometimes multicollinearity leads to bias and inaccuracy within results. Therefore, filtering independent variables, a.k.a. dimension reduction, and feature selection were proposed to overcome this problem (Yuan et al., 2007). Furthermore, similar to other ML techniques, overfitting is still a possibility in MLR. Overfitting is a situation where the prediction model extremely corresponds to the input data that makes the model inapplicable to fit unseen datasets, which leads to providing unreliable results when new data is used. Therefore, research studies suggested some methods for mitigating multicollinearity and overfitting issues in MLR, such as regularized regression techniques. Regularization is a method that minimizes overfitting in regression models by penalizing and shrinking regression coefficients. To this end, Regularized Ridge (RR) and Regularized Lasso (RL) are two famous regularization techniques that were vastly used in the literature (Friedman et al., 2001) that result in removing irrelevant features in the RL and decreasing weights of these features in the RR. An L2 penalty term is added to the cost function of MLR in the RR regression. The corresponding cost function in this algorithm is shown in Equation 7:

$$J(\theta) = \frac{1}{2n}\left[\sum_{i=1}^{n}(y_i - \theta_0 - \sum_{j=1}^{k} x_{ij}\theta_j)^2 + \lambda \sum_{j=1}^{k} \theta_j^2\right] \tag{7}$$

On the other hand, in the LR technique, an L1 penalty term is added to the MLR cost function, as shown in Equation 8:

$$J(\theta) = \frac{1}{2n}\left[\sum_{i=1}^{n}(y_i - \theta_0 - \sum_{j=1}^{k} x_{ij}\theta_j)^2 + \lambda \sum_{j=1}^{k} \|\theta_j\|\right] \tag{8}$$

In the proposed framework of this study, the aforementioned linear models, i.e. MLR, RR, and LR were compared to investigate their performance in the problem.

3.5.2 Nonlinear regression

Linear models are not always capable of capturing the relationship between dependent and independent variables. In such cases, nonlinear regression techniques are used in developing prediction models. In the proposed framework, some of the most famous nonlinear algorithms were leveraged that were widely used in various fields of studies.

3.5.2.1 Support Vector Regression

Support Vector Regression (SVR) is one of the nonlinear ML algorithms whose success in different fields has been highlighted in the literature (Clarke et al., 2005). In a nonlinear SVR, a kernel transformation function is used to map predictors (*x_i*) to a new high-dimensional space. Then, the optimal function *f(x)* is introduced to represent the relationship between the prediction (*y*) and predictors in the transformed space. The most popular kernel functions that are used to map predictors are linear, polynomial, and gaussian kernels, shown in Equations 9 to 11, respectively.

Linear kernel: $$K(x_i, x_j) = x_i^T x_j \tag{9}$$

Polynomial kernel: $$K(x_i, x_j) = (1 + x_i^T x_j)^d \tag{10}$$

Gaussian kernel (RBF): $$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{11}$$

In the abovementioned equations, x_i and x_j are predictors vector spaces, σ is the variance, and d is the polynomial's dimension (Wu et al., 2009). In this study, all three kernel functions were used in the SVR algorithm and the results were reported for the most accurate one in terms of risk scores (RSs) predictions.

3.5.2.2 Artificial Neural Networks

Artificial Neural Network (ANN) is another ML algorithm that has been used in capturing complicated relationships and patterns among datasets. The architecture of the neural network is a major part of creating an ANN model. To this end, Multi-Layer Perceptron (MLP) is a vastly used architecture in the literature of regression models (Cohen et al., 2003; Yilmaz et al., 2011). In the MLP-ANN model, linear combinations of the outputs from each layer nodes are used to produce the inputs for each perceptron in the next layer and, finally, the prediction for the dependent variable (*y*) in the output layer. The vector of inputs in this model is described as Equation 12 (Rynkiewicz, 2012).

$$x = (x(1), \dots, x(d))^T \in \mathbb{R}^d \tag{12}$$

Wherein, *x* is the input's vector and *d* is the number of vectors in the layer. The parameter vector for hidden layer *i* can be presented by Equation 13.

$$w_i = (w_{i1}, \dots, w_{id})^T \in \mathbb{R}^d \tag{13}$$

Therefore, the Multilayer Perceptron (MLP) for *k* hidden layers can be written as Equation 14.

$$f_\theta(x) = \beta + \sum_{i=1}^{k} a_i \emptyset(w_i^T x + b_i) \tag{14}$$

With $\theta = (\beta, a_1, \dots, a_k, b_1, \dots, b_k, w_{11}, \dots, w_{1d}, \dots, w_{k1}, \dots, w_{kd})$ as the parameter vector of the model and $\emptyset$ the bounded transfer function which is usually a sigmoidal function. In an MLP regression model, Equation 14 is used as the regression function. Hence, all parameters are calculated based on Equation 15.

$$Y = f_\theta(x) + \varepsilon \tag{15}$$

Where *Y* is the prediction vector. The MLP regression algorithm was used in this study and Figure 10 shows the utilized MLP architecture. In this figure, the input layer contains 34 predictors that are the contributors to the reduced feature space (weather, traffic, maintenance, and RSs).

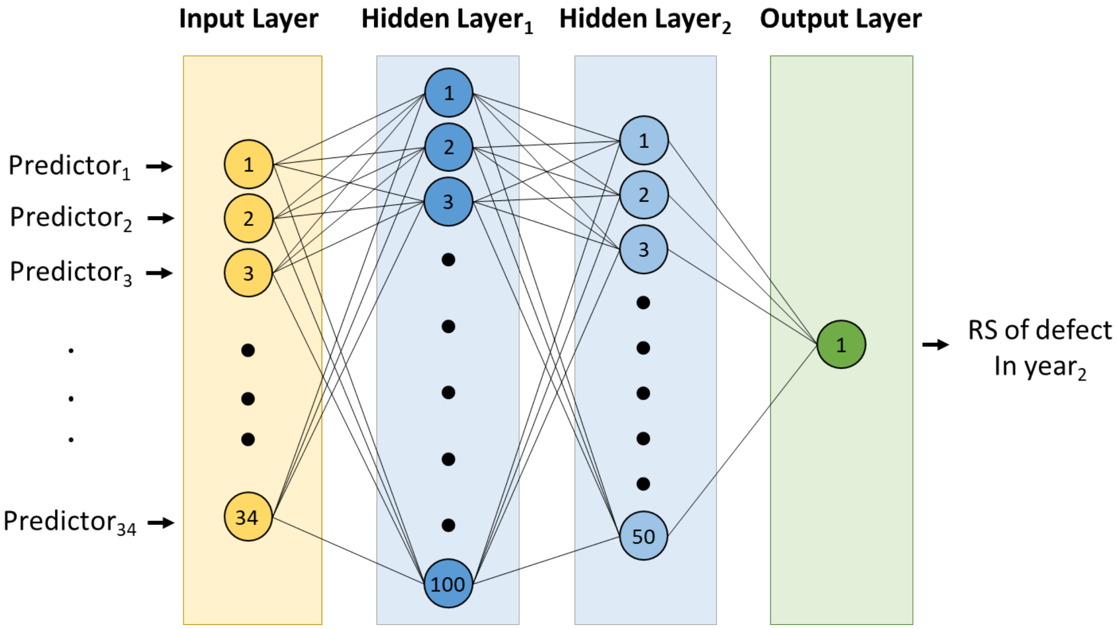


Figure 10 – ANN-MLP structure used in the study

3.5.2.3 Decision Tree Regression

Decision Tree Regression (DTR) is another algorithm that is considered in this study. Due to its intelligibility and simplicity, DTR is among the most popular ML techniques (Tso et al., 2007). In this method, a decision-tree is established with a series of simple rules utilized to split the input dataset into two parts at each node of the tree, as shown in Figure 11. By repetitive process of splitting the data, the desired outcome can be predicted at the final layer of the tree (Tso et al., 2007). The DTR was used for developing a prediction model with the maximum depth of each branch of the tree being set as eight.

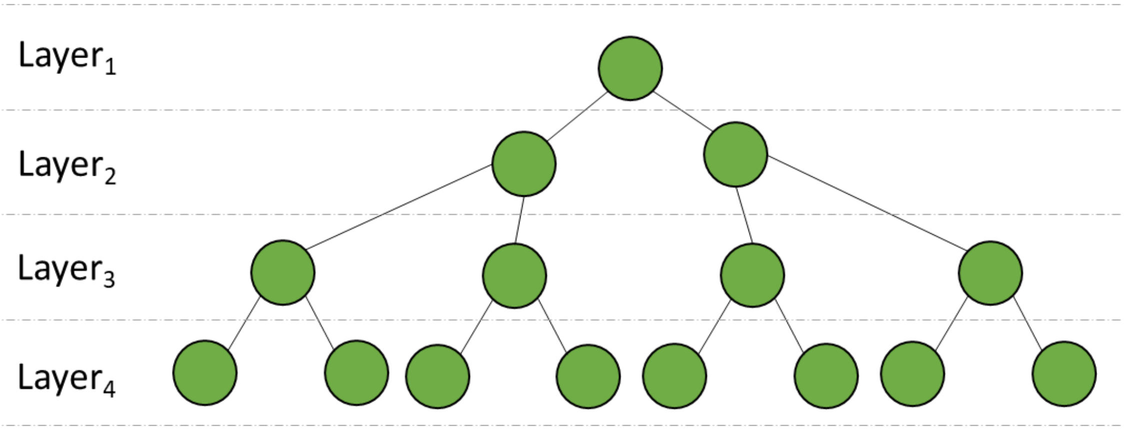


Figure 11 – A decision tree with four layers of depth

3.5.2.4 Adaptive Boosting

The idea of combining a set of weak regressors for building a high-performing model was introduced as ensemble learning. In this technique, more than one regressor is trained, each of which contributes to the final result (Schapire, 2003). In addition, the boosting technique is used to decrease the error of the combination of the constituent models. In this study, Adaptive Boosting (ADB) which is a famous boosting ensemble method was used. This algorithm decreases training errors during the process of learning from the mistakes of sequentially trained constituent models (Karabulut et al., 2014). A decision tree with a maximum depth of five as the base constituent model was utilized in ADB to build a high-performing prediction model using 100 decision trees.

3.5.2.5 Random Forest Regression

Finally, another ML method named Random Forest Regression (RFR) was used to predict future risk scores. The excellent performance of this technique made it a widely used method in developing prediction models (Yaseen et al., 2019). The RFR works based on constructing several decision trees using the bootstrap resampling method. The outcome of the produced decision trees provides the final result by either a voting or averaging approach. Figure 12 shows the general structure of the RFR technique. High stability of the procedure used in RFR has resulted in better performance and prediction accuracy while avoiding overfitting compared to other ML methods such as Artificial Neural Network (ANN) (Breiman, 2001; W.-c. Wang et al., 2015; Yaseen et al., 2019).

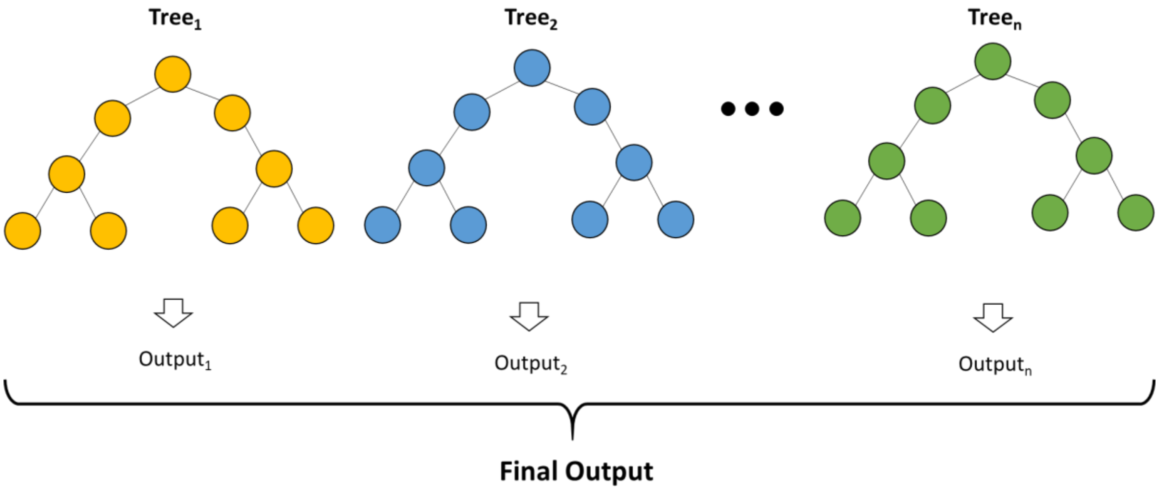


Figure 12 – General structure of a Random Forest Regression (RFR) model

3.6 Model Validation

Two common problematic cases might happen when a machine learning model is developed: underfitting and overfitting. These problems describe the degree to which the model corresponds to a particular set of data (i.e., training set) and the level it might fail to fit unseen data (i.e. the data that is not used in the training process of the model) (Friedman et al., 2001; Hawkins, 2004; Srivastava et al., 2014). Underfitting refers to a situation that the model is incapable to provide acceptable accuracy in predictions within both training and testing datasets. In that case, the model has not learned enough from the data, besides training and testing errors are high and the model outcomes are unreliable. In contrast, overfitting means that the model corresponds too closely to the training dataset but is unable to accurately make predictions for unseen data. Hence, model errors in the training set are low however the model provides predictions with high errors in the scope of the testing set.

To detect underfitting and overfitting problems of a machine learning model, several metrics are addressed in the literature and are commonly used in research studies. For a regression problem which is the case in this study, quantitative metrics that represent the level of errors are utilized to detect underfitting. To this end, the R-squared value between observations and predictions (*R^2*) and Root Mean Squared Error (*RMSE*) are two famous parameters that are vastly used in regression analysis (Lambourne et al., 2010; Thanassoulis, 1993). The greater *R^2* and the less *RMSE* show more accurate predictions. Therefore, these metrics were used to qualify developed prediction models in terms of the underfitting problem.

To find out if a model is overfitted to the training set, k-fold cross-validation is a common method proposed in the literature (Browne, 2000). K-fold cross-validation is a process to assess and validate the performance of the models on unseen data. In this procedure, the dataset is divided into two parts: training and testing sets. Then, after building the model based on the training set, its performance is calculated using the test set. The procedure is performed *k* times and the average score of them is used as the cross-validation score. Figure 13 illustrates the k-fold cross-validation process. This method was used to evaluate developed machine learning models in terms of overfitting and five was considered as the number of folds (*k*) in this study.

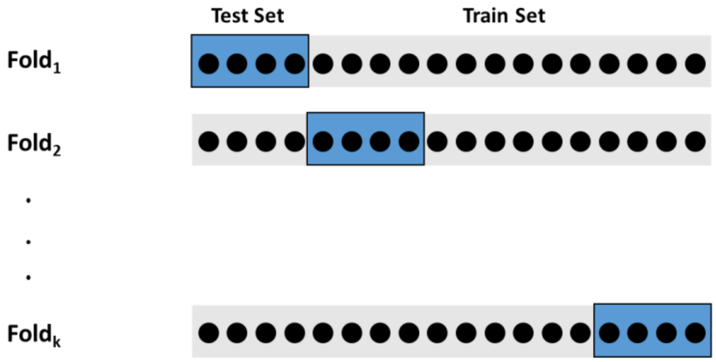


Figure 13 – K-fold cross-validation process

3.7 Model Selection and Implementation

3.7.1 Comparative Study and Model Selection

After developing prediction models based on the eight considered ML algorithms, a comparative study is performed to select the algorithm that provides the best fit to the dataset. To do so, first, the metrics that present the accuracy of predictions in training and testing sets utilizing all considered algorithms are compared. To better compare the models, the bias and variance of their predictions are also taken into consideration. Bias refers to the difference between prediction and actual observation values. Hence, the bias identifies how far off the model predictions are from the correct values. Figure 14 schematically illustrates the meaning of low and high bias for the prediction values of a model. In addition to bias, the variance of the prediction values is important when a model is developed. As it is shown in Figure 14, variance denotes how much the distance between predictions and actual values varies. A low-bias low-variance model is interpreted as a model that provides not only close predictions to the actual values but also a consistent level of accuracy in all prediction values (Suen et al., 2005). In other words, the model is not only accurate but also precise (Berardi et al., 2003). With respect to this, all developed prediction models are compared, and the best model is selected.

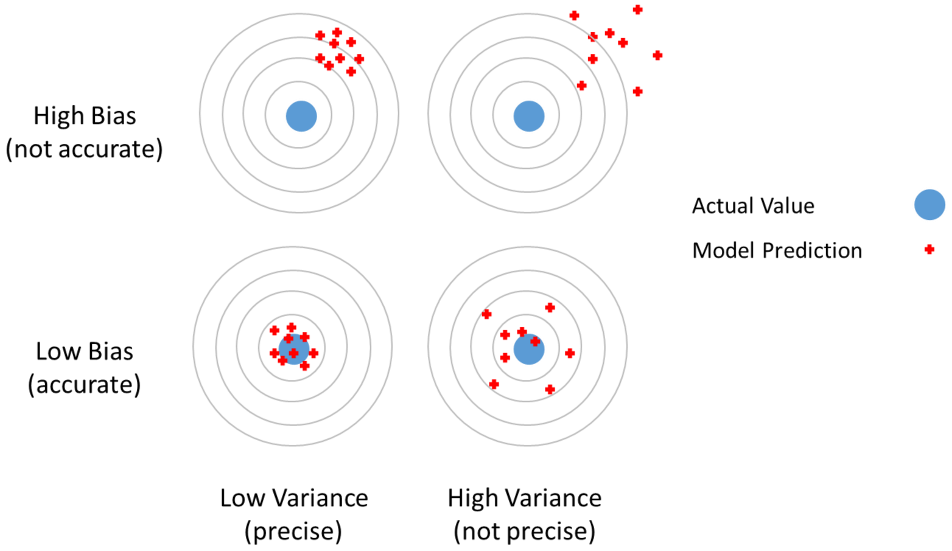


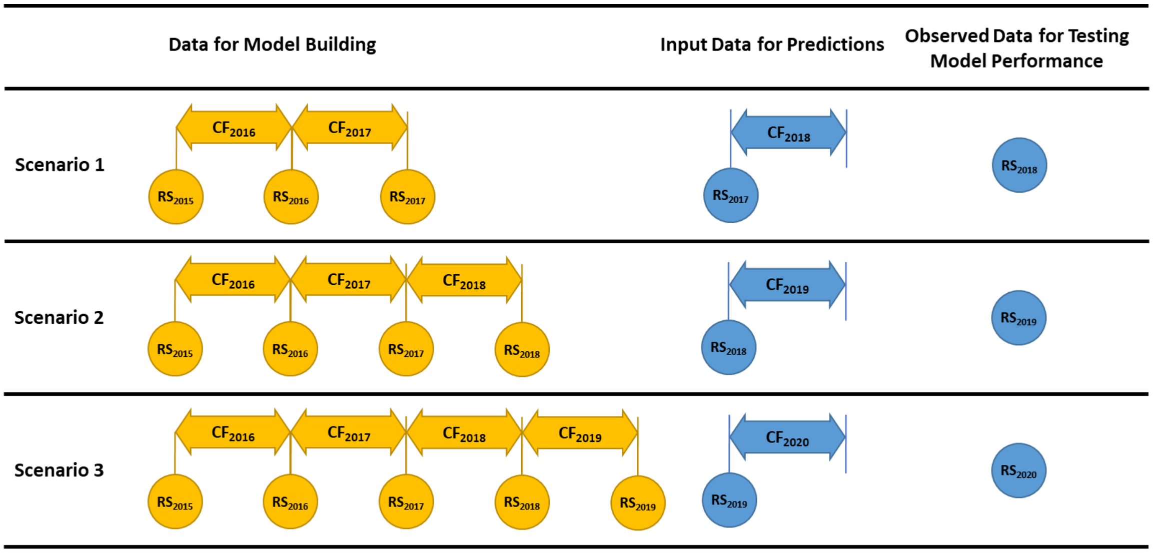
Figure 14 - Bias and variance of a prediction model (Ferrante et al., 2009)

3.7.2 Model Implementation

After developing and validating prediction models and selecting the best option, the next step is dedicated to utilizing the prediction capabilities that the model offers in the decision making process. To this end, the outcomes of the model will be used in preparing risk maps and finding the hotspot of defects over the considered roadways in the next year. Therefore, in this section, the performance of the models in their implementation will be investigated.

To do so, three scenarios were taken into account as shown in Figure 15. This figure shows that the size of the training set was incrementally increased to measure the impact of data availability in developing models. In the first scenario, the data of RSs at the end of FY2015, FY2016, and FY2017 as well as all contributing factors in FY2016 and FY2017 are utilized to build the prediction model. It is worth mentioning that the data is first split into training and testing sets and all validation procedures and comparative study are performed in identifying the best model. Then, the model is used to predict RSs in FY2018 when RSs at the end of FY2017 and contributing factors in FY2018 are the inputs of the model. Later on, R2 and RMSE are used to assess the accuracy of predictions considering the actual observations of RSs at the end of FY2018. This process is repeated in scenario2 by adding the data of FY2018 in building the prediction model and then testing the model performance with FY2019. Finally, in scenario3 another year of data is added to build the model, and predictions are evaluated for FY2020.

The process of adding a new year of data for model development and testing its performance on unseen data in the next year helps to monitor the model performance and the trend of its prediction capability to capture all possible variations of contributing factors in the case study region and being trained thoroughly.



Note: CF_{2016} = Contributing Factors in FY2016 (Weather, Traffic, Maintenance)
RS_{2016} = Observed Risk Scores at the End of FY2016

Figure 15 - Scenarios utilized to assess the selected model performance in its implementation

3.8 Web-based Spatial Visualization

A Geographical Information System (GIS) is a framework that enables practitioners and researchers to gather, manage, and analyze data. GIS integrates various types of data and provides spatial data analysis tools accompanied with visualization capabilities by organizing layers of information into usable maps (Bolstad, 2016). Given these capabilities, GIS reveals deeper insights into data and helps to make maps that highlight patterns, relationships, and situations which ultimately results in smart decisions in different fields.

ESRI's ArcGIS is the most commonly used GIS application for working with geographical data (i.e. the data containing the location attributes). Therefore, it was utilized in this study not only for spatial analysis like KDE and ordinary kriging but also for visualization purposes. In ArcGIS, geographical data are stored as shapefiles that contain the geometric location and attribute information of each feature. In addition, features are represented by points, lines, or polygons that are connected to an attributed table including all corresponding information.

In order to visualize the results of RSs and other attributes of the considered asset items, the data including required features and the coordinates of asset items were imported into the ArcGIS and converted to shapefiles. Therefore, in this study, ArcGIS was used as a data repository to store all attributes of the considered asset items. Then, ArcGIS visualization tools were deployed to come up with illustrative maps. Furthermore, since the web-based version of ArcGIS (i.e. ArcGIS-Online) facilitates cloud-based data sharing and also provides compelling data visualization tools for building maps, it was used as a platform for interactive presentations.

CHAPTER 4: RESULTS AND DISCUSSION

This section provides the results that were accomplished by applying the proposed framework to the selected case study. In this chapter, the results of calculating risk scores utilizing KDE are presented first. Then, in the section prepared for scaling data, the variation of the scaled feature space is illustrated. Next, correlations in the feature space are visualized and the process of removing multicollinearity in the extent of the case study data is explained. Later on, the results of the developed prediction models are presented and compared. Afterward, the implementation of the model is discussed, and the results are highlighted. Finally, spatial visualization in ArcGIS is presented.

4.1 Density Estimation of Defects

According to the proposed methodology, KDE was utilized to come up with a parameter (RS) that represents the density of defects in each segment of the considered roadways. Figure 16 provides the histogram of erosion RSs on paved ditches and the corresponding KDE results at the end of FY2015, FY2016, FY2017, FY2018, FY2019, and FY2020. Additionally, Figure 17 and Figure 18 provide the histogram and spatial distribution of RSs for obstruction and cracking, respectively. Similarly, the RSs of the considered defects were calculated on the selected nearby asset items (mentioned in Table 12) at the end of FY2015 to FY2020 to be used in the prediction.

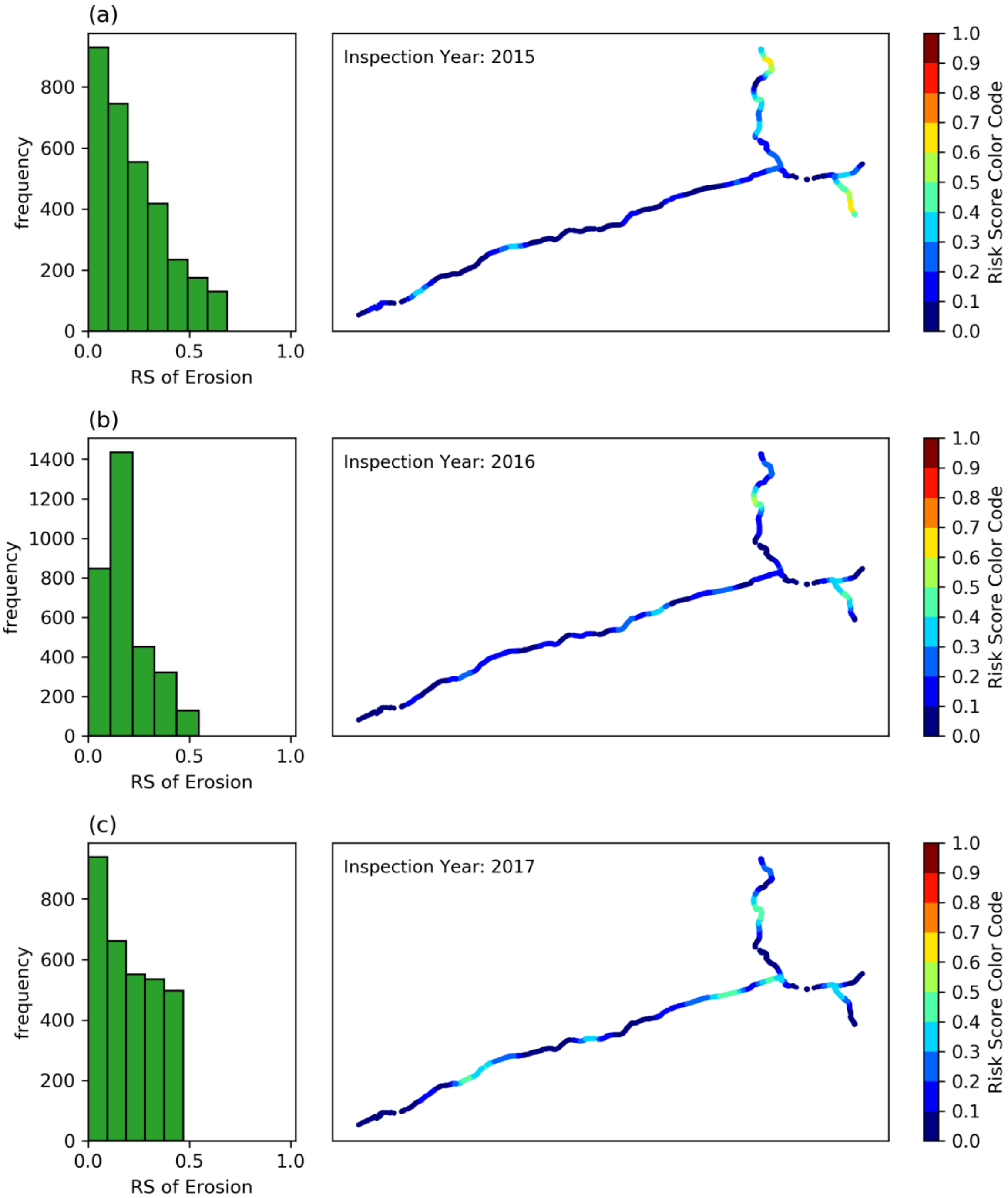


Figure 16 - Histograms and spatial distributions of erosion RSs in different years of inspection (a) FY2015 (b) FY2016 (c) FY2017 (d) FY2018 (e) FY2019 (f) FY2020

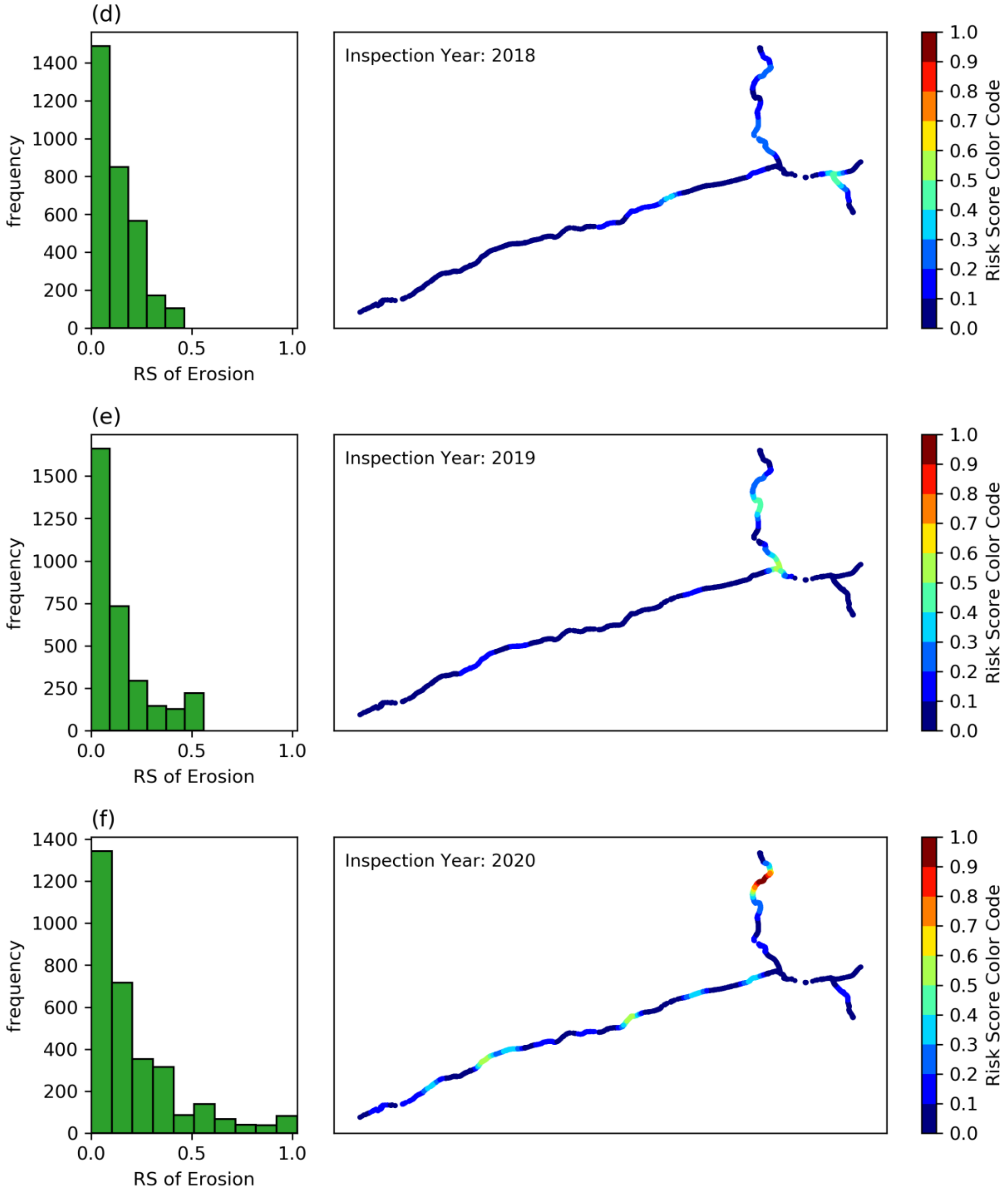


Figure 16 (continued) - Histograms and spatial distributions of erosion RSs in different years of inspection (a) FY2015 (b) FY2016 (c) FY2017 (d) FY2018 (e) FY2019 (f) FY2020

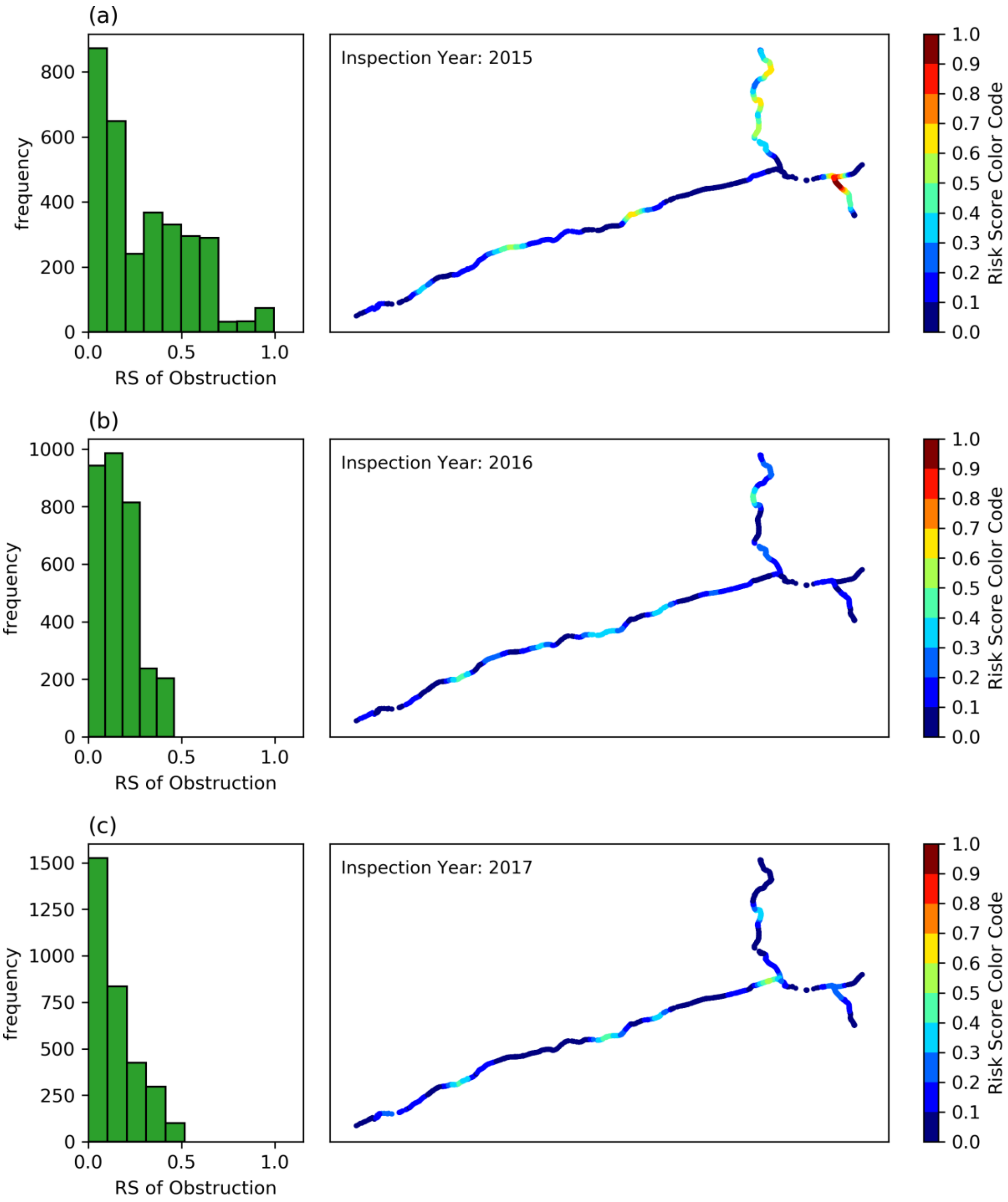


Figure 17 - Histograms and spatial distributions of obstruction RSs in different years of inspection (a) FY2015 (b) FY2016 (c) FY2017 (d) FY2018 (e) FY2019 (f) FY2020

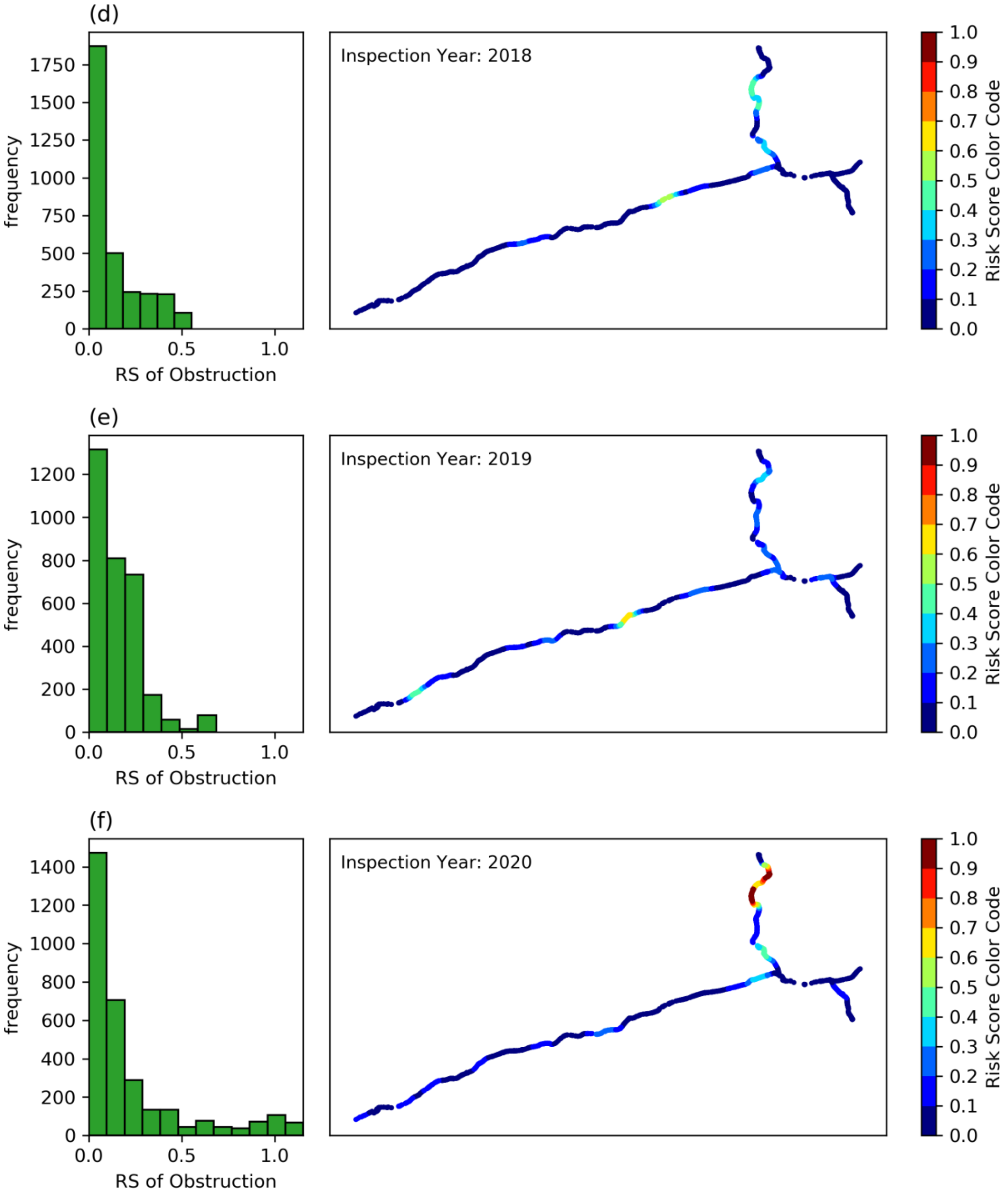


Figure 17 (continued) - Histograms and spatial distributions of obstruction RSs in different years of inspection (a) FY2015 (b) FY2016 (c) FY2017 (d) FY2018 (e) FY2019 (f) FY2020

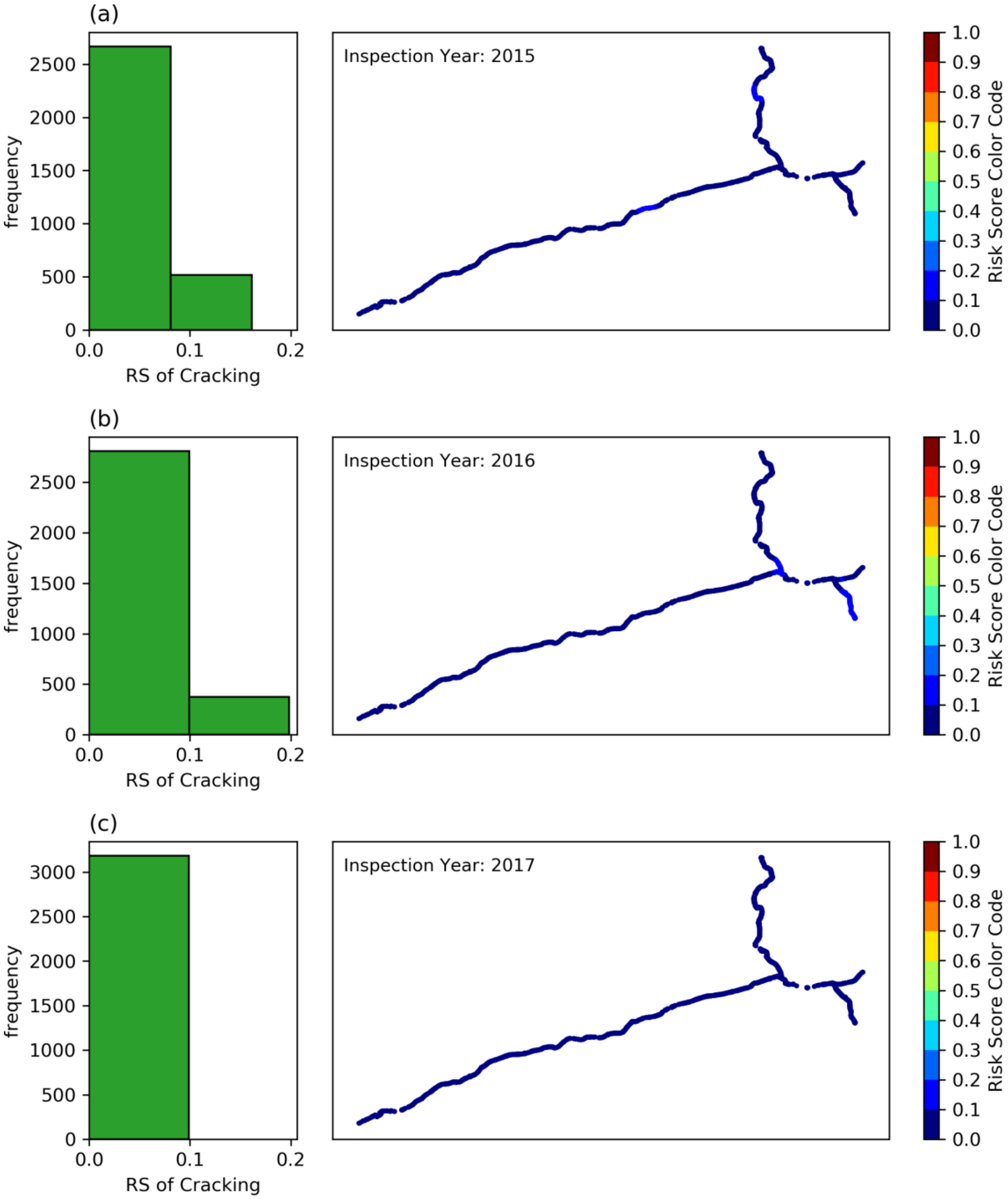


Figure 18 - Histograms and spatial distributions of cracking RSs in different years of inspection (a) FY2015 (b) FY2016 (c) FY2017 (d) FY2018 (e) FY2019 (f) FY2020

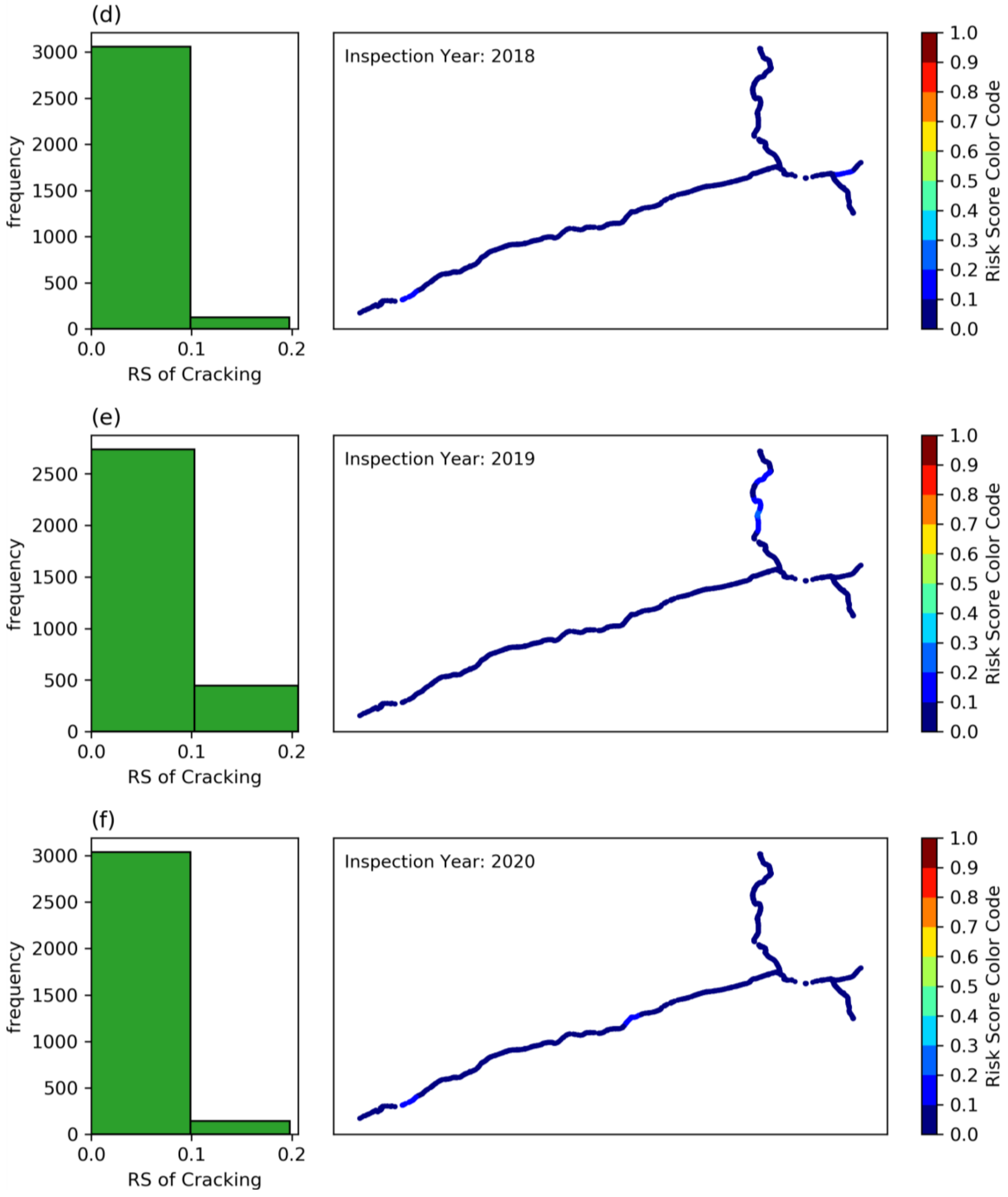


Figure 18 (continued) - Histograms and spatial distributions of cracking RSs in different years of inspection (a) FY2015 (b) FY2016 (c) FY2017 (d) FY2018 (e) FY2019 (f) FY2020

4.2 Scaling Data

The considered contributing factors to the degradation of roadway assets in this study have different ranges and measurement units. Figure 19 provides a series of boxplots that visualize the different variations among continuous features. Later on, the min-max scaler was used to map all features to a range between 0 to 1 to prevent potential future biases of outcomes. Figure 20 shows the boxplots of the scaled features considered in this study.

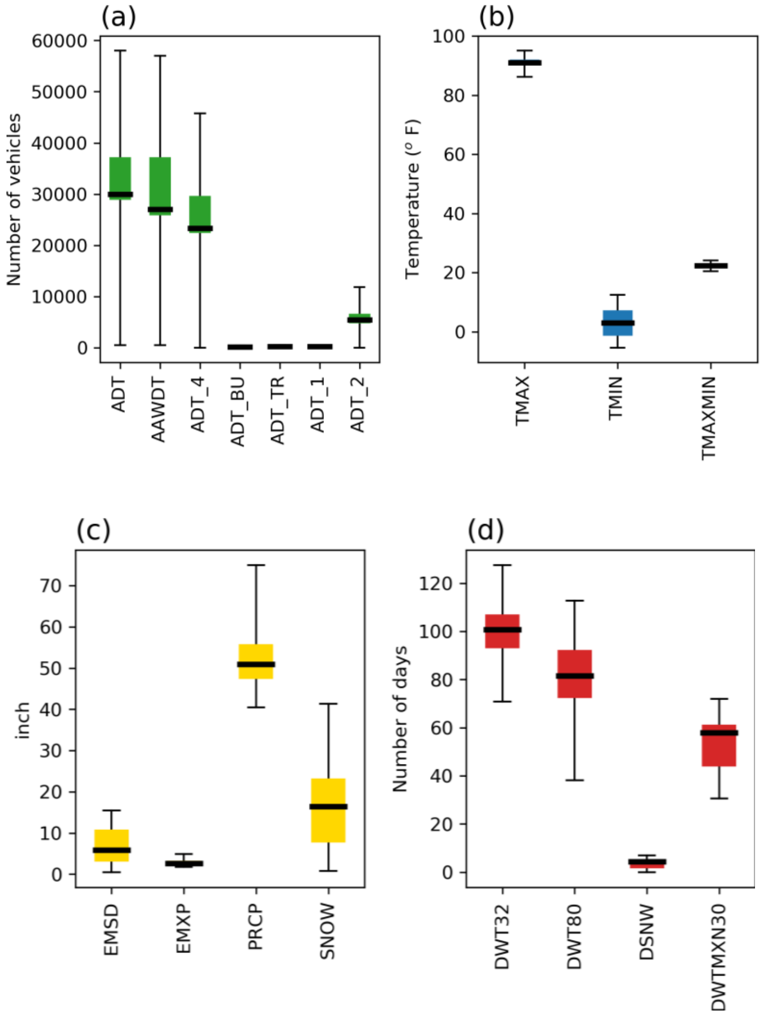
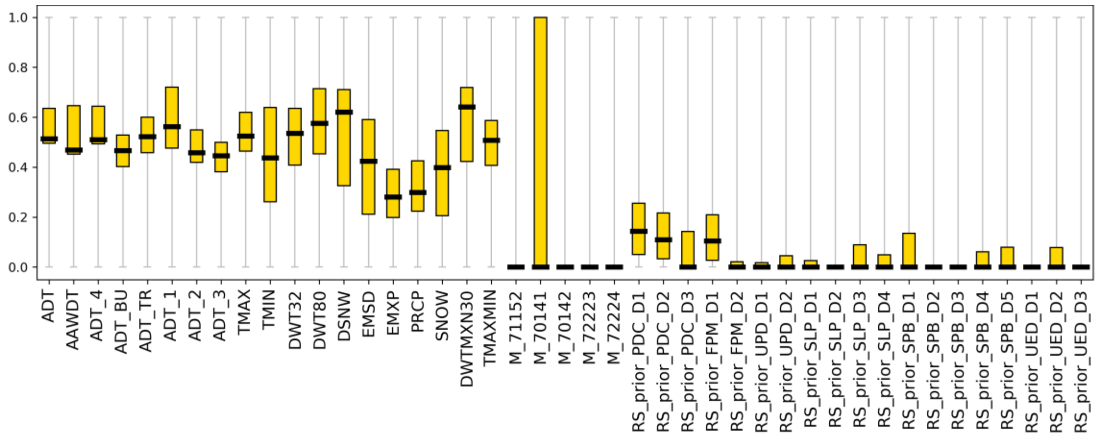


Figure 19 - Boxplots of continuous features (a) traffic features (b) temperature features (c) precipitation features (d) weather features measured with days

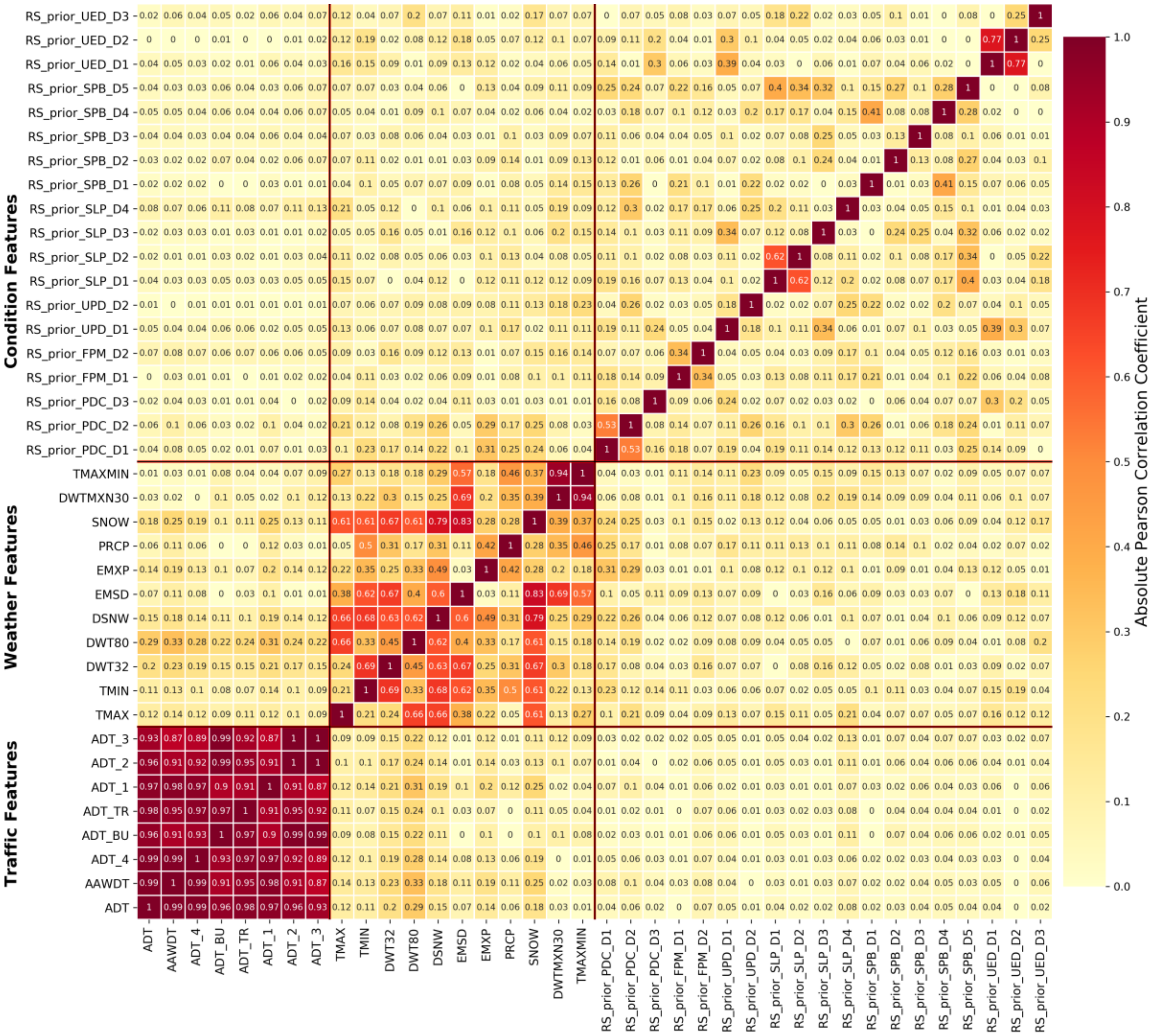


NOTE : RS_prior: Risk score in prior year | PDC: Paved ditch | FPM: Flexible pavement | UPD: Unpaved ditch | SLP: Slope
SPB: Small pipes & box culverts | UED: Under drains and edge drains
D1: Defect #1 of the corresponding asset (e.g. RS_prior_SLP_D1: Risk score of prior year occurence of erosion on slopes)

Figure 20 - Boxplots of scaled features

4.3 Removing Multicollinearity

To detect and remove multicollinearity, the correlations inside the input feature space were investigated. Figure 21 provides the pairwise absolute Pearson Correlation between continuous features. This figure shows that traffic attributes (*ADT, AAWDT, ADT_4, ADT_BU, ADT_TR, ADT_1, ADT_2, and ADT_3*) are highly correlated (i.e., their pairwise absolute Pearson correlation is greater than 0.9). Therefore, the continuous feature space was reduced by keeping *ADT* as the sole representative of traffic features. Furthermore, *TMAXMIN* and *DWTMXN30* are also highly correlated, as shown in Figure 21. Hence, only *TMAXMIN* was kept and *DWTMXN30* was removed from the feature space.



NOTE : RS_prior: Risk score in prior year | PDC: Paved ditch | FPM: Flexible pavement | UPD: Unpaved ditch | SLP: Slope
SPB: Small pipes & box culverts | UED: Under drains and edge drains
D1: Defect #1 of the corresponding asset (e.g. RS_prior_SLP_D1: Risk score of prior year occurence of erosion on slopes)

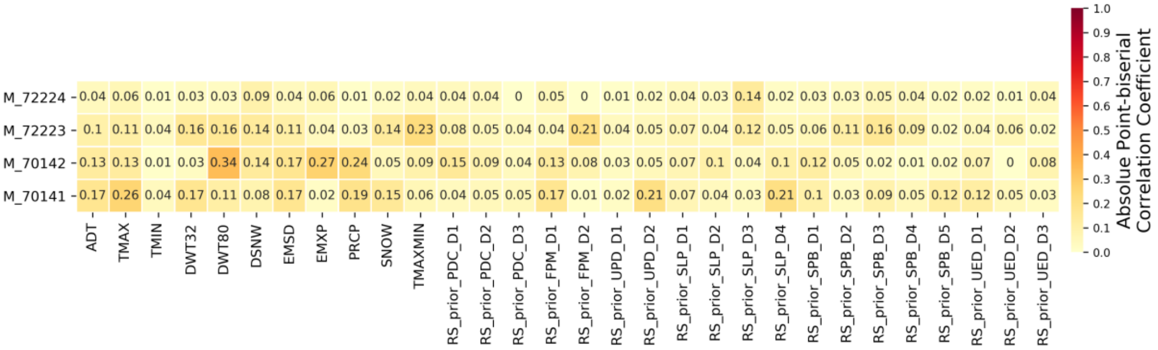
Figure 21 - Absolute Pearson correlation matrix of continuous features

Table 13 provides the chi-square test results, or in other words, the dependencies among categorical features. The results show that M_71152 and M_72224 are highly correlated (i.e., the corresponding p-value is greater than 0.05). Therefore, only M_72224, M_70141, M_70142, and M_72223 were kept and M-71152 was removed for future analysis.

Table 13 - Pairwise Chi-square correlation test results for categorical features

	M_71152	**M_70141**	**M_70142**	**M_72223**	**M_72224**
M_71152	N/A	9.08x10^{-219}	6.60x10^{-147}	1.04x10^{-05}	5.22x10^{-02}
M_70141	9.08x10^{-219}	N/A	0.00	9.14x10^{-260}	8.26x10^{-25}
M_70142	6.60x10^{-147}	0.00	N/A	7.22x10^{-159}	1.99x10^{-42}
M_72223	1.04x10^{-05}	9.14x10^{-260}	7.22x10^{-159}	N/A	0.00
M_72224	5.22x10^{-02}	8.26x10^{-25}	1.99x10^{-42}	0.00	N/A

Finally, Figure 22 presents the absolute point-biserial correlation coefficients between remaining categorical and continuous features. According to the results, none of the features are highly correlated and all features can be considered independent.



	ADT	TMAX	TMIN	DWT32	DWT80	DSNW	EMSD	EMXP	PRCP	SNOW	TMAXMIN	RS_prior_PDC_D1	RS_prior_PDC_D2	RS_prior_PDC_D3	RS_prior_FPM_D1
M_72224	0.04	0.06	0.01	0.03	0.03	0.09	0.04	0.06	0.01	0.02	0.04	0.04	0.04	0	0.05
M_72223	0.1	0.11	0.04	0.16	0.16	0.14	0.11	0.04	0.03	0.14	0.23	0.08	0.05	0.04	0.04
M_70142	0.13	0.13	0.01	0.03	0.34	0.14	0.17	0.27	0.24	0.05	0.09	0.15	0.09	0.04	0.13
M_70141	0.17	0.26	0.04	0.17	0.11	0.08	0.17	0.02	0.19	0.15	0.06	0.04	0.05	0.05	0.17

	RS_prior_FPM_D2	RS_prior_UPD_D1	RS_prior_UPD_D2	RS_prior_SLP_D1	RS_prior_SLP_D2	RS_prior_SLP_D3	RS_prior_SLP_D4	RS_prior_SPB_D1	RS_prior_SPB_D2	RS_prior_SPB_D3	RS_prior_SPB_D4	RS_prior_SPB_D5	RS_prior_UED_D1	RS_prior_UED_D2	RS_prior_UED_D3
M_72224	0	0.01	0.02	0.04	0.03	0.14	0.02	0.03	0.03	0.05	0.04	0.02	0.02	0.01	0.04
M_72223	0.21	0.04	0.05	0.07	0.04	0.12	0.05	0.06	0.11	0.16	0.09	0.02	0.04	0.06	0.02
M_70142	0.08	0.03	0.05	0.07	0.1	0.04	0.1	0.12	0.05	0.02	0.01	0.02	0.07	0	0.08
M_70141	0.01	0.02	0.21	0.07	0.04	0.03	0.21	0.1	0.03	0.09	0.05	0.12	0.12	0.05	0.03

NOTE : RS_prior: Risk score in prior year | PDC: Paved ditch | FPM: Flexible pavement | UPD: Unpaved ditch | SLP: Slope
SPB: Small pipes & box culverts | UED: Under drains and edge drains
D1: Defect #1 of the corresponding asset (e.g. RS_prior_SLP_D1: Risk score of prior year occurence of erosion on slopes)

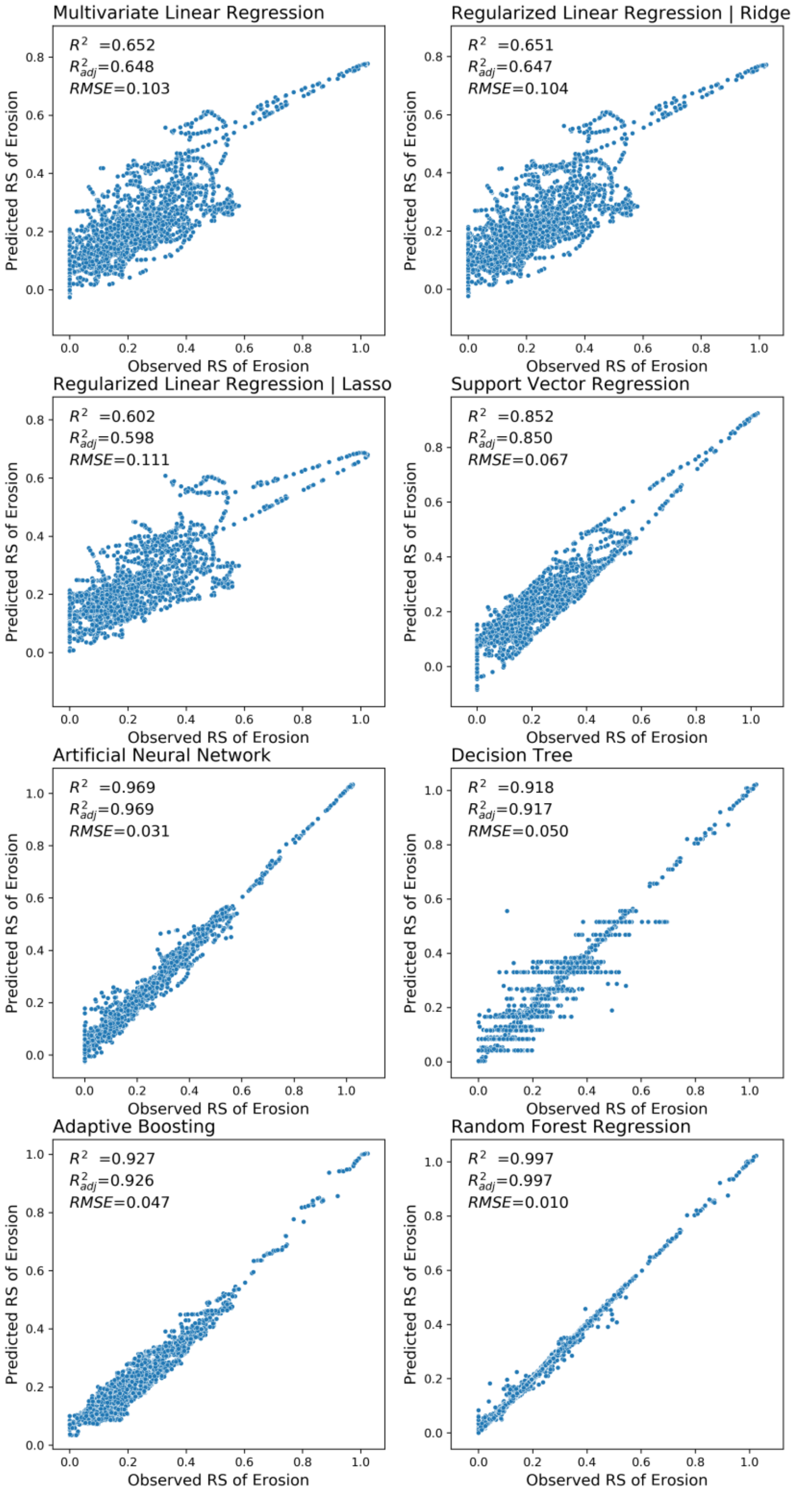
Figure 22 - Absolute point-biserial correlation matrix between continuous and categorical features

4.4 Prediction Models

After reducing feature space and removing multicollinearity, the selected ML algorithms were used to predict RSs of erosion, obstruction, and cracking on paved ditches. The results of the models for erosion, obstruction, and cracking RSs are presented in Figure 23, Figure 24, and Figure 25, respectively. In these figures, the obtained coefficient of determination (R^2), adjusted coefficient of determination (R^2_{adj}), and the Root Mean Square Error (*RMSE*) are reported. In addition, the observed vs. predicted Risk Score (RS) values in all considered algorithms are visualized. Besides, Table 14 provides scores of the developed prediction model (i.e., their R^2) using machine learning algorithms in training and testing sets for three considered defects on paved ditches. This table is utilized to investigate the underfitting problem of the developed models. As it is shown in this table, linear models (i.e., multilinear regression, Ridge, and Lasso) in all cases provided low scores both in training and testing sets. Therefore, the models were incapable of capturing the patterns and relationships between contributors and the prediction. Consequently, this result was interpreted as the nonlinearity in relationships and the need for nonlinear models to capture the relations by learning within the extent of the considered data. Moreover, higher scores of nonlinear models in Table 14 attested to the interpretation.

Table 14 - Scores of the considered models in training and testing sets

Utilized ML algorithm	**Erosion**		**Obstruction**		**Cracking**	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**
Multivariate Linear Regression	0.642	0.652	0.515	0.516	0.317	0.330
Regularized Linear Regression | Ridge	0.641	0.651	0.515	0.516	0.316	0.330
Regularized Linear Regression | Lasso	0.600	0.602	0.479	0.481	0.127	0.150
Support Vector Regression	0.845	0.852	0.871	0.872	-2.575	-2.638
Artificial Neural Network	0.968	0.969	0.982	0.982	0.919	0.911
Decision Tree	0.918	0.918	0.886	0.881	0.951	0.942
Adaptive Boosting	0.926	0.927	0.876	0.877	0.493	0.477
Random Forest Regression	0.999	0.997	0.999	0.997	0.999	0.996



Multivariate Linear Regression

Regularized Linear Regression | Ridge

Regularized Linear Regression | Lasso

Support Vector Regression

Artificial Neural Network

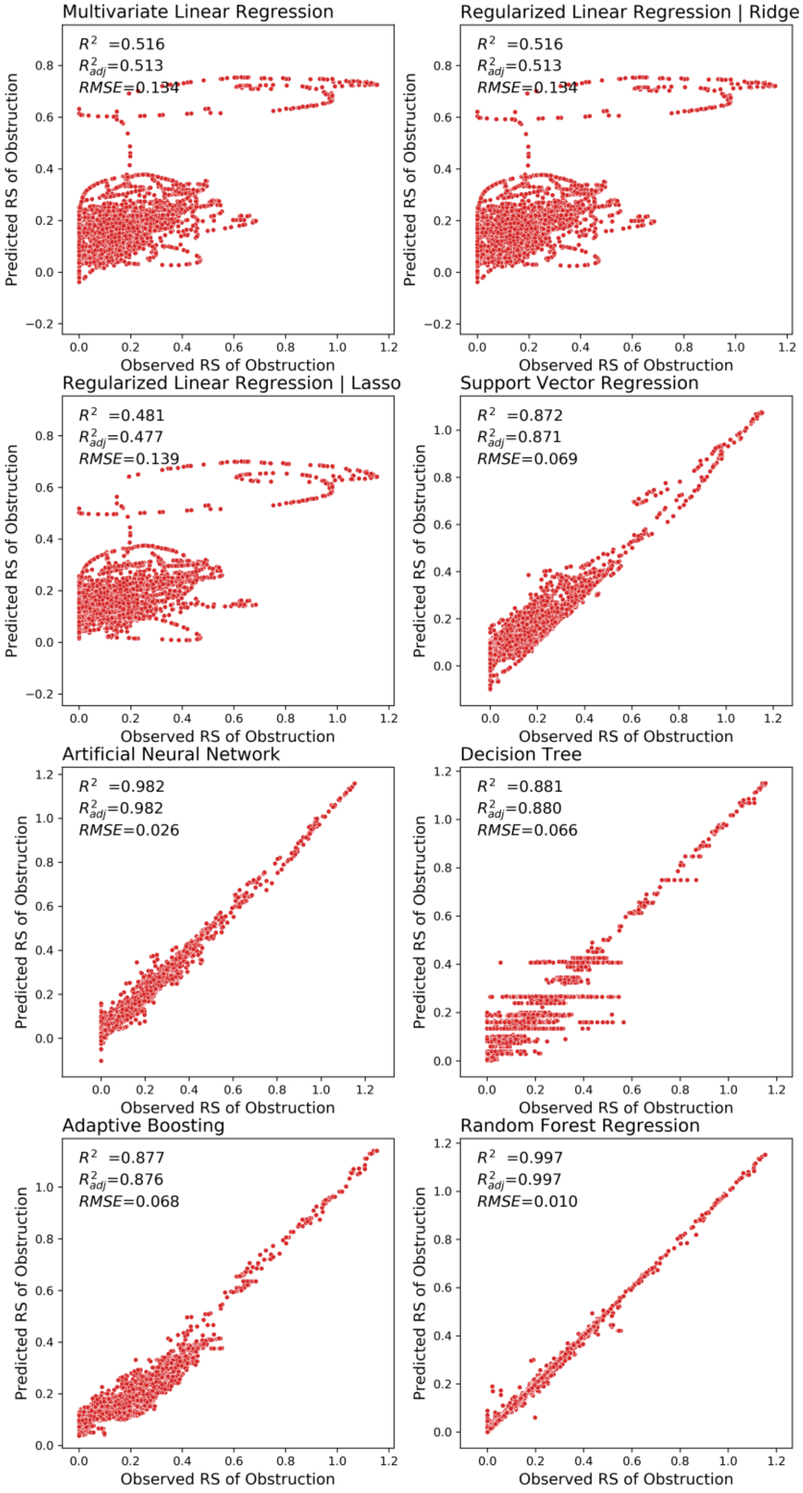
Decision Tree

Adaptive Boosting

Random Forest Regression

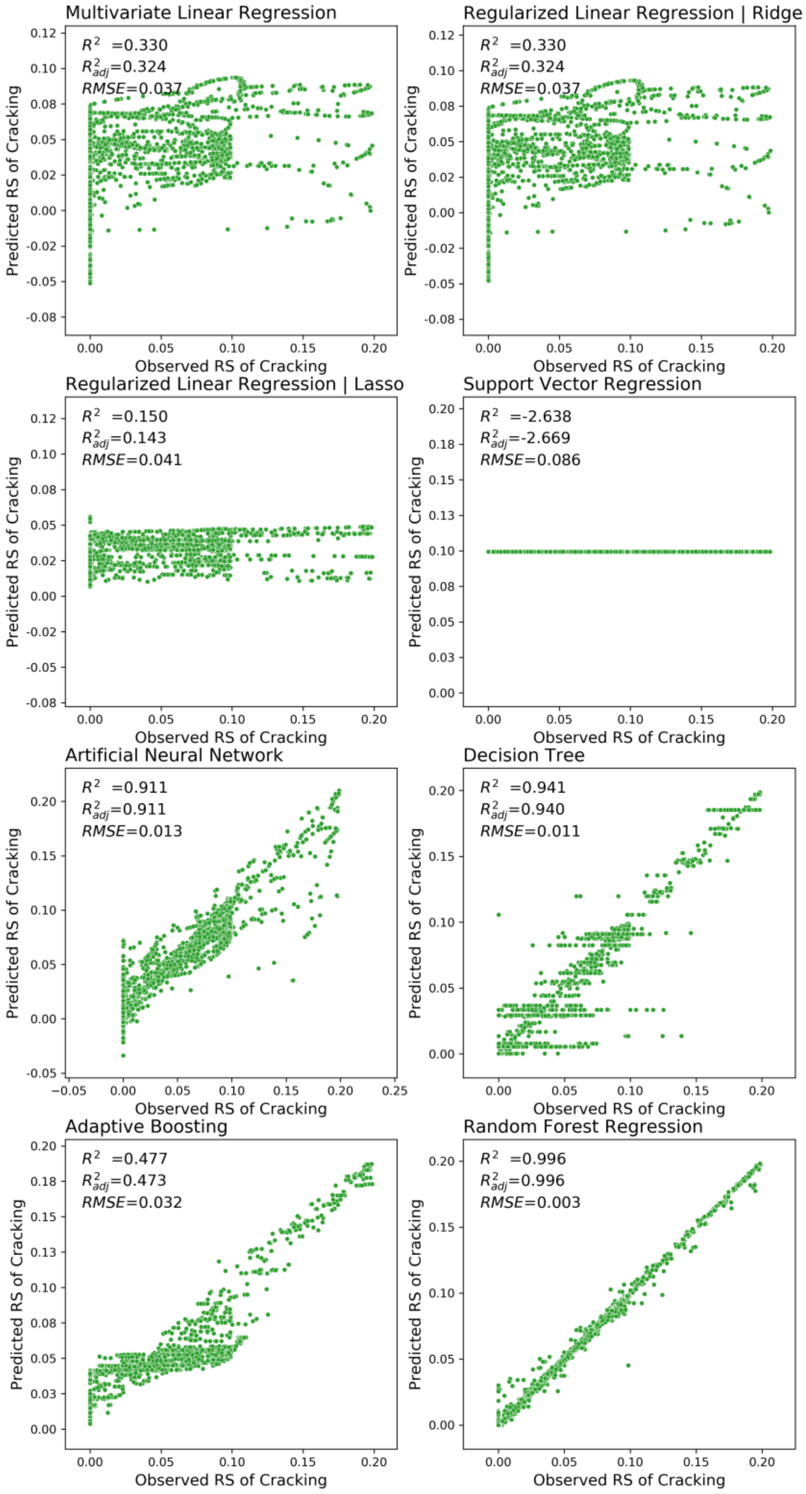
Note: R^2: Coefficient of determination | R^2_{adj}: Adjusted coefficient of determination | $RMSE$: Root Mean Square Error

Figure 23 - Observed versus predicted erosion RSs using considered algorithms on the testing set



Note: R^2: Coefficient of determination | R^2_{adj}: Adjusted coefficient of determination | *RMSE*: Root Mean Square Error

Figure 24 - Observed versus predicted obstruction RSs using considered algorithms on the testing set



Multivariate Linear Regression

Regularized Linear Regression | Ridge

Regularized Linear Regression | Lasso

Support Vector Regression

Artificial Neural Network

Decision Tree

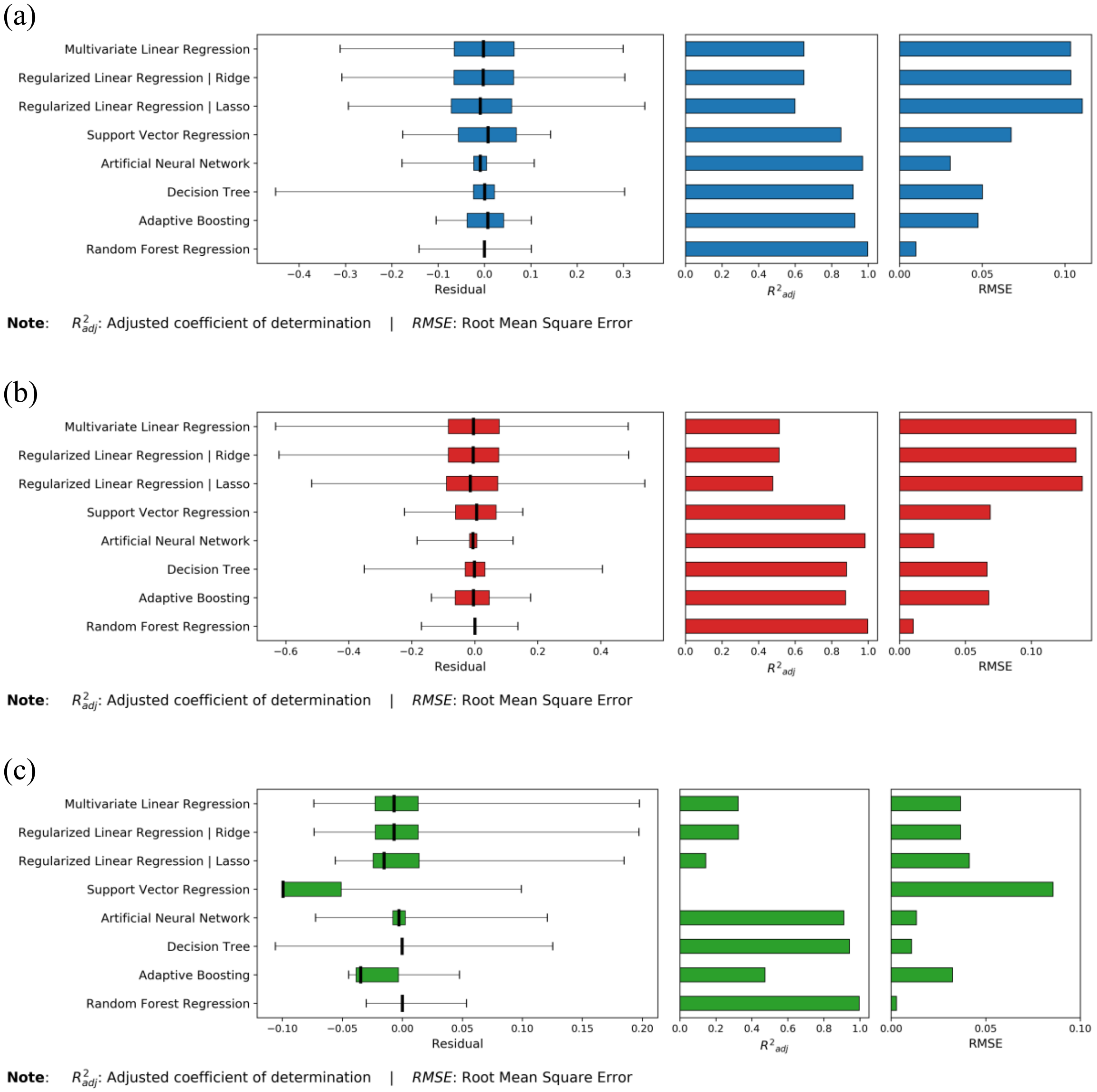
Adaptive Boosting

Random Forest Regression

Note: R^2: Coefficient of determination | R^2_{adj}: Adjusted coefficient of determination | $RMSE$: Root Mean Square Error

Figure 25 - Observed versus predicted cracking RSs using considered algorithms on the testing set

Additionally, a summary of the accuracy metrics for the corresponding prediction models is provided in Figure 26. The result of k-fold cross-validation is also reported in Table 15. In this table, for each ML algorithm, the minimum and maximum scores in the five folds of training and testing sets are provided.



Note: R^2_{adj}: Adjusted coefficient of determination | *RMSE*: Root Mean Square Error

Note: R^2_{adj}: Adjusted coefficient of determination | *RMSE*: Root Mean Square Error

Note: R^2_{adj}: Adjusted coefficient of determination | *RMSE*: Root Mean Square Error

Figure 26 - Comparison of models' accuracy metrics on testing set (a) prediction models for erosion RSs (b) prediction models for obstruction RSs (c) prediction models for cracking RSs

Table 15 - Results of cross-validation scores for all considered ML algorithms

Algorithm	**Erosion**		**Obstruction**		**Cracking**	
	Min Score	**Max Score**	**Min Score**	**Max Score**	**Min Score**	**Max Score**
Multivariate Linear Regression	0.614	0.672	0.480	0.546	0.285	0.350
Regularized Linear Regression | Ridge	0.615	0.67	0.481	0.546	0.286	0.349
Regularized Linear Regression | Lasso	0.583	0.623	0.453	0.502	0.103	0.162
Support Vector Regression	0.834	0.865	0.873	0.877	-2.997	-2.224
Artificial Neural Network	0.973	0.984	0.984	0.990	0.914	0.937
Decision Tree	0.893	0.927	0.829	0.891	0.933	0.963
Adaptive Boosting	0.919	0.939	0.872	0.910	0.390	0.624
Random Forest Regression	0.996	0.999	0.998	0.999	0.997	0.999

The results unveiled that in all cases, the RFR provided more accurate outcomes with respect to R^2 and *RMSE* values. Additionally, Figures 27 to 28 reveal that the predicted values using RFR are very close to the observed values in the considered dataset. Moreover, the RFR cross-validation scores show the narrow range of scores in all five folds of validation, which can be interpreted as the lack of overfitting, and further confirms the significant performance of RFR on unseen data. Therefore, given the fact that all measurements pointed out the RFR as the best model for the considered case study, the RFR was selected for further analyzing and discussing the results.

One of the best attributes of RFR is its capability in calculating the contribution (i.e., importance) of each feature in the regression by providing a metric called importance score. To measure the importance of each contributing factor, most methods rely on the decrease in the accuracy when a permutation to a specific feature is performed. In this approach, when a feature is permuted, its original relationship within the decision trees (shown in Figure 12) with the final output is disturbed. Therefore, using the permuted feature along with the other non-permuted features might result in a decrease in the accuracy of predictions. This descent in accuracy is believed to be a realistic way of finding the importance of each feature. The more accuracy decrease can be interpreted as more contribution of that feature into regression (Strobl et al., 2007). This metric was used here to measure the importance of the considered features in the regression analysis. With utilizing RFR, the main goal is to let the model decide the most significant contributors among the wide range of potential candidates that were included in the framework, instead of subjectively selecting them beforehand. In this way, the important contributors might vary from asset to asset, which shows that the proposed framework respects the difference among the different highway assets' nature.

The importance of each considered contributing factor was investigated to interpret their contribution to the regression with the aforementioned attribute of RFR. Figure 27 provides the obtained results in erosion, obstruction, and cracking predictions. This figure shows that paved ditch erosion RS in the prior year, the erosion of neighboring unpaved ditches, and maximum annual daily temperature (*TMAX*) contributed the most to the predicted erosion RS. In addition, the importance of the annual average of daily max-min temperature difference (*TMAXMIN*) and the number of freezing days (*DWT32*) are considerable. The results confirm the interrelations between nearby assets and their importance on one another's conditions. More importantly, the outcomes also highlight the higher contribution of short-term precipitation factors (e.g. *EMXP: the maximum daily precipitation*, and *EMSD: the maximum annual daily snow)* in the erosion of paved ditches in comparison to the long-term average annual precipitation (i.e. *PRCP* and *SNOW*). Finally, the results underline the slight contribution of two of the maintenance works (*M_72223*: *Concrete Patching / Repair* and *M_70142: Machine cleaning*) in erosion RS of paved ditches out of the considered maintenance tasks.

Similarly, the importance of contributing factors in the prediction model of obstruction RSs on paved ditches was investigated. According to Figure 27, like erosion, the maximum annual daily temperature (*TMAX*) has a bold contribution to the values of obstruction RSs calculated by the model. Also, the contribution of the number of freezing days (*DWT32*), the annual average of daily max-min temperature difference (*TMAXMIN*), and total annual snow depth were significant. Furthermore, in this case, long-term precipitation features (i.e. *PRCP* and *SNOW*) had more contribution in predicting obstruction RSs compared to the short-term precipitation features (*EMSD* and *EMXP*). In addition, the contribution of the prior year drain outlet defect (*RS_prior_UED_D1*) in the vicinity of paved ditches was bold. The reason for this contribution could be attributed to the downstream blockage resulted from defected under drains and edge drains outlet and settlement of debris and obstruction in the upstream ditches. The figure also highlights the contribution of the condition of other neighboring assets, such as erosion on unpaved ditches and lower-slope issue on slopes on calculating paved ditch obstruction RSs.

Figure 27(c) reveals that *TMAX* and *TMAXMIN* that represent the temperature features and correspond to temperature harshness in a region contributed significantly in predicted cracking RSs. Besides, the next rank belongs to *EMXP* which is a short-term precipitation feature. This figure also unveils the importance of prior year cracking, erosion, and obstruction RSs in the next year cracking RSs on paved ditches. Ultimately, the results show that the condition of nearby assets contributed to the predicted RSs as well. It is worth mentioning that the contribution of traffic (*ADT*) in all three considered defects RSs is low, which seems rational.

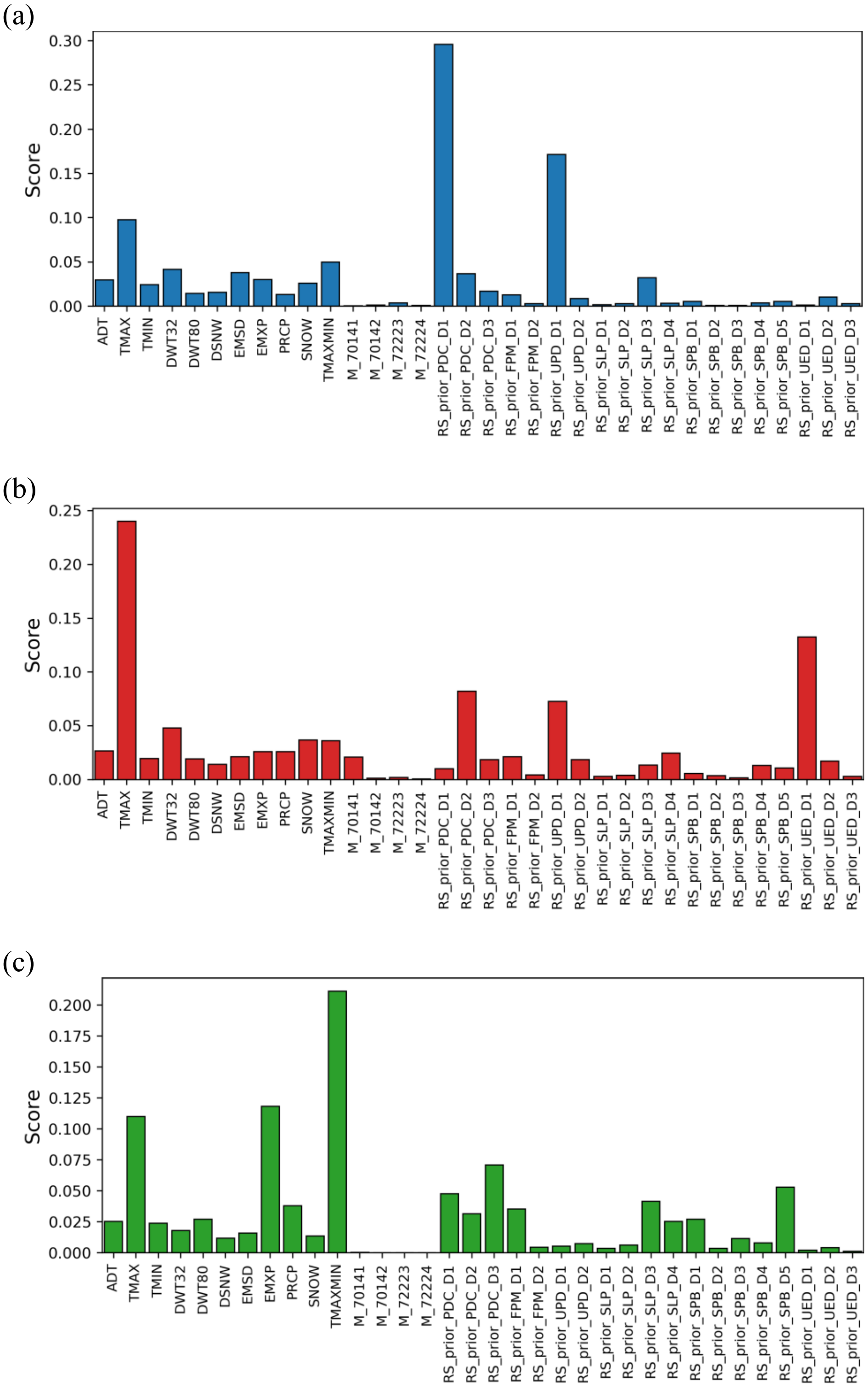


Figure 27 - Importance feature scores in the RFR model for predicting RSs of defects (a) erosion (b) obstruction (c) cracking

4.5 Model Implementation

With respect to the scenarios introduced in Figure 15, the performance of the prediction models in forecasting RSs of erosion, obstruction, and cracking on paved ditches was investigated and the results are provided in this section.

4.5.1 Erosion Prediction

Figure 28 illustrates R^2 and *RMSE-scaled* values for prediction models developed for RSs of erosion on paved ditches based on the scenarios explained in the methodology (section 4.6.2). Since the range of RS values in different years are not equal, to come up with a comparable metric for residuals, *RMSE* values were scaled based on the range of RSs in the corresponding year. Hence, *RMSE-scaled* in each year was calculated by dividing *RMSE* value by the range of RSs (i.e. maximum-minimum difference). As it is shown, the trend of R^2 values is increasing while that of *RSME* is decreasing. It shows that adding a new year of data to the training set of scenario1 improved the accuracy of the predictions in scenario2 and this trend continues for scenario3 that reached the value of 0.65 for R^2. The improvement looks more significant given that adding only two years of data in the training process improved R^2 from -1.1 to 0.65.

Figure 29 displays the spatial distribution of erosion RSs that provides a comparison between observed and predicted RSs at the end of FY2020. As shown in this figure, the locations of higher RSs are almost the same in two cases. However, to better compare observation and prediction values, Figure 30 illustrates the longitudinal profile of erosion RSs on case study roadways. In this figure, the match between the majority of observed and predicted RSs peaks is obvious. To quantify the match, Figure 31 shows the percentage of points that are similar when a threshold is assumed for defining hotspot and coldspot of defects. The threshold is calculated based on the Jenks method that clusters points into similar groups in terms of their attributes (North, 2009). Hence, the threshold is identified based on the similarity between datapoints instead of selecting it subjectively. In Figure 31, the value of RSs is considered to cluster datapoints and calculating the threshold (as shown 0.3 in the figure). As it is shown, the percentage of the match between observation and prediction was 81.9 percent which shows good accuracy in predicting the location of hotspots.

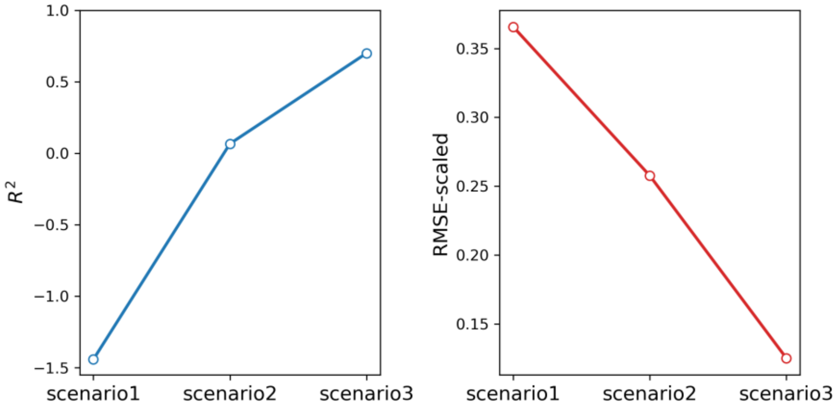


Figure 28 - Accuracy metrics of prediction model for RS of erosion on paved ditches (a) R-squared (b) Scaled Root Mean Squared Error

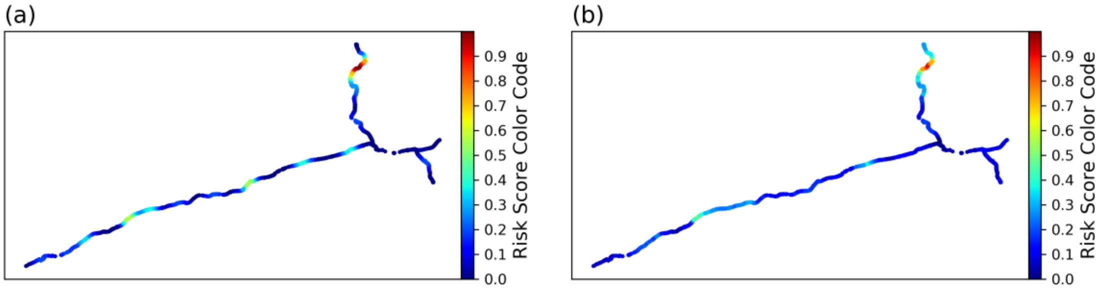


Figure 29 - RSs of erosion on paved ditches at the end of FY2020 (a) observation (b) prediction

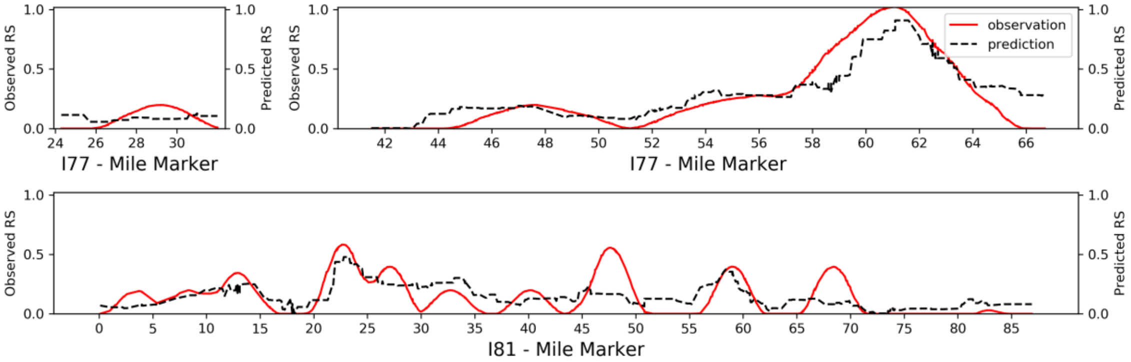


Figure 30 - Longitudinal distribution of RSs of erosion on paved ditches all over case study roadways at the end of FY2020

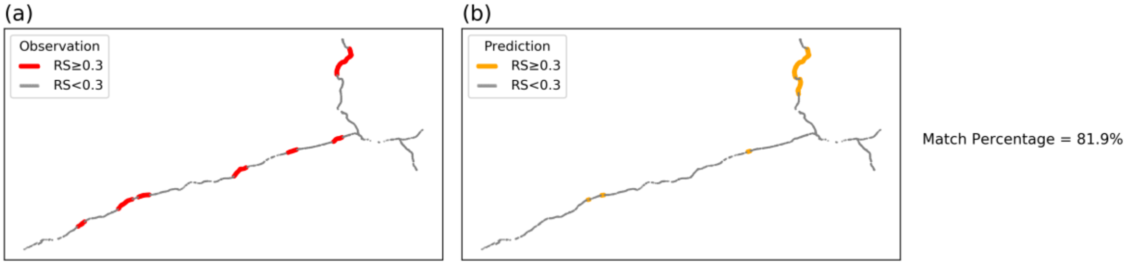


Figure 31 - Match percentage of observed versus predicted RSs of erosion on paved ditches at the end of FY2020

4.5.2 Obstruction Prediction

Like erosion, the developed models for obstruction RSs in three considered scenarios were assessed. Figure 32 illustrates the accuracy metrics of the models in scenarios 1 to 3. As it is shown, like erosion, predictions for obstruction were improved when an additional year of data has been used in the training process. Moreover, the R^2 of predictions after adding two years changed from -0.24 to a value of 0.7 at the end of FY2020. Figure 33 illustrates the spatial distribution of obstruction RSs in two cases of observed and predicted values. Again, like erosion, the location of higher obstruction RSs in those two cases are almost similar. Furthermore, Figure 34 better present the similarity between observations and predictions. Finally, Figure 35 presents the match between actual and forecasted hotspots and coldspots as 96.2 percent.

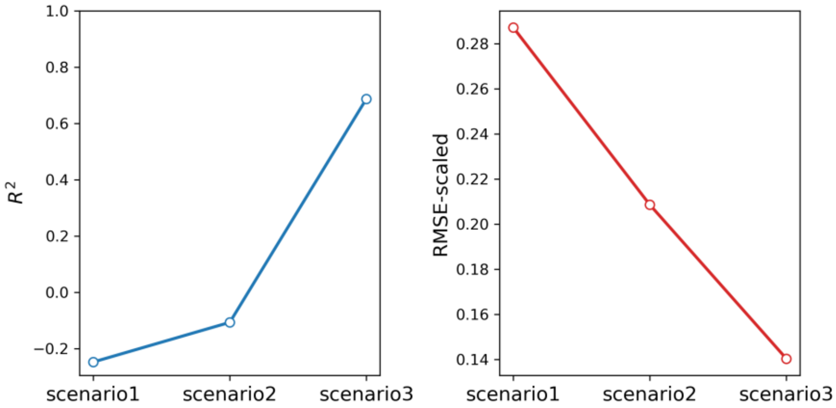


Figure 32 - Accuracy metrics of prediction model for RS of obstruction on paved ditches (a) R-squared (b) Scaled Root Mean Squared Error

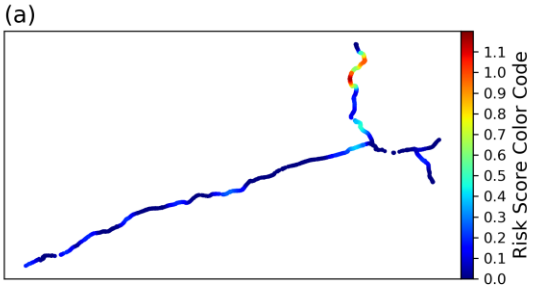


Figure 33 - RSs of cracking on paved ditches at the end of FY2020 (a) observation (b) prediction

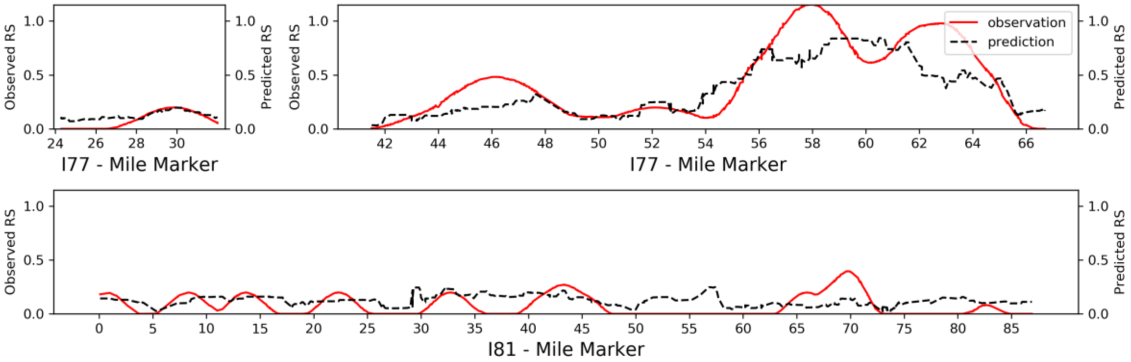


Figure 34 - Longitudinal distribution of RSs of obstruction on paved ditches all over case study roadways at the end of FY2020

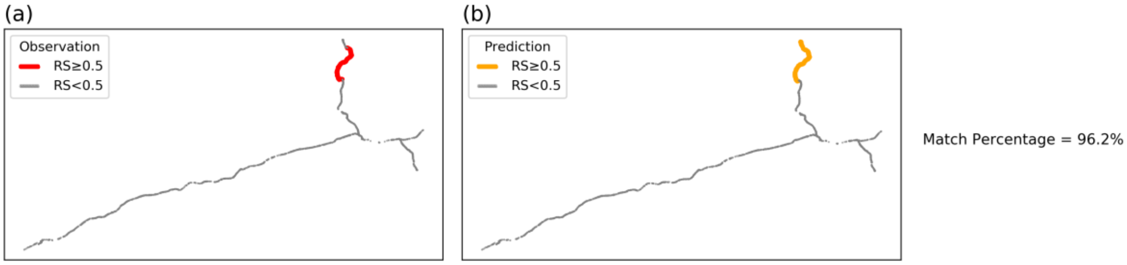


Figure 35 - Match percentage of observed versus predicted RSs of obstruction on paved ditches at the end of FY2020

4.5.3 Cracking Prediction

For cracking, the trend of improvement in the accuracies by adding years of new data in the training process was still increasing. However, the trend had a slighter slope compared to erosion and obstruction cases. As shown in Figure 36, The R^2 of the scenario3 for cracking was 0.25 which is much less than that of erosion and obstruction cases. Although the prediction model in scenario3 had low accuracy in terms of its R^2 value it provides a reasonable prediction of hotspots of cracking as shown in Figures 37 and 38. The results show that the percentage of the match between observed and forecasted cracking hotspots and coldspots was 96.1 percent, as shown in Figure 39.

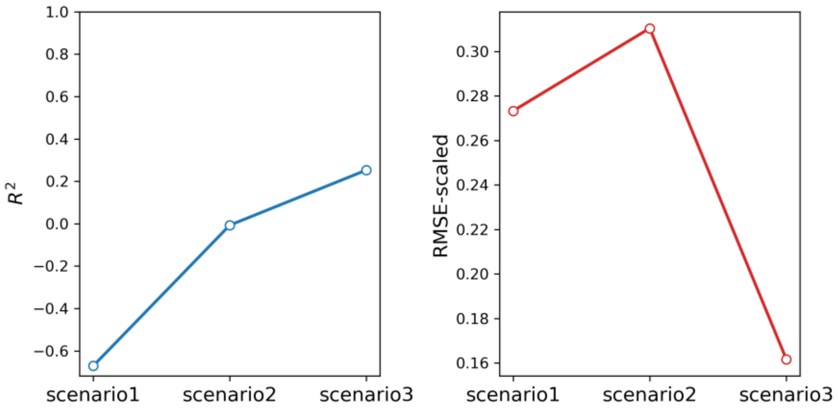


Figure 36 - Accuracy metrics of prediction model for RS of cracking on paved ditches (a) R-squared (b) Scaled Root Mean Squared Error

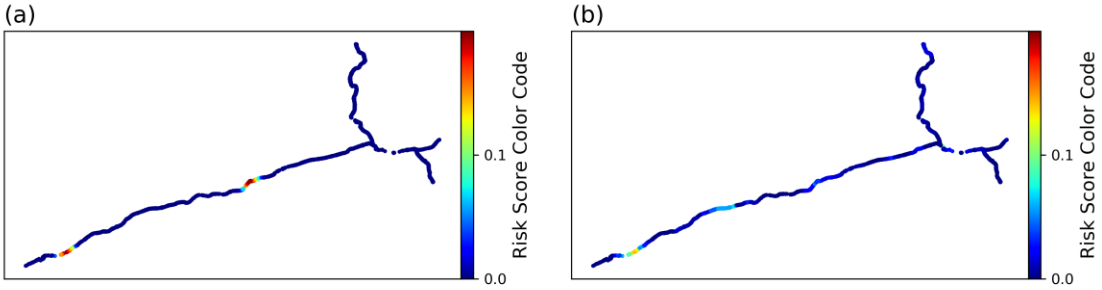


Figure 37 - RSs of erosion on paved ditches at the end of FY2020 (a) observation (b) prediction

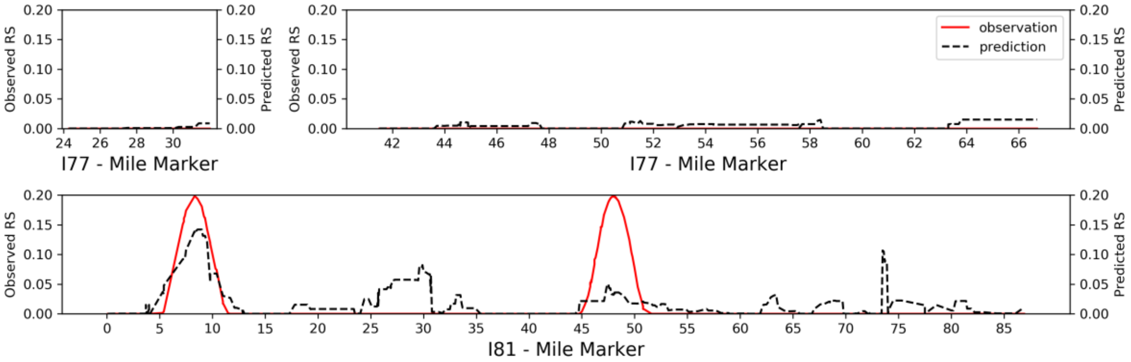


Figure 38 - Longitudinal distribution of RSs of cracking on paved ditches all over case study roadways at the end of FY2020

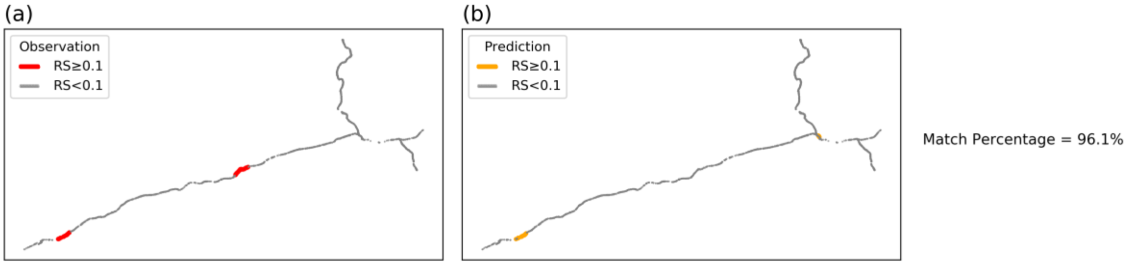


Figure 39 - Match percentage of observed versus predicted RSs of cracking on paved ditches at the end of FY2020

4.6 Web-based Spatial Visualization

In order to spatially visualizing the results of RSs and other attributes of the considered asset items in roadway segments, ArcGIS was deployed. Figures 40 to 43 present produced maps from ArcGIS to show the distribution of multiple contributing factors in the case study. Such that, Figure 40 shows the annual minimum daily temperature in FY2017, Figure 41 presents the number of freezing days in FY2016, Figure 42 illustrates total annual snow depth in FY2018, and Figure 43 reveals total annual precipitation depth in FY2019. Furthermore, the location of machine cleaning works in FY2018 is highlighted in Figure 44. Moreover, the ArcGIS was utilized to produce the map of risk scores in different years. To this end, Figure 45 is an example of illustrating erosion RSs on paved ditches in 2015.

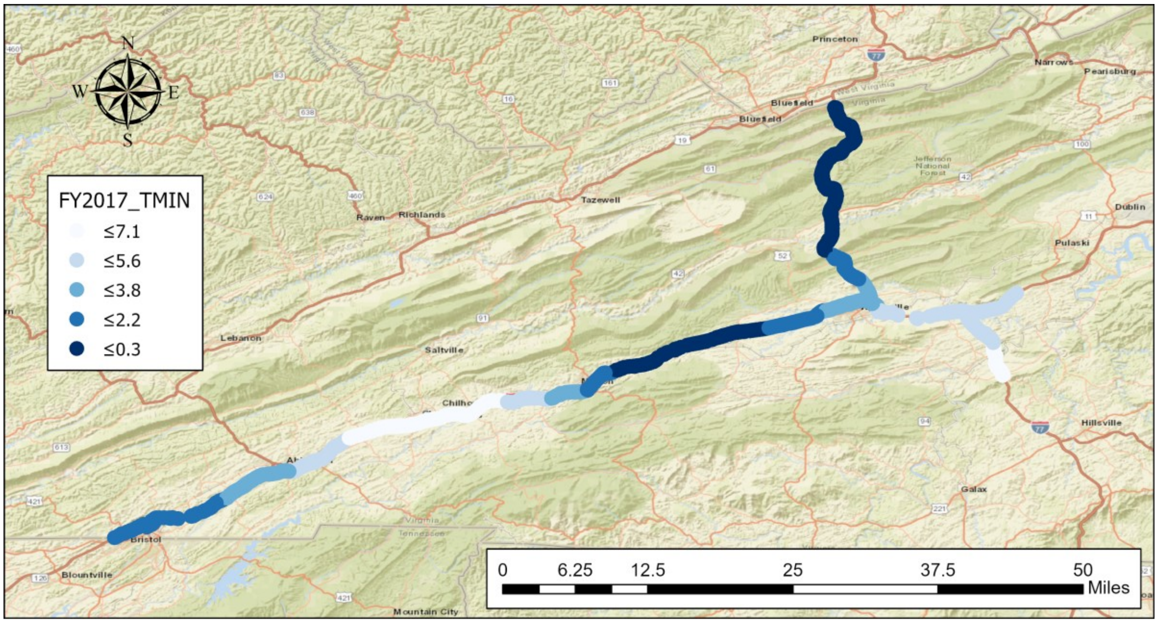


Figure 40 – Annual minimum daily temperature (º F) in FY2017 over the case study

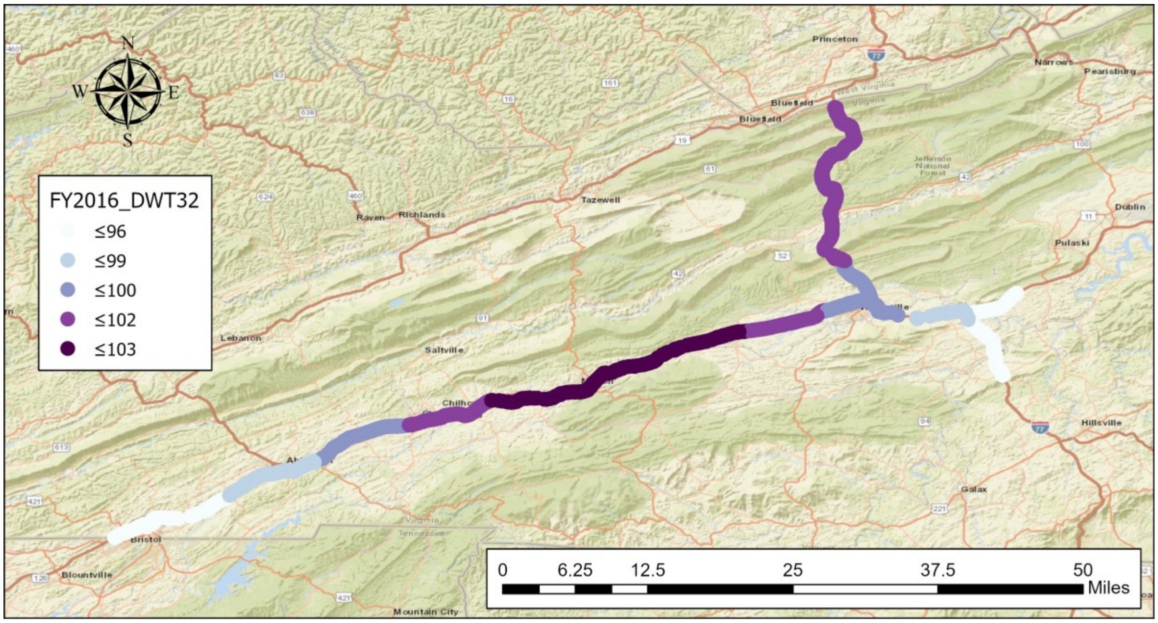


Figure 41 – Number of freezing days in FY2016 over the case study

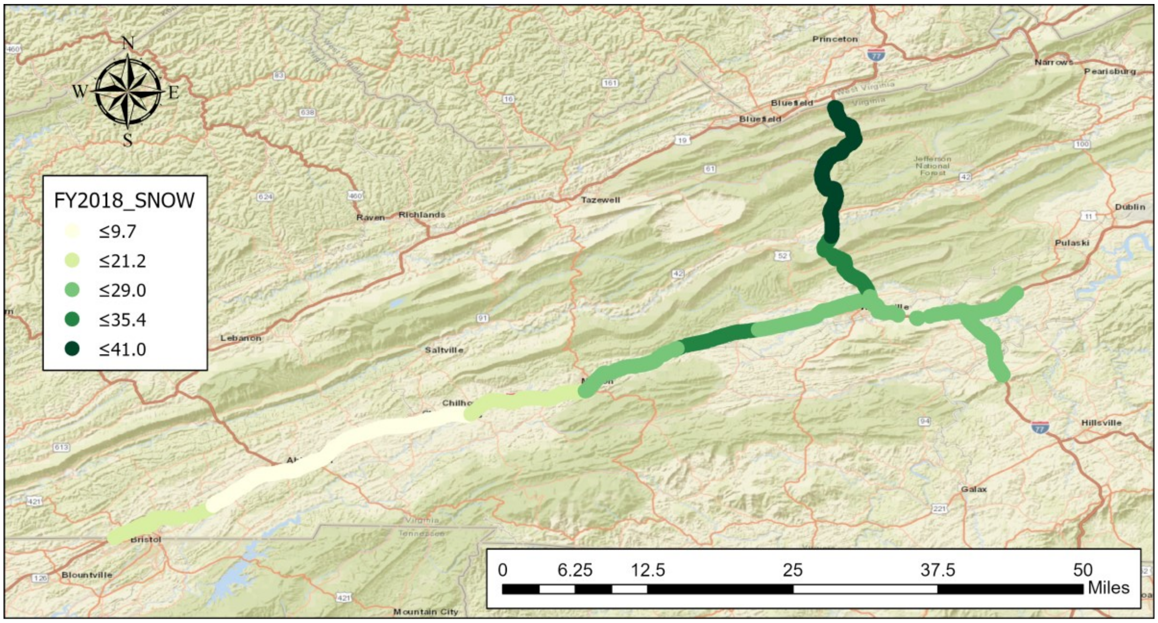


Figure 42 – Total annual snow depth (inch) in FY2018 over the case study

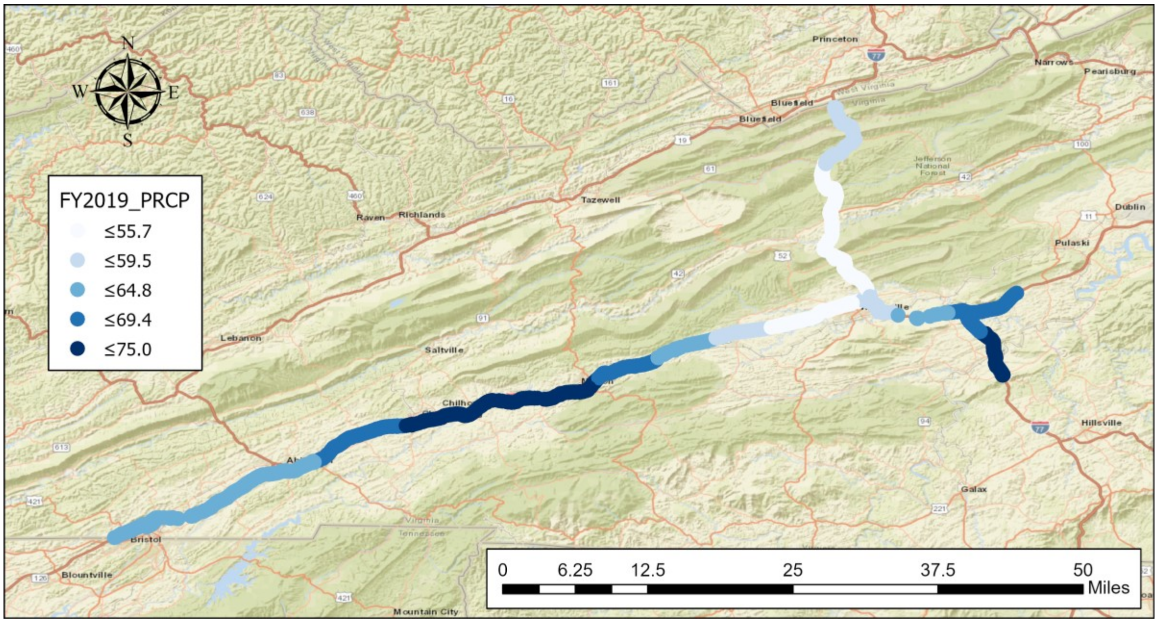


Figure 43 – Total annual precipitation depth (inch) in FY2019 over the case study

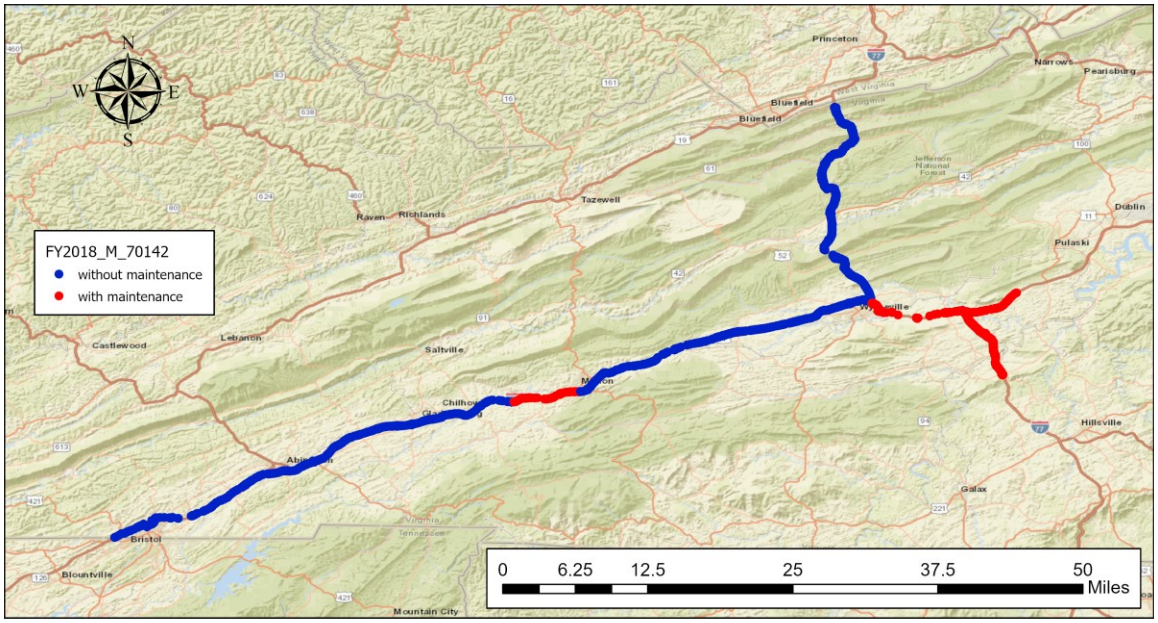


Figure 44 – Location of maintenance with code 70142 (machine cleaning) in FY2018 over the case study

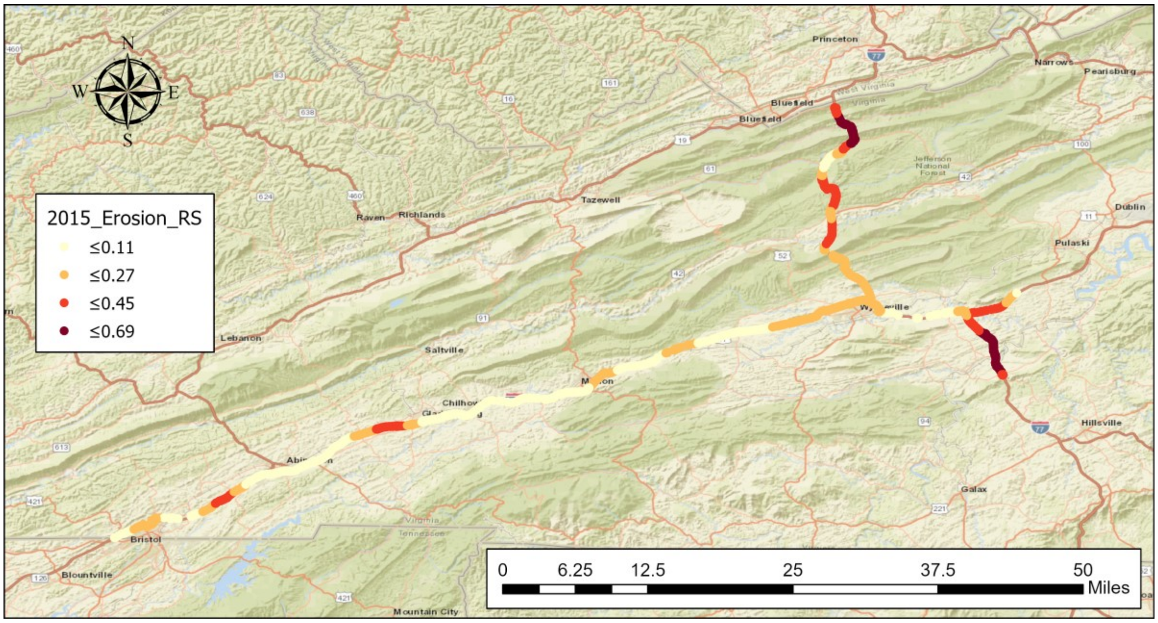
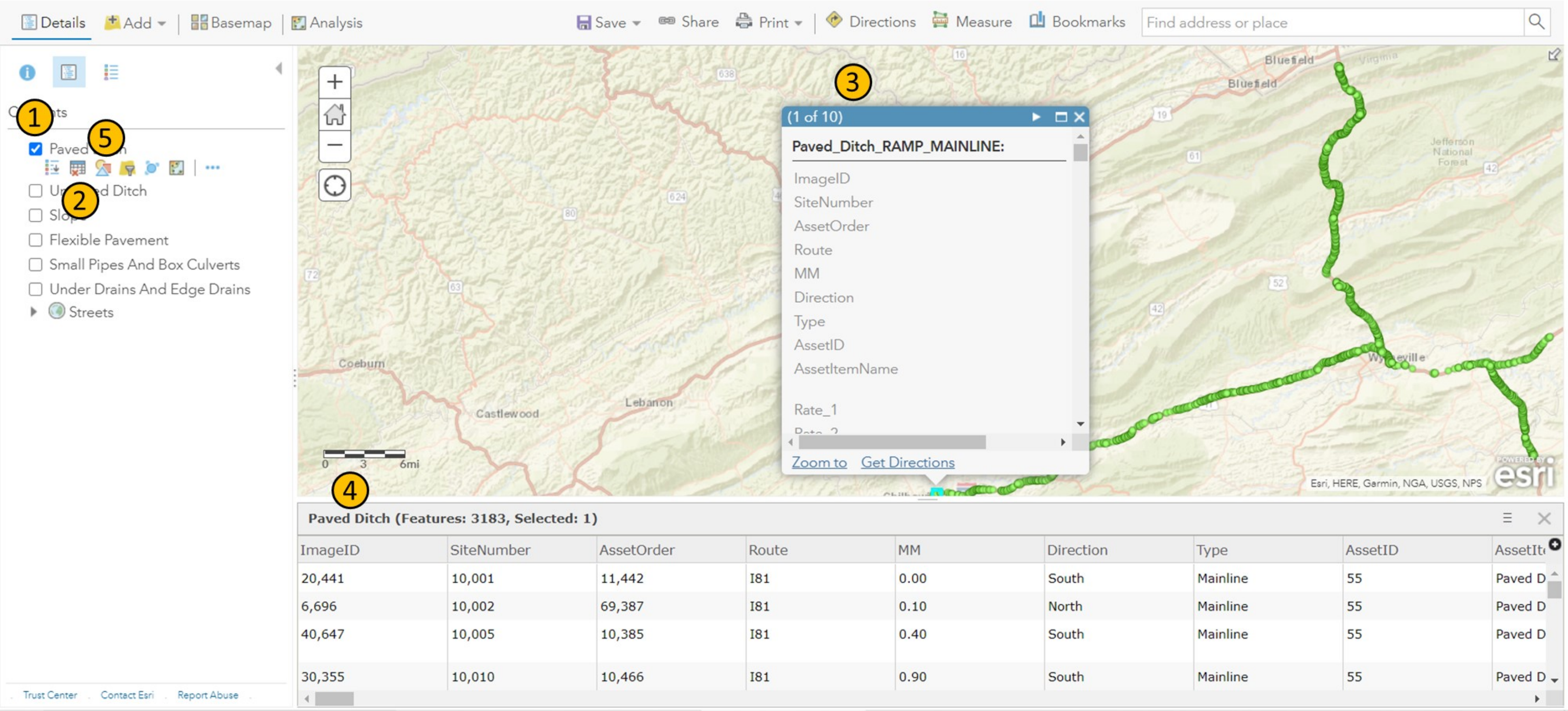


Figure 45 – Risk scores of erosion on paved ditches in 2015 over the case study

ArcGIS Online was also utilized for presentation purposes. Appendix A illustrates how the shapefiles containing geographical locations and all attributes of asset items (weather, traffic, inspection, maintenance, risk scores) were imported into the ArcGIS Online environment. As specified in this appendix, first all created shapefiles were added to ArcGIS Online. Then, they were added to the map viewer that provides interactive visualization and presentation capabilities. All attributes of each point are accessible through the attribute table of this asset type or by clicking on that point, as shown in Figure 46. In addition, visualizing the spatial distribution of each attribute can be fulfilled by selecting and changing the options for visualization, as presented in Appendix A.



(1 of 10)

Paved_Ditch_RAMP_MAINLINE:

ImageID
SiteNumber
AssetOrder
Route
MM
Direction
Type
AssetID
AssetItemName
Rate_1

Zoom to Get Directions

Paved Ditch (Features: 3183, Selected: 1)

ImageID	SiteNumber	AssetOrder	Route	MM	Direction	Type	AssetID	AssetIt
20,441	10,001	11,442	I81	0.00	South	Mainline	55	Paved D
6,696	10,002	69,387	I81	0.10	North	Mainline	55	Paved D
40,647	10,005	10,385	I81	0.40	South	Mainline	55	Paved D
30,355	10,010	10,466	I81	0.90	South	Mainline	55	Paved D

1 Selecting a certain asset type
2 Opening asset's attribute table
3 Attribute information of a selected data point
4 Attribute table of the selected asset
5 Visualization tool

Figure 46 - Accessing attribute information of asset items in ArcGIS Online

CHAPTER 5: CONCLUSION

This study provides a framework for a multi-asset hotspot prediction model that points out the susceptibility of road segments to a set of selected defects. It is empowered by combining a risk factor generator (Kernel Density Estimation) and a machine learning algorithm selected from a pool of candidates to provide the most accurate results. This combination enables the proposed framework to predict the probability of future defects given a wide range of historical information, including weather, traffic, and historical inspection and maintenance data.

The proposed framework provided significant accuracy in the extent of the case study data for forecasting the risk scores of erosion, obstruction, and cracking on paved ditches based on real observations. The outcomes corroborated the interrelation between adjacent assets and their contribution to future defects. For instance, the contribution of outlet defects of the downstream under drains and edge drains on the obstruction on upstream paved ditches was unveiled. As another example, the contribution of the prior year erosion on unpaved ditches and lower-slope issue on slopes in the next year erosion on paved ditches was revealed.

The comparative study showed that the hotspot prediction in the extent of the case study followed a nonlinear pattern, with Random Forest Regression (RFR) being the most accurate algorithm in this problem. The already proven performance of RFR in unbalanced data with categorical features resulted in its outperforming other selected linear and nonlinear algorithms in the case study. Not only its performance in terms of R^2 was the most accurate among the selected algorithms, but also the variation among the residuals was the least, which further corroborates it being the best fit.

Even though the proposed framework is a multi-asset risk score predictor, it respects the different essence of each selected asset and investigates the impact of contributors on different asset types appropriately. For example, the findings of this study show that while prior year erosion on paved ditches contributed the most to the predicted erosion, obstruction predictions were mostly influenced by the annual maximum temperature feature. The methodology also identified the annual average min-max daily temperature difference and annual maximum daily temperature as the most influential parameters in predicting next year's cracking, which makes sense. Furthermore, evidence of maintenance impacting both erosion and obstruction was identified in the results despite its negligible influence on cracking.

The efficiency of the developed prediction models was also investigated in a way that they will be deployed by agencies to forecast RSs. The results unveiled that adding one year of data to the training dataset increases the accuracy of the predictions and this trend continues by feeding more years of data when the model is trained. Therefore, after a few years, the model is capable to capture possible scenarios of the variation of each contributing factor in the region of the case study and to provide reliable predictions.

This study provides highway asset managers with a method to predict and explore parts of roadways that are prone to a certain defect. Hence, agencies can plan and prioritize maintenance activities based on the outcomes of the models. Furthermore, the proposed methodology substitutes the independent investigation of each asset’s deterioration with an integrated estimator of defects’ probability for various assets. Therefore, this study can be leveraged to provide a holistic view of the future condition of a roadway system in terms of probable defects of multiple asset types. Consequently, it has the potential of benefiting risk mitigation plans for the whole highway infrastructure. In addition, the framework can be used in locating the segments with higher risk assessments in terms of multiple defects. As a result, maintenance plans can be enriched by such information and be optimized accordingly. Moreover, the proposed framework can be applied to other linear or network infrastructures such as sewers, water networks, and railroads.

The proposed framework also helps agencies to prioritize their future inspections with more concentration on the locations with a higher probability of defects. Consequently, after a few years, more data will be available in susceptible parts of roadways and accurate maintenance decisions can be made utilizing more available data.

By providing prediction models considering several contributors to the degradation, the outcome of this study can be utilized as a tool to examine different scenarios of future variation of the contributors. For example, what-if analysis can be performed to find out what are the impacts of temperature increase resulted from climate change in a region. Furthermore, the implications of variations in traffic patterns resulted from urban developments, and their impacts on RSs can be investigated by using the developed prediction models.

5.1 Limitations of Study and Recommendations for Future Works

As a limitation of this study, only five years of inputs were used, and six asset types were considered due to data availability issues. However, the proposed framework does not have any constraints on the scope of inputs. Therefore, it is suggested that future studies consider applying the methodology on other road assets and cover a more extended scope of time. Besides, the time period of available data (five years) might not be long enough to contain the most probable range of variation for each contributing factor in the region of study. Hence, it is recommended to investigate the minimum time period of historical data required for thoroughly training the model to capture the possible variation of contributing factors in each region.

In this study, due to the limitations in available data, some of the characteristics of assets such as their size and age were not considered in developing prediction models. Since the characteristics of assets are believed to impact their degradation trend, it is recommended to take into consideration these factors in forecasting risk scores of defects as a future research study. Also, the characteristics of the soils on which or inside which road assets are built or installed directly impact the condition of the assets, however, the data were not accessible in this study. Therefore, considering the characteristics of soil layers at the location of assets is suggested for future investigations of prediction models. Furthermore, the parameters that represent the topography of each region (e.g. slope of roadways) can be added to the set of potential contributors in future studies.

Data-driven prediction models are built upon the historical data and are valid within the variation range of the constructive features. Hence, their forecasting performance outside this scope might be questionable. In this study, a data-driven approach was provided to come up with prediction models based on historical weather, traffic, condition, and maintenance data. Since the models were built based on the trends and patterns inside the input dataset during the considered timeframe (i.e. five years), changing patterns and trends in future results in forecasts utilizing extrapolations. Changing traffic patterns, and climate change in the future are two examples of this scenario. Recently, weather extremes such as floods, tropical cyclones, heat waves, and heavy storms are becoming more intense and frequent compared to the past due to climate warming. Additionally, traffic patterns have changed recently due to the COVID-19 pandemic. Therefore, it is recommended to investigate how the change of patterns inside the feature space might impact the accuracy of the proposed framework and examine how to incorporate this change into predictions. Also, integrating expert opinions and knowledge-based judgments into the prediction stage for improving the accuracy of forecasts can be examined when future variations of the features are different from the historical data.

In this study, eight machine learning algorithms were proposed to develop prediction models in two main groups: linear and non-linear regression. Sometimes, these algorithms are clearly interpretable and users can simply explain how they work, how they produce forecasts, and what are the most influential parameters. Linear regression is an example of this category of algorithms known as white-box models. However, some machine learning algorithms, such as ANN, are more complicated to understand and interpret. These kinds of algorithms are known as black-box models. Since the interpretation of the model helps practitioners to better understand the impact of each feature on the prediction results, white-box models gained more attention and deployed widely in the industry. However, sometimes black-box models outperform white-box algorithms, hence, an investigation on the trade-off between the accuracy and interpretability of the models is recommended for future studies.

The results of this study unveiled that Random Forest Regression (RFR) provided the highest accuracy in the case study problem when most of the Risk Scores (RSs) were small values or zero. Other techniques are also utilized in the same cases such as Zero-inflated Poisson regression. Hence, examining the performance of these algorithms is recommended in a future study.

REFERENCES

AASHTO. (2011). *AASHTO transportation asset management guide: A focus on implementation*. Washington D.C.: AASHTO.

Abaza, K. A. (2017). Empirical Markovian-based models for rehabilitated pavement performance used in a life cycle analysis approach. *Structure and Infrastructure Engineering, 13*(5), 625-636.

Aksoy, S., & Haralick, R. M. (2001). Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern recognition letters, 22*(5), 563-582.

Al-Mansour, A. I., Sinha, K. C., & Kuczek, T. (1994). Effects of routine maintenance on flexible pavement condition. *Journal of Transportation Engineering, 120*(1), 65-73.

Anderson, C. J., Claman, D., & Mantilla, R. (2015). Iowa’s bridge and highway climate change and extreme weather vulnerability assessment pilot.

Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention, 41*(3), 359-364.

Anyala, M., Odoki, J., & Baker, C. (2014). Hierarchical asphalt pavement deterioration model for climate impact studies. *International Journal of Pavement Engineering, 15*(3), 251-266.

Bannour, A., El Omari, M., Lakhal, E. K., Afechkar, M., Benamar, A., & Joubert, P. (2017). Optimization of the maintenance strategies of roads in Morocco: calibration study of the degradations models of the highway development and management

(HDM-4) for flexible pavements. *International Journal of Pavement Engineering*, 1-10.

Berardi, V. L., & Zhang, G. P. (2003). An empirical investigation of bias and variance in time series forecasting: modeling considerations and error evaluation. *IEEE Transactions on Neural Networks, 14*(3), 668-679.

Bolstad, P. (2016). *GIS fundamentals: A first text on geographic information systems*: Eider (PressMinnesota).

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/A:1010933404324

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology, 44*(1), 108-132.

Bujang, M. A., Sa'at, N., & Bakar, T. M. I. T. A. (2017). Determination of minimum sample size requirement for multiple linear regression and analysis of covariance based on experimental and non-experimental studies. *Epidemiology, Biostatistics and Public Health, 14*(3).

Chainey, S., & Ratcliffe, J. (2013). *GIS and crime mapping*: John Wiley & Sons.

Chen, & Liu, X. (2019). Roadway Asset Inspection Sampling Using High-Dimensional Clustering and Locality-Sensitivity Hashing. *Computer-Aided Civil and Infrastructure Engineering, 34*(2), 116-129.

Chen, D., Cavalline, T., Ogunro, V., & Thompson, D. (2014). Development and validation of pavement deterioration models and analysis weight factors for the NCDOT pavement management system. *Rep. No. FHWA/NC/2011-01, Federal Highway Administration (FHWA), Washington, DC.*

Chen, D., & Mastin, N. (2016). Sigmoidal models for predicting pavement performance conditions. *Journal of Performance of Constructed Facilities, 30*(4), 04015078. doi:10.1061/(ASCE)CF.1943-5509.0000833

Chimba, D., Emaasit, D., Allen, S., Hurst, B., & Nelson, M. (2014). Factors affecting median cable barrier crash frequency: new insights. *Journal of Transportation Safety & Security, 6*(1), 62-77.

Chopra, T., Parida, M., Kwatra, N., & Chopra, P. (2018). Development of Pavement Distress Deterioration Prediction Models for Urban Road Network Using Genetic Programming. *Advances in Civil Engineering, 2018*.

Choummanivong, L., & Martin, T. (2013). *Probabilistic road deterioration model development* (1925037290)

Clarke, S. M., Griebsch, J. H., & Simpson, T. W. (2005). Analysis of support vector regression for approximation of complex engineering analyses.

Coffey, S., & Park, S. (2016). Observational study on the pavement performance effects of shoulder rumble strip on shoulders. *International Journal of Pavement Research and Technology, 9*(4), 255-263.

Cohen, S., & Intrator, N. (2003). *A study of ensemble of hybrid networks with strong regularization.* Paper presented at the International Workshop on Multiple Classifier Systems.

Craig III, W. N., Sitzabee, W. E., Rasdorf, W. J., & Hummer, J. E. (2007). Statistical validation of the effect of lateral line location on pavement marking retroreflectivity degradation. *Public Works Management & Policy, 12*(2), 431-450.

da Silva, A. S. A., Stosic, B., Menezes, R. S. C., & Singh, V. P. (2019). Comparison of Interpolation Methods for Spatial Distribution of Monthly Precipitation in the State of Pernambuco, Brazil. *Journal of Hydrologic Engineering, 24*(3), 04018068.

de Amorim Borges, P., Franke, J., da Anunciação, Y. M. T., Weiss, H., & Bernhofer, C. (2016). Comparison of spatial interpolation methods for the estimation of precipitation distribution in Distrito Federal, Brazil. *Theoretical and applied climatology, 123*(1-2), 335-348.

Elwakil, E., Eweda, A., & Zayed, T. (2014). Modelling the effect of various factors on the condition of pavement marking. *Structure and Infrastructure Engineering, 10*(1), 93-105.

Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of safety research, 40*(5), 341-351.

Ferrante, L., & Cameriere, R. (2009). Statistical methods to assess the reliability of measurements in the procedures for forensic age estimation. *International journal of legal medicine, 123*(4), 277-283.

Ford, K. M., Arman, M., Labi, S., Sinha, K. C., Thompson, P., Shirole, A., & Li, Z. (2012). *Estimating Life Expectancies of Highway Assets - Volume 2: Final Report*

Forsyth, R. A., Wells, G. K., & Woodstrom, J. H. (1987). *Economic impact of pavement subsurface drainage*.

Frazier, A. G., Giambelluca, T. W., Diaz, H. F., & Needham, H. L. (2016). Comparison of geostatistical approaches to spatially interpolate month-year rainfall for the Hawaiian Islands. *International Journal of Climatology, 36*(3), 1459-1470.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1): Springer series in statistics New York.

FTA. (2004). *Risk analysis methodologies and procedures*. Washington DC.

Gao, L., Aguiar-Moya, J. P., & Zhang, Z. (2012). Bayesian analysis of heterogeneity in modeling of pavement fatigue cracking. *Journal of Computing in Civil Engineering, 26*(1), 37-43. doi:10.1061/(ASCE)CP.1943-5487.0000114

Gaull, B., Michael-Leiba, M., & Rynn, J. (1990). Probabilistic earthquake risk maps of Australia. *Australian Journal of Earth Sciences, 37*(2), 169-187.

Ghabchi, R., Zaman, M., Khoury, N., Kazmee, H., & Solanki, P. (2013). Effect of gradation and source properties on stability and drainability of aggregate bases: a laboratory and field study. *International Journal of Pavement Engineering, 14*(3), 274-290.

Haimes, Y. Y. (2005). *Risk modeling, assessment, and management* (Vol. 40): John Wiley & Sons.

Halmen, C., Trejo, D., & Folliard, K. (2008). Service Life of Corroding Galvanized Culverts Embedded in Controlled Low-Strength Materials. *Journal of Materials in Civil Engineering, 20*(5), 366-374. doi:10.1061/(ASCE)0899-1561(2008)20:5(366)

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences, 44*(1), 1-12.

Henning, T. F., Alabaster, D., Arnold, G., & Liu, W. (2014). Relationship between traffic loading and environmental factors and low-volume road deterioration. *Transportation Research Record, 2433*(1), 100-107.

Hong, F., & Prozzi, J. A. (2010). Roughness model accounting for heterogeneity based on in-service pavement performance data. *Journal of Transportation Engineering, 136*(3), 205-213.

Hunt, R. E. (1992). Slope failure risk mapping for highways: methodology and case history. *Transportation Research Record*(1343).

Immaneni, V. P., Hummer, J. E., Rasdorf, W. J., Harris, E. A., & Yeom, C. (2009). Synthesis of sign deterioration rates across the United States. *Journal of Transportation Engineering, 135*(3), 94-103. doi:10.1061/(ASCE)0733-947X(2009)135:3(94)

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666.

Karabulut, E. M., & Ibrikci, T. (2014). Analysis of cardiotocogram data for fetal distress determination by decision tree based adaptive boosting approach. *Journal of Computer and Communications, 2*(9), 32-37.

Karimzadeh, A., Sabeti, S., Burde, A., Tabkhi, H., & Shoghli, O. (2020). *Spatial-Temporal Deterioration of Multiple Highway Assets: A Correlational Study*. Paper presented at the ASCE Construction Research Congress (CRC) - 2020, Tempe, Arizona.

Karimzadeh, A., Sabeti, S., & Shoghli, O. (2020). *Clustering-based Similarity Detection of Pavement Segments Considering Multiple Contributors to Deterioration*. Paper presented at the ASCE Construction Research Congress (CRC-2020), Tempe, Arizona.

Karimzadeh, A., Sabeti, S., Tabkhi, H., & Shoghli, O. (2020). *Condition Prediction of Highway Assets Based on Spatial Proximity and Interrelations of Asset Classes: A*

Case Study. Paper presented at the 37th International Symposium on Automation and Robotics in Construction (ISARC), Kitakyshu, Japan.

Karimzadeh, A., & Shoghli, O. (2020). Predictive Analytics for Roadway Maintenance: A Review of Current Models, Challenges, and Opportunities. *Civil Engineering Journal, 6*(3), 602-625.

Karlaftis, A. G., & Badr, A. (2015). Predicting asphalt pavement crack initiation following rehabilitation treatments. *Transportation Research Part C: Emerging Technologies, 55*, 510-517.

Kuter, N., & Kuter, S. (2018). Investigation of wildfire at forested landscapes: A novel contribution to nonparametric density mapping at regional scale. *Applied Ecology and Environmental Research, 16*(4), 4701-4716.

Labi, S., & Sinha, K. C. (2003). Measures of short-term effectiveness of highway pavement maintenance. *Journal of Transportation Engineering, 129*(6), 673-683.

Lambourne, K., & Tomporowski, P. (2010). The effect of exercise-induced arousal on cognitive task performance: a meta-regression analysis. *Brain research, 1341*, 12-24.

Leggetter, C., & Woodland, P. C. (1994). *Speaker adaptation of continuous density HMMs using multivariate linear regression.* Paper presented at the Third International Conference on Spoken Language Processing.

Lin, Y. C., Paul, A., Corotis, R. B., & Liel, A. B. (2015). Framework methodology for risk-based decision making for transportation agencies. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 1*(3), 04015006.

Lu, D. (2020). Pavement Flooding Risk Assessment and Management in the Changing Climate.

Luo, Z., & Chou, E. Y. (2006). Pavement condition prediction using clusterwise regression. *Transportation Research Record, 1974*(1), 70-77.

Madaio, M., Chen, S.-T., Haimson, O. L., Zhang, W., Cheng, X., Hinds-Aldrich, M., . . . Dilkina, B. (2016). *Firebird: Predicting fire risk and prioritizing fire inspections in Atlanta.* Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Malyuta, D. A. (2015). *Analysis of Factors Affecting Pavement Markings and Pavement Marking Retroreflectivity in Tennessee Highways.* University of Tennessee at Chattanooga.

Markow, M. J. (2007). *Managing selected transportation assets: Signals, lighting, signs, pavement markings, culverts, and sidewalks*. Washington D.C.

Massada, A. B., Radeloff, V. C., Stewart, S. I., & Hawbaker, T. J. (2009). Wildfire risk in the wildland–urban interface: a simulation study in northwestern Wisconsin. *Forest Ecology and Management, 258*(9), 1990-1999.

Millington, J., Romero-Calcerrada, R., Wainwright, J., & Perry, G. (2008). An agent-based model of Mediterranean agricultural land-use/cover change for examining wildfire risk. *Journal of Artificial Societies and Social Simulation, 11*(4), 4.

Mills, L. N. O., Attoh-Okine, N. O., & McNeil, S. (2012). Developing pavement performance models for Delaware. *Transportation Research Record, 2304*(1), 97-103.

NASEM. (2019). *Critical Issues in Transportation 2019*. The National Academies of Science, Engineering & Medicine: The National Academies Press.

North, M. A. (2009). *A method for implementing a statistically significant number of data classes in the Jenks algorithm.* Paper presented at the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery.

Pantuso, A., Flintsch, G. W., Katicha, S. W., & Loprencipe, G. (2019). Development of network-level pavement deterioration curves using the linear empirical Bayes approach. *International Journal of Pavement Engineering*, 1-14.

Plouffe, C. C., Robertson, C., & Chandrapala, L. (2015). Comparing interpolation techniques for monthly rainfall mapping using multiple evaluation criteria and auxiliary data sources: A case study of Sri Lanka. *Environmental Modelling & Software, 67*, 57-71.

Proctor, G., & Varma, S. (2012). *Risk-Based Transportation Asset Management: Evaluating Threats, Capitalizing on Opportunities: Report 1: Overview of Risk Management*

Prozzi, J. A., Serigos, P. A., Kim, M. Y., & Xu, H. (2017). *Deterioration Modelling of Preventive Maintenance Treatments for Flexible Pavements*

Radopoulou, S. C., & Brilakis, I. (2016). Improving road asset condition monitoring. *Transportation Research Procedia, 14*(0), 3004-3012.

Rahman, M. K., Crawford, T., & Schmidlin, T. W. (2018). Spatio-temporal analysis of road traffic accident fatality in Bangladesh integrating newspaper accounts and gridded population data. *GeoJournal, 83*(4), 645-661.

Ré, J., Miles, J., & Carlson, P. (2011). Analysis of in-service traffic sign retroreflectivity and deterioration rates in Texas. *Transportation Research Record: Journal of the Transportation Research Board*(2258), 88-94.

Renn, O. (2008). *Risk governance: coping with uncertainty in a complex world*: Earthscan.

Rynkiewicz, J. (2012). General bound of overfitting for MLP regression models. *Neurocomputing, 90*, 106-110.

Saha, P., Ksaibati, K., & Atadero, R. (2017). Developing Pavement Distress Deterioration Models for Pavement Management System Using Markovian Probabilistic Process. *Advances in Civil Engineering, 2017*.

Saliminejad, S., & Gharaibeh, N. G. (2016). Proximity-based outlier detection method for roadway infrastructure condition data. *Journal of Computing in Civil Engineering, 30*(1), 04015001.

Sanabria, N., Valentin, V., Bogus, S., Zhang, G., & Kalhor, E. (2017). *Comparing Neural Networks and Ordered Probit Models for Forecasting Pavement Condition in New Mexico*. Paper presented at the Transportation Research Board 96th Annual Meeting, Washington D.C.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification* (pp. 149-171): Springer.

Shoghli, O., & De La Garza, J. M. (2016). A Multi-Objective Decision-Making Approach for the Sustainable Maintenance of Roadways. In *Construction Research Congress 2016* (pp. 1424-1434).

Shoghli, O., & De La Garza, J. M. (2017). Multi-Asset Optimization of Roadways Asset Maintenance. In *Computing in Civil Engineering 2017* (pp. 297-305).

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26): CRC press.

Sitzabee, W. E., White, E. D., & Dowling, A. W. (2012). Degradation modeling of polyurea pavement markings. *Public Works Management & Policy, 18*(2), 185-199.

Sohn, J. (2006). Evaluating the significance of highway network links under the flood damage: An accessibility approach. *Transportation research part A: policy and practice, 40*(6), 491-506.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research, 15*(1), 1929-1958.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics, 8*(1), 25.

Suen, Y. L., Melville, P., & Mooney, R. J. (2005). *Combining bias and variance reduction techniques for regression trees.* Paper presented at the European Conference on Machine Learning.

Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association, 98*(463), 750-763.

Swargam, N. (2004). *Development of a neural network approach for the assessment of the performance of traffic sign retroreflectivity.* (M.S. Thesis), Lousiana State University, Civil and Environmental Engineering, Baton Rouge, LA.

Thakali, L., Kwon, T. J., & Fu, L. (2015). Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation, 23*(2), 93-106.

Thanassoulis, E. (1993). A comparison of regression analysis and data envelopment analysis as alternative methods for performance assessments. *Journal of the operational research society, 44*(11), 1129-1144.

Tighe, S., He, Z., & Haas, R. (2001). Environmental deterioration model for flexible pavement design: an Ontario example. *Transportation Research Record, 1755*(1), 81-89.

Titus-Glover, L. (2019). Unsupervised extraction of patterns and trends within highway systems condition attributes data. *Advanced Engineering Informatics, 42*, 100990.

Toole, T., Martin, T., Roberts, J., Kadar, P., & Byrne, M. (2007). *Guide to asset management part 5H: performance modelling*.

Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy, 32*(9), 1761-1768.

VDOT. (2014). *Bundled Interstate Maintenance Services (BIMS): Instructions, Asset and Activity Codes for Reports Manual* https://www.bidnet.com/bneattachments?/563874634.pdf

Vicente-Serrano, S. M., Saz-Sánchez, M. A., & Cuadrat, J. M. (2003). Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature. *Climate research, 24*(2), 161-180.

Wang, C., Wang, Z., & Tsai, Y.-C. (2016). Piecewise Multiple Linear Models for Pavement Marking Retroreflectivity Prediction Under Effect of Winter Weather Events. *Transportation Research Record: Journal of the Transportation Research Board*(2551), 52-61.

Wang, H., & Wang, Z. (2017). Deterministic and probabilistic life-cycle cost analysis of pavement overlays with different pre-overlay conditions. *Road Materials and Pavement Design*, 1-16.

Wang, J., & Wang, X. (2011). *An ontology-based traffic accident risk mapping framework.* Paper presented at the International Symposium on Spatial and Temporal Databases.

Wang, W.-c., Chau, K.-w., Qiu, L., & Chen, Y.-b. (2015). Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. *Environmental research, 139*, 46-54.

Wolters, A. S., & Zimmerman, K. A. (2010). *Current practices in pavement performance modeling project 08-03 (C07): task 4 report final summary of findings*

Wright, L., Chinowsky, P., Strzepek, K., Jones, R., Streeter, R., Smith, J. B., . . . Perkins, W. (2012). Estimated effects of climate change on flood vulnerability of US bridges. *Mitigation and Adaptation Strategies for Global Change, 17*(8), 939-955.

Wu, C.-H., Tzeng, G.-H., & Lin, R.-H. (2009). A Novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Systems with Applications, 36*(3), 4725-4735.

Yacout, S., & Ouali, M. S. (2019). *Using Artificial Intelligence for Block Maintenance of Pavement Segments with Similar Degradation Profile.* Paper presented at the 2019 Annual Reliability and Maintainability Symposium (RAMS).

Yang, S.-I., Frangopol, D. M., & Neves, L. C. (2004). Service life prediction of structural systems using lifetime functions with emphasis on bridges. *Reliability Engineering & System Safety, 86*(1), 39-51.

Yaseen, Z. M., Sulaiman, S. O., Deo, R. C., & Chau, K.-W. (2019). An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology, 569*, 387-408.

Yilmaz, I., & Kaynar, O. (2011). Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. *Expert Systems with Applications, 38*(5), 5958-5966.

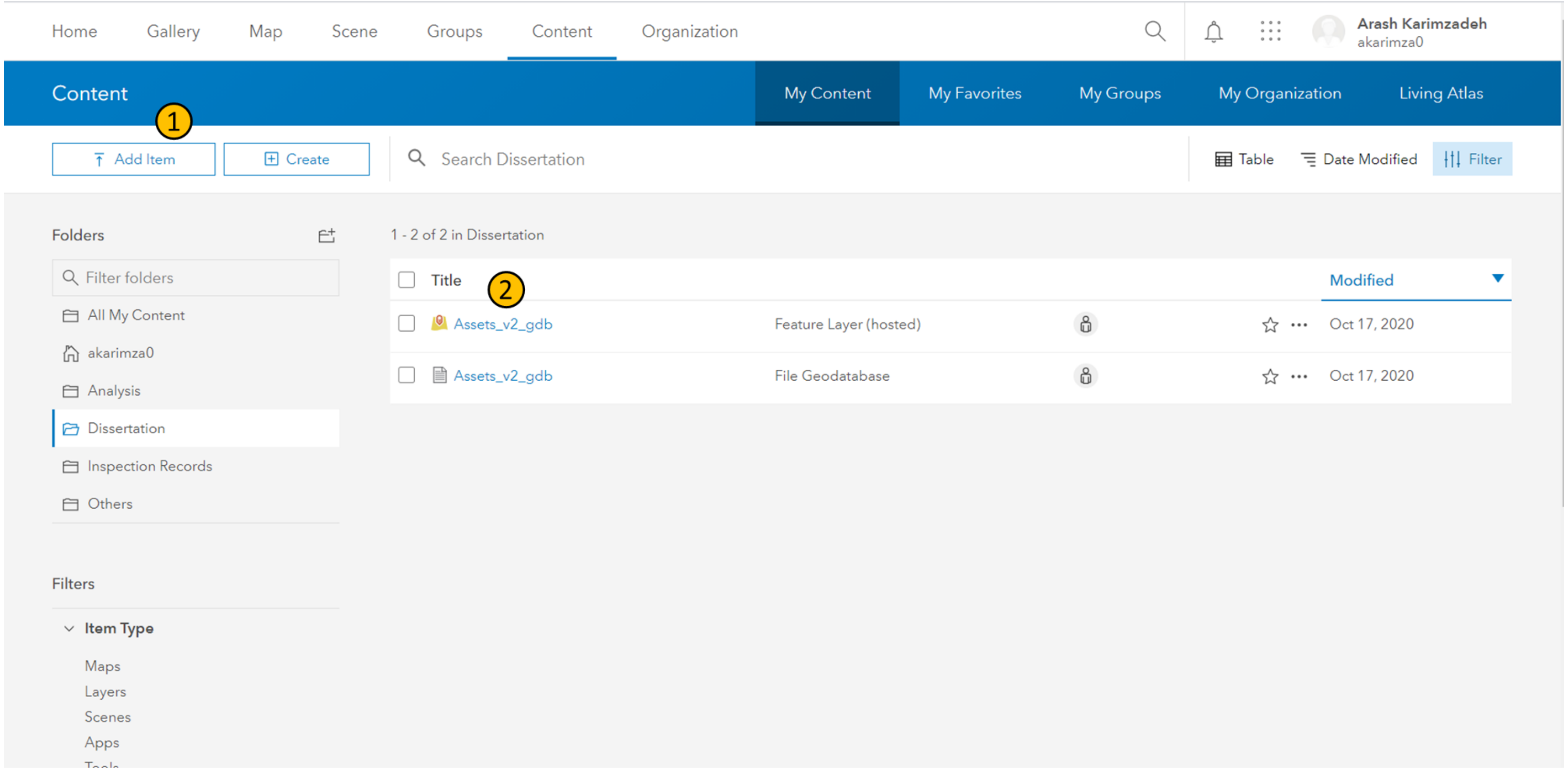
Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q. P., & Lillard Jr, J. W. (2014). A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology, 4*(5), 9.

Yuan, M., Ekici, A., Lu, Z., & Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*(3), 329-346.

Zhu, A. X., Lu, G., Liu, J., Qin, C. Z., & Zhou, C. (2018). Spatial prediction based on Third Law of Geography. *Annals of GIS, 24*(4), 225-240.

APPENDIX A: IMPORTING AND VISUALIZING DATA IN ARCGIS ONLINE

In this appendix, the process of importing asset items data into ArcGIS online is provided in Figure A1. In addition, visualization process of the imported data is shown in Figure A2. After importing all shapefiles, Figure A3 shows an example of how the asset items are visualized in ArcGIS Online. In this figure, data points belonging to the paved ditch asset type are shown. Besides, the environment is equipped with a tool to select a base map as the background of the visualizations, as shown in Figure A3. Moreover, visualizing the spatial distribution of each attribute of each point can be fulfilled by selecting and changing the options for visualization. Figure A4 represents the spatial distribution of erosion risk scores on paved ditches in 2015 all over the case study roadways using ArcGIS online.



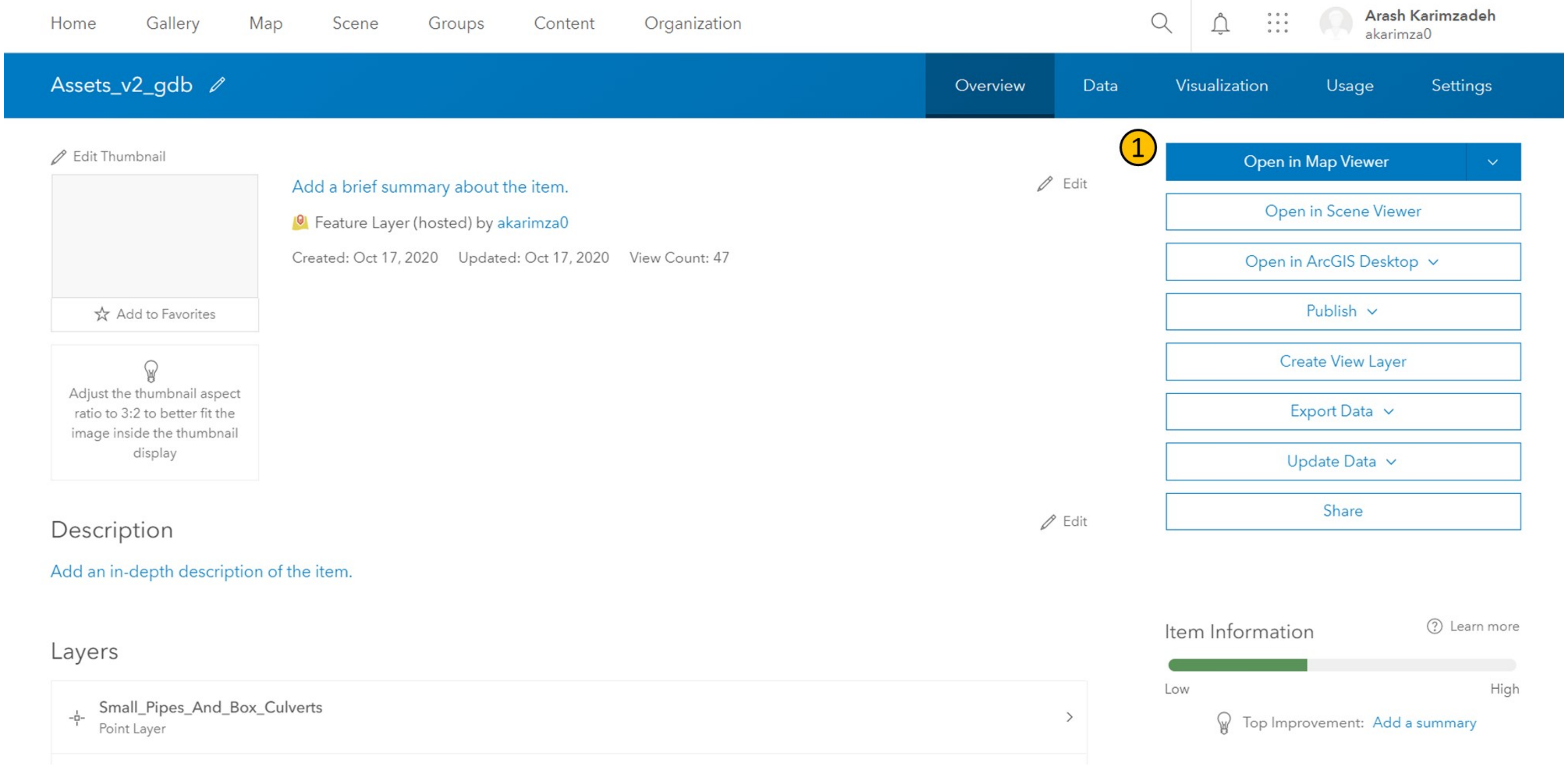
1 - 2 of 2 in Dissertation

Title			Modified
Assets_v2_gdb	Feature Layer (hosted)	☆ •••	Oct 17, 2020
Assets_v2_gdb	File Geodatabase	☆ •••	Oct 17, 2020

1 Importing shapefiles to ArcGIS Online

2 Geodatabase file containing all considered asset types

Figure A1 - Importing shapefiles of the considered asset types into ArcGIS Online



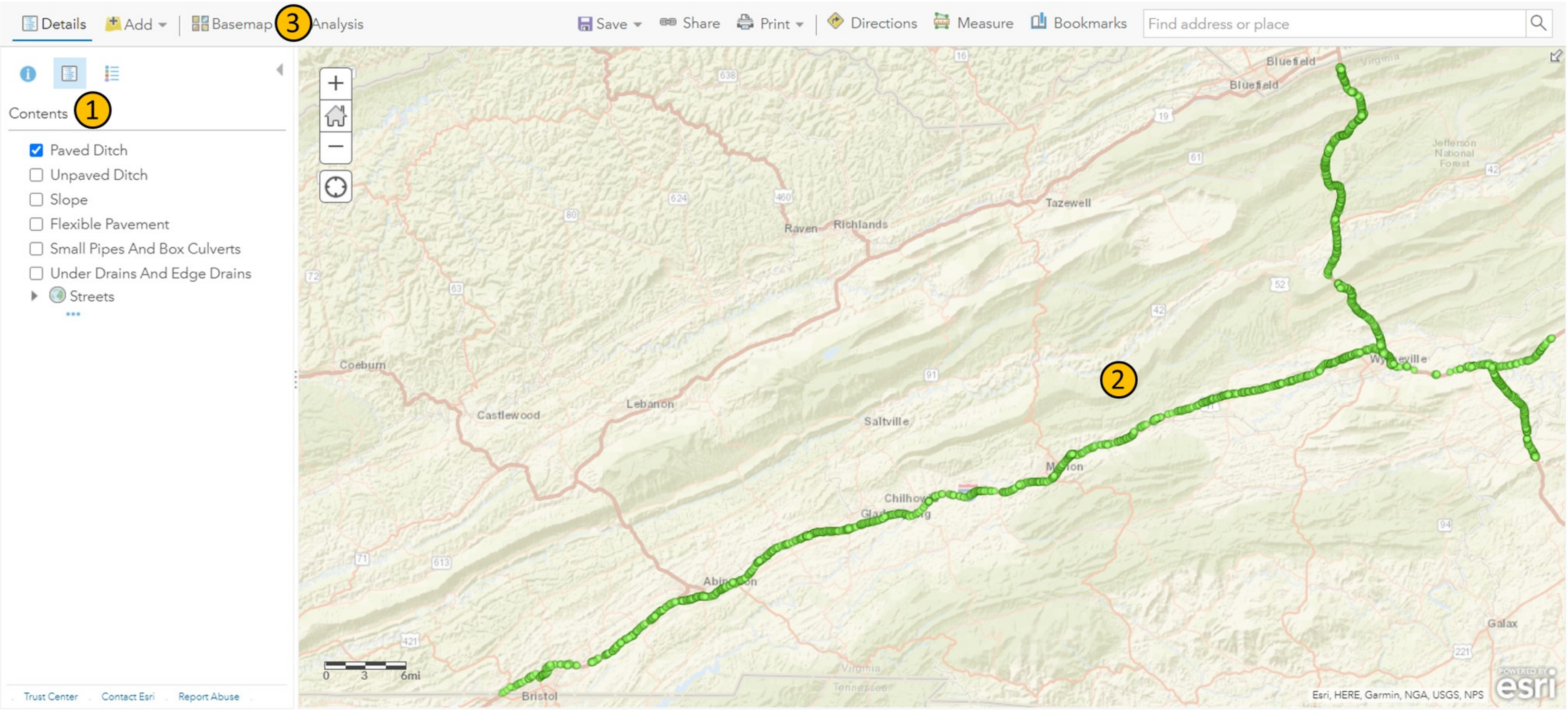
Feature Layer (hosted) by akarimza0

Created: Oct 17, 2020 Updated: Oct 17, 2020 View Count: 47

Small_Pipes_And_Box_Culverts
Point Layer

1 Creating maps for visualization purposes

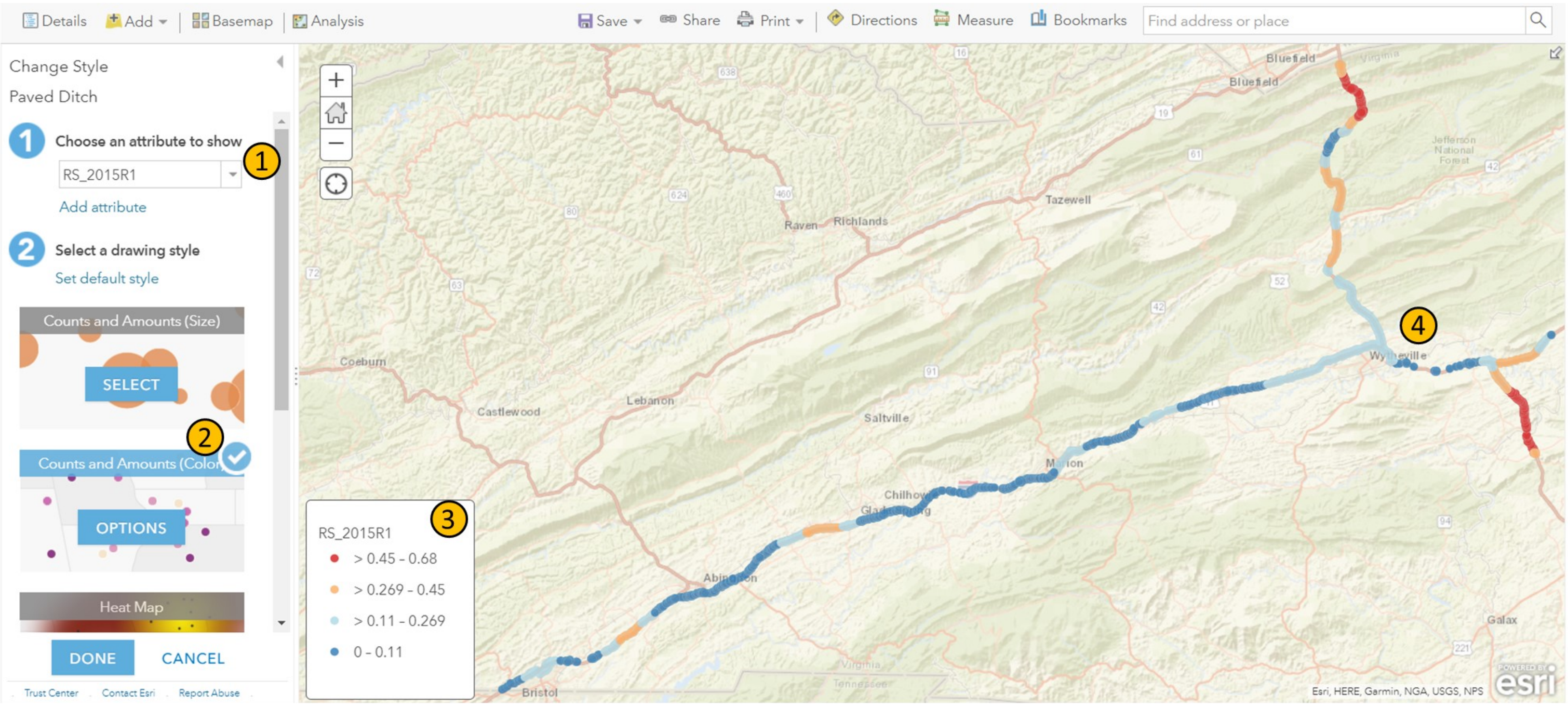
Figure A2 - Generating maps in ArcGIS Online for visualizations



☑ Paved Ditch
☐ Unpaved Ditch
☐ Slope
☐ Flexible Pavement
☐ Small Pipes And Box Culverts
☐ Under Drains And Edge Drains
▸ Streets

1 Layers of asset types (shapefiles) 2 Asset items 3 Selecting basemaps

Figure A3 - Visualizing attribute information in ArcGIS Online



Change Style
Paved Ditch

1 Choose an attribute to show
RS_2015R1
Add attribute

2 Select a drawing style
Set default style
Counts and Amounts (Size)
Counts and Amounts (Color) ✓

RS_2015R1

●	> 0.45 - 0.68
●	> 0.269 - 0.45
●	> 0.11 - 0.269
●	0 - 0.11

1 Selecting an attribute for presentation
2 Options for visualization
3 Legend
4 Visualization of datapoints with respect to the selected attribute

Figure A4 - Visualizing each attribute of asset types in ArcGIS Online