Machine learning for homogeneous grouping of pavements

Kanan Mukhtarli

A Thesis
in the Department
of
Buiding, Civil and Environmental Engineering

Submitted
For the Degree of
Master of Applied Science (Civil Engineering) at
Concordia University
Montreal, Quebec, Canada

January 2020

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Kanan Mukhtarli

Entitled: Machine learning for homogeneous grouping of pavements.

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Sciences (Civil Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

______________________________ Chair
Dr. C. Alecsandru

______________________________ External-to-program
Dr. Jia Yuan Yu (CIISE)

______________________________ Thesis Co-Supervisor(s)
Dr. L. Amador

______________________________ Thesis Co-Supervisor(s)
Dr. M. Nik-Bakht

______________________________ BCEE Examiner
Dr. F. Nasiri

______________________________ BCEE Examiner
Dr. C. Alecsandru

Approved by ____
Dr. Michelle Nokken, Graduate Program Director

____
Dr. Amir Asif, Dean, Gina Cody School of Engineering and Computer Science

Date: February 19th , 2020

Abstract

Machine learning for homogeneous grouping of pavements.

Kanan Mukhtarli

Rapid pavement deterioration is a major problem in areas with harsh weather conditions or high traffic loading. Despite many studies focused on the pavement management systems, there is not, to the date, a robust method explaining how to process large amounts of pavement data to create homogeneous groups for rehabilitation-related decision making. This thesis employs machine learning to develop an approach capable of partitioning pavement data with a close response to casual factors like traffic and weather conditions and considering its performance through international roughness index and deflections. Two different methods: K-means and Self Organizing Maps (SOM) clustering techniques were tested to understand the correlation between daily factors and pavements deterioration. The goodness of clustering was tested using extrinsic and intrinsic evaluation methods. It was concluded from the results that SOM clustering provided better results as it relies on a soft clustering method where one point can represent two clusters at the same time. Moreover, it became obvious from the methodology that including the previous year's data has very little to no effect on homogeneous groups. Techniques discussed and developed in this study can help road asset managers with decision making for the maintenance and rehabilitation of pavement. Moreover, future researchers can use the results of this study to further develop the idea of building decision support systems for pavement rehabilitation.

I would like to dedicate this thesis to my loving family and my supportive spouse.

Declaration

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

This thesis contains fewer than 22,000 words including appendices, bibliography, footnotes, and equations and has fewer than 70 figures and tables.

Kanan Mukhtarli

February 2019

Acknowledgment

This dissertation was completed in January 2020 at Concordia University, Montreal, Quebec, Canada. Firstly, I would like to express my deepest gratitude to Dr. Luis Amador and Dr. Mazdak Nik-bakht for their kind encouragement and support during the period of my MASc. study. Under their guidance, I gained invaluable experience in performing scientific research and learned how to overcome various type of research problems.

Secondly, I am very grateful to my dearest friend Reza Ghobadpour and Anuj Bisani for their consistent support in every step of the hardship.

Finally, I must express my very profound gratitude to my parents and to my spouse for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis.

Table of Contents

List of Figures

List of Tables

Nomenclature

Acronyms and Abbreviations

PMS – Pavement Management Systems

IRI – International Roughness Index

FWD – Falling Weight Deflectometer

JMP – Jump Statistical Analysis Software

AADT – Annual Average Daily Traffic

ESAL – Equivalent Single Axis Load

GIS – Geographic Information System

SOM – Self Organizing Maps

BMU – Best Matching Unit

LR – Linear Regression

VCI – Visual Condition Index

MLP-BP ANN –Multilayer Perceptron Artificial Neural Networks trained by the back-propagation algorithm

PSI – Present Serviceability Index

WIM – Weight in Motion

UTC – Coordinated Universal Time

.csv – Comma delimited Excel file

.mdb – Access database file

. gdb – A GDB file is a database file created by ArcGIS

CCC – Cubic Clustering Criterion

RASE – Root Average Squared Error

AAE – Average Absolute Error

UCR – University of Costa Rica

tanH – Hyperbolic Tangent Activation Function

SAS – Statictical Analysis Systems

Chapter 1

1. Introduction

1.1 Background

Canada's roadway system provides mobility and accessibility to thousands of individuals. The roadway network is not only important to the nation's overall economic vitality by providing for the movement of freight and commodities, but it also provides societal benefits as well (e.g., access to schools, services, and work; leisure travel; and general mobility) (Van Dam, et al., 2015). Pavements are an integral part of this roadway network. Pavements are expected to provide a smooth and durable all-weather traveling surface that benefits a range of vehicles and users (Van Dam, et al., 2015).

A pavement management system PMS is a valuable tool for handling highway transportation infrastructure (Kulkarni, 2003). The benefits of having a pavement management are well documented and include:

- Enhanced planning ability at all levels, including strategic, network, and project.
- Decision making based on observed and forecasted conditions rather than opinions.
- The ability to generate alternate scenarios for future pavement conditions based on different budget scenarios or management approaches.

Besides, improved PMS leads to better rehabilitation and reconstruction works which at the end provides a safer environment in terms of accidents and casualties. (King, 2014) investigated the effect of road roughness on traffic speed and road safety in Southern Queensland, Australia on his research. The study found a strong relationship between higher crash rates and increased pavement roughness. Crash rates involving light vehicles were 9 times more affected by increasing roughness than crashes involving heavy freight vehicles. The study recommended that traffic authorities managing rural roads need to reduce roughness to an IRI (International Roughness Index) value of 120 inches per mile (in/mile) in order to provide a safer road environment.

(Hu, 2013) developed mathematical relationships between IRI and driving comfort and safety (driving workload). The author developed threshold IRI values on-road segments at different risk levels for driving comfort and safety. They also concluded that standard trigger values of IRI for pavement maintenance are beyond the comfort and safety threshold for both car and truck drivers.

Another research conducted in Qatar concludes about the correlation between road conditions and safety. It is obvious that major surface defects may adversely affect the vehicle path or unpredictable impact on vehicle control. Features of poor to high road conditions have been introduced based on the frequency of potholes of 10 to 100-meter length. Defects were considered as deformations, pot-holes, and edge defects. (ASHGHAL, 2016)

1.2 Problem statement

There are several studies that deal with pavement management systems, deterioration modeling and data mining applied in pavement management. However, there is a need for an approach to develop homogeneous groups of pavements.

1.3 Research goal and tasks

The overall goal of this research is to develop an approach that is capable of clustering pavements into homogenous groups based on their performance, considering casual factors like climate, traffic loading and pavement condition.

The goal of the study is followed through five specific tasks:

Task 1 -To identify a method to collect, combine and analyze data gathered from the field.

Task 2 -To examine the methods used to clean and pre-process pavement condition data.

Task 3 -To cluster data into homogeneous groups.

Task 4 -To investigate the performance of the pavement within each homogeneous group.

Task 5 -To assess the results, visualize and provide suggestions accordingly.

As a result, homogeneous groups allow to build robust decision-making tools to help transportation agencies and manucipalities to take proactive measures on the deterioration of assets. These groups enhance the accuracy of prediction models and the asset management policy. With the help of clustering it is possible to better explain the pavement performance and ensure meaningful candidates for maintenance and rehabilitation. Moreover, it provides more insight in evaluation of the past M&R actions and if it was successful.

1.4 Research significance

Development of an approach to create homogeneous groups considering pavement deterioration with 2 main clustering techniques and determining which method provides better results.

1.5 Organization of the thesis

This thesis is presented in five (5) chapters as follows:

Chapter 1 - Explains the problem statement and presents the objective and structure of the thesis.

Chapter 2 - Consists of a review of concepts related to data collection used in this thesis prediction models and clustering techniques.

Chapter 3 - Contains the methods employed in collecting, analyzing and processing the data, building the prediction models for extrinsic evaluation of clustering techniques and two (2) different clustering methods.

Chapter 4 - Explores and analyzes results of the two clustering methods that are being implemented for a case study of Costa Rica pavements from the national road network.

Chapter 5 - Contains the conclusions and recommendations part that can be used for future research work.

Chapter 2

2. Literature review

2.1 Pavement management

Pavement Management Systems (PMS) require inspecting and collecting pavement related data, predicting the deterioration of pavements through performance models, and optimizing the Maintenance, Rehabilitation and Reconstruction (M&R&R) activities over a given planning horizon. Performance models are a core component of a PMS. These models are also used as an input in project design procedures (Madanat, Nakat, & Sathaye, 2005). Pavement behavior and performance is highly variable due to many factors, such as pavement structural design, climate, traffic, materials, subgrade, and construction quality. These factors contribute to changes in pavement performance that are reflected in the results of a pavement condition survey. Minimizing the impact of data variability on pavement condition data helps ensure that survey results reflect real changes in pavement performance rather than variations in data due to poor data quality (Pierce, 2014). Pavement condition data quality supports a wide variety of decisions and has direct and indirect impacts on agency processes (Xu, Bai, & Sun, 2014). Some of the major uses of pavement condition data include:

- Characterizing current condition.
- Developing models of predicted pavement deterioration.
- Projecting future conditions.
- Developing treatment recommendations, timing, and cost.
- Preparing and prioritizing annual and multi-year work programs.
- Allocating resources between regions and/or assets.

- Analyzing the impacts of various budget and treatment scenarios.
- Analyzing the performance of different pavement designs and/or materials.

In practice, the level of accuracy of pavement condition data has often proved to be difficult to achieve in network-level data collection. Deterioration is a function of environmental exposure, structural traffic loading, structural capacity and frequency and type of preventive maintenance. External factors include the number of freeze/thaw cycles, traffic loading (for pavements) and type of waste transported (for sewers) which can break and indirectly damage the road structure. Intrinsic factors include materials' type and construction methods. Maintenance factors include the type and frequency of maintenance treatments.

In many infrastructure assets, the rate of deterioration is expected to gradually increase with time. A typical deterioration curve (with maintenance activities) is given in Figure 2-1.

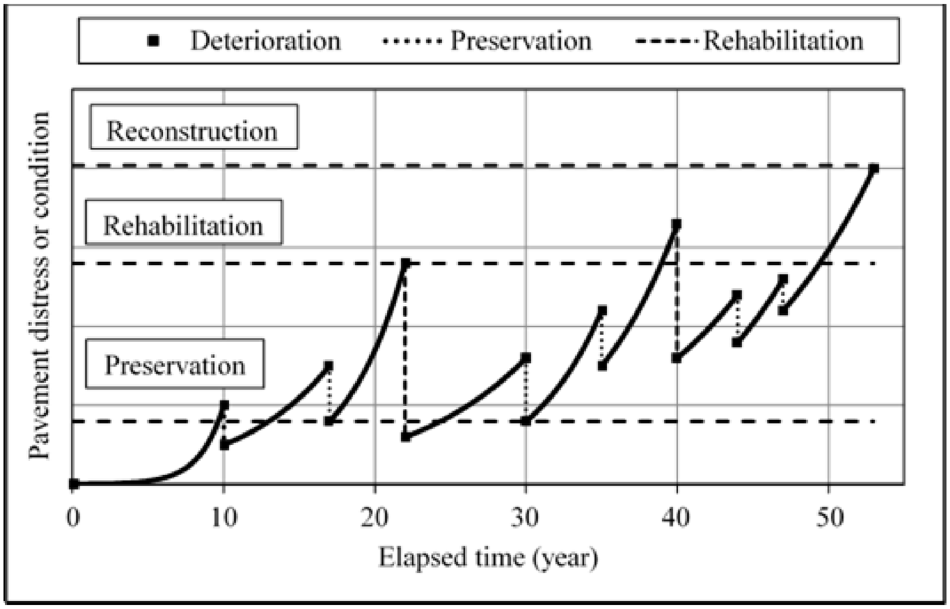


Figure 2-1. Typical pavement deterioration curve (Pavement Deterioration vs Time / Traffic) (James et all., 2014)

However, deterioration does not always occur in this way. Concave up deterioration curves are found when pavements have been designed to a higher standard than required for traffic alone, and primarily deteriorate due to weather/climate factors (Haas, 1997). Also, a single damage event may cause an asset to deteriorate very rapidly or almost instantaneously. To create a deterioration model, the factors that affect the infrastructure's condition must be quantified. For example, the causes of pavement deterioration are well-known and include environmental, traffic and structural factors. Environmental factors can include measurements of the number of freeze-thaw cycles, temperature, humidity, precipitation, water table depth; traffic factors typically include measurements of Average Annual Daily Traffic (AADT) and re-expressing it into axle-load spectra or Equivalent Single Axle Loads (ESALs). Structural factors can include pavement type, strength, and thickness. Construction and maintenance techniques also influence pavement deterioration. However, these factors can be more difficult to quantify and are not included in many deterioration models.

The remainder of this chapter explains the main concepts necessary for understanding the analysis undertaken in the upcoming chapters.

- GIS (Geographic Information System) applied in pavement management
- pavement condition measurement (Performance Indicators)
- data mining applied in pavement management systems
- prediction models
- data collection

2.2 GIS applied in pavement management

GIS is defined as "a system of computer hardware, software and procedures designed to support the capture, management, manipulation, analysis, modeling, and display of spatially referenced data for solving complex planning and management problems." (FEMA, 2003).

GIS deals with the two basic types of data, vector data types and raster data types, both of which refers the data to a geographical coordinate system (e.g., latitude/longitude or state plane coordinates) instead of the milepost or reference-point system traditionally used in transportation. We call this, geospatial data.

- Vector data is composed of discrete coordinates that can be used as points or connected to create lines and polygons.
- Raster data represent features as a matrix of cells within rows and columns in continuous space.

With the rapid increase of advanced information technology, many investigators have successfully integrated the GIS (Geographic Information System) into PMS for storing, retrieving, analyzing, and reporting information needed to support pavement-related decision making (Zhou, 2011). Such an integration system is thus called G-PMS (Lee, 1996). The main characteristic of a GIS system is that it links data/information to its geographical location, i.e., geographical coordinate system (e.g., latitude/longitude or state plane coordinates) instead of the milepost or reference-point system traditionally used in transportation, which is fundamental when integrating separate databases (Medina, 1999). GIS is also capable of rapidly retrieving data from the database and automatically generating customized maps to meet specific needs such as identifying maintenance locations. The attribute data in the pavement management system can be stored in the GIS database by location and attribute. So, a G-PMS can be enhanced with features and functionality by using a geographic information system (GIS) to perform pavement management operations, create maps of pavement condition, provide cost analysis for the recommended maintenance strategies, and long-term pavement budget programming (Zhou, 2011). Until today there is no research that adequately documents the use of GIS systems to build and map datasets in pavement management systems. In this thesis, ArcGIS was one of the most important platforms to map and join the datasets based on longitude and latitude.

2.3 Pavement condition measurement (performance indicators)

Pavement condition and performance generally can be described by four primary data (namely deflection; surface distress; serviceability; and surface friction), and two of these categories are discussed in the following sub-sections (MDOT, 2016).

2.3.1 Structural adequacy: deflections

Structural adequacy describes the load-bearing capacity of the pavement. Measuring structural adequacy involves the evaluation of deflection data within a context of pavement properties and performance demand. Deflection data collection requires specialized measurement equipment called a deflectometer. Structural adequacy is valuable in forecasting the condition of pavement under predicted loading scenarios. Early static deflection devices could measure deflection at only one point. Nowadays, Falling-weight deflectometer (FWDs) can measure deflection under the load and at a number of locations away from the load, resulting in a much larger basin. Datasets collected for this thesis included deflection values for 9 points, time, force acting on pavement, and etc., however, it is not enough to evaluate the pavement condition without calculating the deflection basin area because together they represent the condition of pavement and soil as well. Deflection basin parameters are widely used for three major applications:

- to check the structural integrity of in-service pavements
- to relate to critical pavement response
- to calculate the in-situ layer moduli of the pavements.

Figure 2-2 shows the cross-section of the deflection basin area and how the forces act during the test the formula to calculate the deflection basin area.

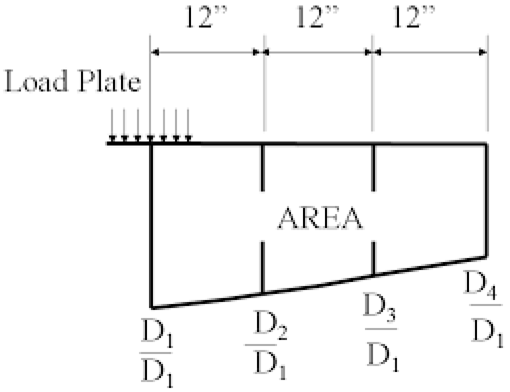


Figure 2-2. Representation of the cross-section illustrating the deflection basin area where D is the deflection measured by Geophones *(Zaniewski, Hossain, & John, 1991)*

Table 2-1. Relationship of the surface deflection - deflection basin area and pavement condition *(AASHTO, 1993)*

FWD Based Parameter		**Generalized Conclusions***
Area	**Maximum Surface Deflection (D_0)**	
Low	Low	Weak structure, strong subgrade
Low	High	Weak structure, weak subgrade
High	Low	Strong structure, strong subgrade
High	High	Strong structure, weak subgrade

* Some exceptions can be observed.

Table 2-1 illustrates the relationship between maximum deflection value and the deflection basin are to interpret the condition of the pavement. Pavement Condition was considered one of the most important influencing attributes to create homogeneous groups.

2.3.2 Surface distress and serviceability: international roughness index

One of the main internationally used pavement condition indicators in PMS is the International Roughness Index (IRI). The IRI was developed in 1986 by the World Bank and was based on the extension of the National Cooperative Highway Research Program (NCHRP) concept. The IRI was first introduced in the International Road Roughness Experiment held in Brazil (Sayers, 1995). Traffic conditions, especially ESAL, have the highest significance in contributing to the IRI value because of ESAL numbers greatly affect the changes in the surface conditions of pavements. Thus, the prediction of IRI value demanding an appropriate design of traffic.

Surface distress was traditionally assessed via visual sampling of the pavement surface. According to the final report of Federal Highway Administration (FHWA, 2018), a few metrics necessary for reporting are as follows: percentage cracking, rutting, and faulting. Considering the thresholds of the recent studies, IRI data was categorized as good, fair and poor to create homogeneous groups and to build prediction models as described in Table 2-2.

Table 2-2. International Roughness Index thresholds. *(Abudinen, Carvajal-Muñoz, & Fuentes, 2016)*

Road Condition	IRI value	Pavement Maintenance
Good	IRI ≤ 3	Routine
Fair	5 < IRI ≤ 13	Periodic
Poor	13 > IRI	Reconstruction

2.3.3 Traffic loading

Equivalent single axle load (ESAL) is a quantity that is related to pavement damage caused by a standard axle load of 80 kilonewtons (kN) (18,000 pound force (lbf) carried by a single axle with dual tires (Hajek, Selezneva, Mladenovic, & Jiang, 2005). Equivalent Single Axle Loads

(ESALs) are generally accepted as a way to represent the damage to pavement from its traffic loading. ESALs are typically calculated as a percentage of average annual daily traffic (AADT) using Equation 2-1.

$$ESALs = 182.5 \times AADT \times TP \times TF \qquad \text{Equation 2-1}$$

Where AADT is the average annual daily traffic, TP is the percentage of heavy vehicles and combinations, and TF is the truck factor.

According to (Rifai, Sigit, Correia, & Pereira, 2015), the ESAL possesses the highest importance value in the contribution towards the IRI value because the number of ESAL greatly affects the changes in the surface condition of the pavement. In his research results of homogeneous grouping stated that ESAL contribution to modeling was 26.33%, Age 13.40 %, IRI 11.54%. Thus, group contribution of traffic on the pavement deterioration model was 39.73% in total.

MOPT (Ministry of Public Works and Transport of Costa Rica) reported that over the time the traffic problem of the country has compounded as demand for cars doubled from about 700,000 to nearly 1.5 million between 2000 and 2014. This growth doesn't only affect the city planning and pollution but affects the condition of pavement as well. Traffic (ESAL) is one of the most important factors influencing the IRI of the pavement (MOPT, 2019).

2.3.4 Environmental exposure: rain

Asphalt damaged by moisture will endure distress that can lead to raveling, cracking, stripping and rutting. Water that seeps down into the structure of the pavement as a result of rain, water flow or groundwater will be absorbed by the pavement and wear away at the bond between the pavement's aggregate and the asphalt binder. This is precisely why the structure of the pavement is so important. The proper structural design allows for the pavement to eliminate as much moisture as possible in a timely manner. Thus, rain has a negative impact on pavement structure (IRI), therefore, it is necessary to include rain in the dataset before building homogeneous groups. The temperatures are mostly determined by the elevation and other geographical factors.

2.4 Data mining

Several data mining techniques have been developed over the last decade in the artificial intelligence community. Generally, the data mining techniques can be categorized into two categories: Supervised learning (which includes classification and regression – prediction); Unsupervised learning (which includes clustering and association); and reinforcement learning. (Tan, Bao, & Dong, 2007). Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in a dataset. This process helps in extracting and refining useful knowledge from large datasets. The extracted information can be used to form a prediction or classification model, identify trends and associations, refine an existing model, or provide a summary of the datasets being mined. A review by (Kohavi, 2000) states that data mining serves two goals namely insight and prediction. Insight leads to identifying patterns and trends that are useful. Prediction leads to identifying a model that gives reliable predictions based on input data.

2.4.1 Data mining background and knowledge discovery

Data mining algorithms can follow three different learning approaches: supervised, unsupervised and semi-supervised (Neelamegam & Dr.Ramaraj, 2013). There is a broad spectrum of engineering problems where computational intelligence is becoming an essential part of many advanced systems. Such problems arise in data processing, which is faced with huge data explosion, due to automatic data collection systems and the possibility for combining data from many sources over data networks (Barai, 2003). Basic steps involved in data mining and knowledge discovery are as follows and detailed explanations can be referred elsewhere (Fayyad, 1996).

1. Understanding of the application domain
2. Collection of the target dataset
3. Data cleaning and preprocessing
4. Data Warehousing
5. Selection of task-relevant data selection
6. Selection of data mining task
7. Selection of data mining tool - Artificial neural networks, Genetic Algorithms, Decision trees, Nearest neighbor method, Rule induction, Data visualization
8. Data mining - relationship identification - Classes, Clusters, Associations, Sequential patterns
9. Interpretation of results
10. Consolidation of discovered knowledge

2.4.2 Data mining techniques

Within the context of the reviewed literature, applications of data mining have been widely used in various enterprises ranging from public health-care, construction industry, food industry, finance, etc. Each field can be supported by different data mining techniques and tools which generally include Clustering, Classification, Regression, and Neural Networks.

2.4.3 Main clustering methods

Clustering is the task of identifying a finite set of data points (called clusters) to describe a dataset. This involves seeking to identify a finite set of categories and grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters (Neelamegam

& Dr.Ramaraj, 2013). Clustering can be further divided into hierarchical clustering and centroid-based clustering. K-means is the most often utilized centroid-based clustering algorithm. K-means calculates the k subsets in the data by iteratively calculating the arithmetic mean (centroid) for the k estimated clusters. With each calculation, it adjusts the clusters until they no longer change. Hierarchical and K-means clustering methods work well when clusters are well separated, but when clusters overlap, assigning each point to one cluster is problematic. In the overlap areas, there are data points from several clusters sharing the same space. In such cases, it is essentially important to use Self Organizing Maps (SOM) rather than K-means clustering if an accurate estimate of the total population in each group is desired. It is because the SOM algorithm is working based on cluster membership probabilities which can assign one overlapping data point into several clusters.

2.4.3.1 K-means clustering

K-means clustering is one of the least difficult and best-known unsupervised machine learning algorithms. The target of K-means is basic: grouping similar information and finding basic relationships. To achieve this goal, K-means looks for a fixed number (k) of clusters in a dataset. The basic algorithm is very simple (Amandeep & Navneet, 2013):

1. Select K random points as centroids.
2. Form K clusters by assigning each point to its closest centroid.
3. Recompute the centroid of each cluster until centroid does not change.

Properties of the K-means algorithm include (Amandeep & Navneet, 2013):

1. Large datasets can be efficiently processed.
2. It often terminates at local optima.
3. It works only on numeric values.

4. The shape of the identified cluster is convex.

A centroid is the imaginary or real location representing the center of the cluster (Arat, 2019). Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies K number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroid displacement as small as possible. The ‘means’ in the “K-means” refers to averaging of the data; that is, finding the centroid (Arat, 2019).

In previous studies, Sunitha (2012) used the K-means clustering technique to create homogeneous groups for the rural road network in India that behave similarly when meeting certain criteria. The attributes considered in that study are the condition of shoulders, drains, cross drainage structures, and camber, and pavement distresses, namely, potholes, crack area, and edge break, collected at every 200 m section.

Qing Li (2016) investigated the relationship between vehicle emission and IRI using the K-means clustering method. In his study, four categories A, B, C, and D, with the combination of two-step clustering modeling were clustered based on the on-road collected vehicle emissions and the pavement roughness. Results show that the relationship between the pavement roughness and vehicle emissions (traffic) is nonlinear with an R^2 value of 0.69 (Li, Qiao, & Yu, 2016).

In another study, Ting-Wu Ho (2010) was able to develop a technique to identify the location and type of cracks using image recognition. This software used K-means clustering and classification algorithms from data mining. Using this data mining methods, he was able to cluster the pixels as distress regions by scanning the images (Ting-Wu Ho, 2010).

2.4.3.2 Self Organizing Maps

The Self Organizing Maps (SOMs) technique was developed by Teuvo Kohonen (1989). The original SOM was cast as a learning process, as the original neural network algorithms. The version of SOM implemented in this thesis is a variation on K-means clustering to understand which clustering method is rather providing better results. The goal of a SOM is not only to form clusters but also to shape them in a particular layout on a cluster grid, such that points in clusters that are near each other in the SOM grid are also near each other in multivariate space (Kohonen, 1990). In classical K-means clustering, the structure of the clusters is arbitrary, but in SOMs the clusters have a grid structure. This grid structure helps to interpret the clusters in two dimensions: clusters that are close are more similar than distant clusters.

In previous studies, Senthan Mathavan (2014) used self-organizing maps for the same purpose of crack detection by reading images (like the prior studies). The main focus was on highly textured road images that makes crack detection very difficult. Road images are split into smaller rectangular cells, and a representative dataset is generated for each cell by analyzing image texture and color properties. Texture and color properties are combined with a Kohonen map to distinguish crack areas from the background. Using this technique, cracks were detected to a precision of 77% (R^2 of 0.77) (Mathavan, 2014).

The present thesis considers more detailed datasets (with features including Traffic, Temperature, Precipitation, and Deflection related attributes) to create homogenous groups using K-means, and SOM clustering methods to observe which method builds better groups that act similarly.

Also, in this thesis, for the first time, SOM will be used to create homogeneous groups as an alternative to the K-means clustering method. SOM comprises neurons in the grid, which gradually adapt to the intrinsic shape of our data. The final results allowed us to visualize datapoints and identify hidden patterns among clusters. To simply explain how SOM works, the below steps will help (Choudhury, 2019).

Step 1. Randomly position the grid's neurons in the data space.

Step 2. Select one data point, either randomly or systematically cycling through the dataset in order

Step 3. Find the neuron that is closest to the chosen data point. This neuron is called the Best Matching Unit (BMU).

Step 4. Move the BMU closer to that data point. The distance moved by the BMU is determined by a learning rate, which decreases after each iteration.

Step 5. Move the BMU's neighbors closer to that data point as well, with farther away neighbors moving less. Neighbors are identified using a radius around the BMU, and the value for this radius decreases after each iteration.

Step 6. update the learning rate and BMU radius, before repeating Steps 1 to 4. Iterate these steps until positions of neurons have been stabilized.

2.5 Pavement deterioration prediction

2.5.1 Logistic regression prediction model

The logistic regression model also referred to as a logit model, is commonly used to predict the presence or absence of an outcome with predictor variables (Powers & Xie, 2008). For example, one can use it to project a certain price, based on other factors such as availability, consumer demand, and competition. The logit transformation (Wang & Rennolls, 2005) converts a probability measurement between 0 and 1 into values in the interval (−∞, ∞). The logit transformation is defined as (Powers & Xie, 2008):

$$Logit(p) = ln[\frac{p}{1-p}] \qquad \text{Equation 2-2}$$

where Logit (p) = the natural log of the odds, ln = the natural logarithm, and p = the probability of success.

More specifically, logistic regressions' main focus is to help uncover the exact relationship between two (or more) variables in a given dataset. Like all regression analyses, the logistic regression is a predictive analysis (SS, 2019). Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables (SS, 2019). Recent studies implemented regression models to predict the pavement deterioration has a range of R^2 of 0.68 and 0.76.

In this thesis logistic models are used to be able to predict the IRI to understand the pavement deterioration trend and verify the goodness of the clustering.

(Dae, Chi, & Kim, 2018) conducted a study using a logistic regression model to predict network-level sections that would be selected for pavement-preservation projects. A large number of samples were used to develop a logistic regression model. The model results indicated that all predictors, except the truck AADT, were significant at the 95% confidence level. These predictors included the total AADT, speed limit, condition score, and changes in condition score since last year.

In another recent study, (Heidari, Najafi, & Alavi, 2018) conducted similar research where they considered Traffic (AADT), Pavement Condition, Precipitation, and Road qualify as four (4) main categories. The study concluded that ESAL and Pavement thickness has a strong effect on the pavement deterioration model. As a result, Linear Regression (LR) and Artificial Neural Networks (ANNs) were able to classify the 82% to 89% of the pavement condition precisely based on the 185 road segments covering the length of 50 km.

2.5.2 Neural networks

The use of artificial neural networks (ANNs) has tremendously increased in several areas of engineering over the last three decades. In the literature of pavement management, neural networks have been used under seven different categories: (1) prediction of pavement condition and performance; (2) pavement management and maintenance strategies; (3) pavement distress forecasting; (4) structural evaluation of pavement systems; (5) pavement image analysis and classification; (6) pavement materials modeling; and (7) other miscellaneous transportation infrastructure applications (Ceylan, Bayrak, & Gopalakrishnan, 2014).

Neural networks are data processing computational tools that are capable of solving complex nonlinear relations. Like humans, they have the flexibility to learn from examples by means of interconnected elements, namely neurons. Neural networks have been found to be very powerful and versatile computational tools for determining and predicting the future condition and performance of the existing pavement systems (Panerati, Schnellmann, Patience, Beltrame, & Patience, 2019).

In one of the studies, Attoh-Okine (1994) applied a back-propagation type ANN to develop a pavement roughness progression model. A neural network model was trained using synthetically generated roughness data. The ANN prediction results were found to be more satisfactory when the pavement condition database considered was large enough. However, it was reported that the ANN model may not produce as good results with real datasets as it gave for the simulated dataset (Attoh-Okine, 1994).

Van der Gryp et al. (1998) introduced a one-hidden layer feed-forward ANN model to estimate the overall pavement condition based on the visual condition index (VCI) that ranges from 0 to 10, where 0 indicates worst and 10 indicates excellent pavement surface condition. The reported simulations made it difficult to conclude on the effectiveness of the ANN (Gryp, Bredenhann, Henderson, & Rohde, 1998).

Lin (2003) developed a multilayer perceptron ANN trained by the back-propagation algorithm MLP-BP ANN (14 input nodes, 2 hidden layers with 6 nodes each, and one output node) to predict IRI based on pavement distresses (Lin, 2003).

In ANN, the activation function defines the output of the node based on the input or set of inputs. In modern computer circuits, this function is either one (1) or zero (0). The activation function (also called a transfer function), can be linear or nonlinear function. There are different types of activation functions (Sibi, 2005). The activation function f(.) is also known as a squashing function. There are various types of activation functions such as; Piecewise Linear Function (Linear Function), Hyperbolic Tangent Function, Gaussian, etc. Sigmoid and hyperbolic tangent is the most widely used because their differentiable nature makes them compatible with backpropagation algorithm (BP) (Hussein, 2015).

Equation of Piecewise Linear Function is shown below in Equation 2-3. (Hussein, 2015)

$$g(net) = \begin{cases} 1: & if\ net \geq \frac{1}{2} \\ net: & if\ \frac{1}{2} > net > -\frac{1}{2} \\ 0: & if\ net \leq -\frac{1}{2} \end{cases} \qquad \text{Equation 2-3}$$

By varying the domain of the net input values over which the above function exhibits linear characteristics, the two extremes of this activation function can be derived (Witten & Frank, 2000). The one extreme happens when the domain of the net input values for which this function is linear is infinite; then an activation function that is linear everywhere is being dealt with. The other extreme occurs when the domain of the net values for which activation function is linear shrinks to zero; in that case, threshold activation function comes into play (Sibi, 2005).

In many applications, hyperbolic tangent function (tanH) is used as the activation function, so that the output y will be in the range from -1 to +1 rather than 0 to +1 (Özkan, 2003). The hyperbolic tangent function is defined as the ratio between the hyperbolic sine and the cosine functions or expanded as the ratio of the half difference and the half sum of two exponential functions in the points x and –x as follows:

$$tanh(x) = \frac{sinh\ (x)}{cosh\ (x)} = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \qquad \text{Equation 2-4}$$

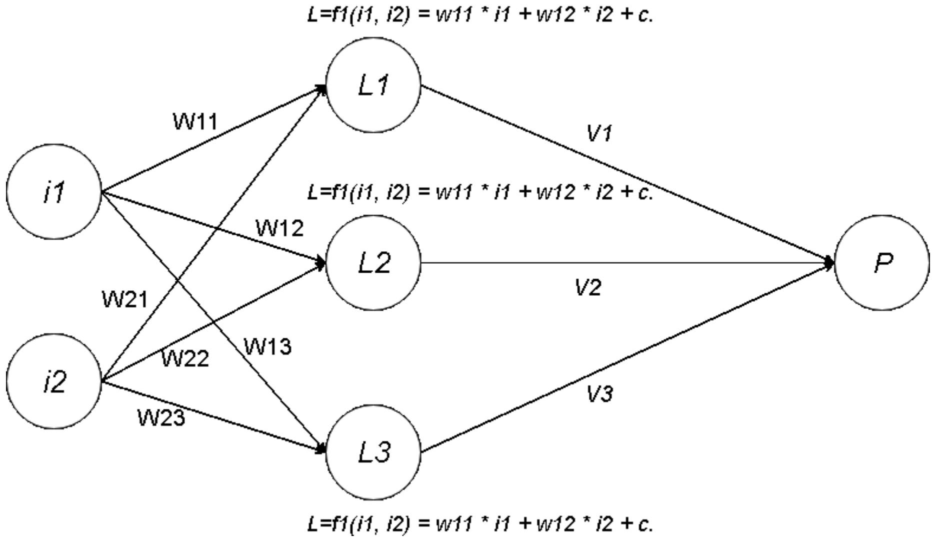

L=f1(i1, i2) = w11 * i1 + w12 * i2 + c.
L1
W11
V1
i1
L=f1(i1, i2) = w11 * i1 + w12 * i2 + c.
W12
L2
P
W21
V2
W13
W22
i2
V3
W23
L3
L=f1(i1, i2) = w11 * i1 + w12 * i2 + c.

Figure 2-3. Graphical Representation of a neural network with two inputs (i1 and i2), one hidden layer (w11 through w23), three hidden nodes (L1 through L3), and one output (P)

The results of the neural network models are interpreted and compared to the other models using R^2, RASE (Root Average Square Error) and AAE (Average Absolute Error) values. The RASE is a quadratic scoring rule which measures the average magnitude of the error. Since the errors are squared before they are averaged, the RASE gives a relatively high weight to large errors. This means the RASE is most useful when large errors are particularly undesirable. RASE is calculated with the below formula:

$$RASE = \sqrt{\frac{SSE}{n}} \qquad \text{Equation 2-5}$$

Where:

SSE – Sum of Squared Errors.

N – The number of observations.

Inevitably, The RASE will always be larger or equal to the AAE and the greater difference between them, the greater the variance in the individual errors in the sample. Both the AAE and RASE can range from 0 to ∞. The model is assumed better with lower values of both RASE and AAE.

2.5.3 Partitioning – decision tree model

A decision tree consists of two types of nodes: (1) decision nodes and (2) chance nodes, and several alternatives, shown as branches at each of these nodes. The analysis of a decision tree requires the estimation of probabilities and costs of different outcomes at each chance node. The costs would include construction cost, maintenance cost, and user cost associated with a given PSI (Present Serviceability Index) level.

There are many variations of partitioning like decision trees, CARTTM, CHAIDTM, C4.5, C5, etc (Holdaway, 2014). Decision trees are commonly used due to advantages such as the following.

- it is good for exploring relationships without having a good prior model,
- it handles large problems easily, and
- the results are interpretable.

The analysis of a decision tree requires the estimation of probabilities and costs of different outcomes at each chance node. The costs would include construction cost, maintenance cost. The goodness of partitioning is measured with R^2 value where higher value leads to a better result.

On the other hand, the purpose of using partitioning models in this thesis is mainly for confirming the goodness of clustering methods, rather than building a prediction model. In previous studies, partitioning has not been used for extrinsic evaluation of K-means and SOM clustering

2.6 Literature gaps

Previously several studies were conducted, applying data mining for PMS of different regions. However, each study had different limitations such as the inadequacy of data, merging the datasets, clustering techniques, etc. In his data-mining study, Lea (2004) emphasized merging datasets as the main challenge. Handling such large datasets still remains as the main issue to merge and process (Lea, 2004).

More recently, in a study which is more similar to the present thesis, Sunitha (2012) conducted research on low-volume rural roads with different attributes to construct pavement deterioration models. The main drawback of this study was the availability and processing of the datasets. While processing the data, large portions of the dataset were filtered due to inconsistencies, as the location of data collection didn’t match and makes it impossible to build the homogeneous groups. (V. Sunitha, 2012). In the present thesis, a method of merging such datasets is introduced and applied, which creates further research opportunities.

In previous research works, the size of datasets used was not too large. Thus, as a major drawback rules or identified patterns generated based on a small dataset was not reliable. In this thesis, relatively larger datasets of country-wide road networks are analyzed by investigating the relationships of IRI, deflection and environmental factors with pavement deterioration which has not yet been done.

Additionally, all of the closely related studies mentioned above only used K-means clustering as a base technique to form homogeneous groups, however, in this thesis a new method of SOM clustering is introduced to evaluate such data. Moreover, two methods of evaluation of homogeneous groups were introduced in the study.

Another major gap in all of the previous studies related to pavement management systems was that the conclusions were drawn without checking the quality of clustering techniques. This was the case for both of the very similar studies conducted by (V. Sunitha, 2012) and (Wang K. &., 2010) where they analyzed the hierarchical and gray clustering methods to learn the pavement deterioration process.

Chapter 3

3. Methodology

3.1 Introduction

This chapter explains the method proposed to merge and create datasets in order to define the homogeneous groups, and evaluate them. The flow of this chapter is given in the form of a process tree in Figure 3-1. Collection of data is explained in Section 3.2 following with the methods of cleaning and merging of the separate datasets in Section 3.3 and the methods in which the data is analyzed are explained in Sections 3.4-3.5.

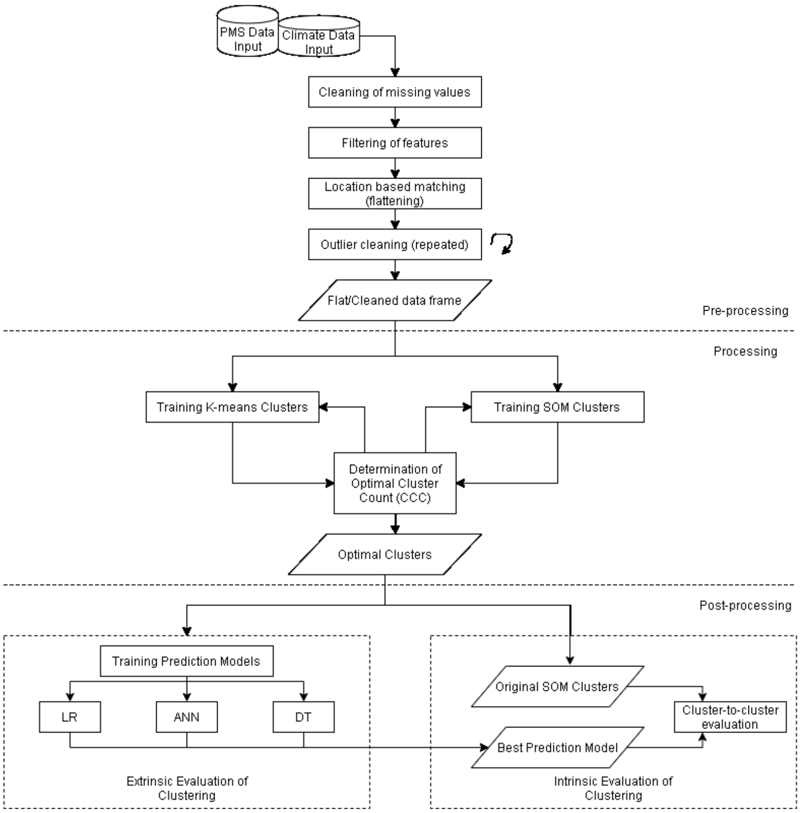

PMS Data Input
Climate Data Input
Cleaning of missing values
Filtering of features
Location based matching (flattening)
Outlier cleaning (repeated)
Flat/Cleaned data frame
Pre-processing
Processing
Training K-means Clusters
Training SOM Clusters
Determination of Optimal Cluster Count (CCC)
Optimal Clusters
Post-processing
Training Prediction Models
LR
ANN
DT
Extrinsic Evaluation of Clustering
Original SOM Clusters
Cluster-to-cluster evaluation
Best Prediction Model
Intrinsic Evaluation of Clustering

Figure 3-1. Graphical representation of the flow chart of methodology.

3.2 Data collection

The data collected from the field should properly be used to determine the impact of the influencing daily factor on the future condition of the pavement.

3.2.1 Traffic

For the pavement management purposes, the preferred practice to collect the traffic data is to use weigh-in-motion (WIM) scales at non-enforcement locations to measure actual pavement loads as opposed to legally enforced loads. In order to accurately predict future pavement performance, engineers need to know how heavy are loads being applied to a pavement. The equivalent load most commonly used in pavement design in the U.S. is the 18,000 lb. (80 kN) equivalent single axle load (ESAL). In order to convert the AADT data to ESAL's necessary assumptions needs to be taken into consideration. Equation 3-1 is used to convert the AADT data to ESAL with the assumptions given in Table 3-1.

$$ESALi = (AADTi) * (Fd) * (Gjt) * (fi) * 365 \qquad \text{Equation 3-1}$$

Table 3-1. Important assumptions for the ESAL calculation

Assumptions for ESAL calculation

No. of Years to Project Traffic (yrs)	1
Directional Distribution Factor (%)	50
Design Lane Distribution Factor (%)	100
Growth Rate (%)	2
Truck Factor (ESALs/Truck)	1.7

3.2.2 International roughness index

The use of high-speed longitudinal pavement profile equipment has become the generally accepted industry standard for measuring pavement roughness. The measurement technique is based on using an inertial profiler, which measures the change in longitudinal profile in the wheel paths at or near the speed limit. Inertial profiler's work principle has been shown in Figure 3-2 and usually, the IRI data is collected for every 100 meters.

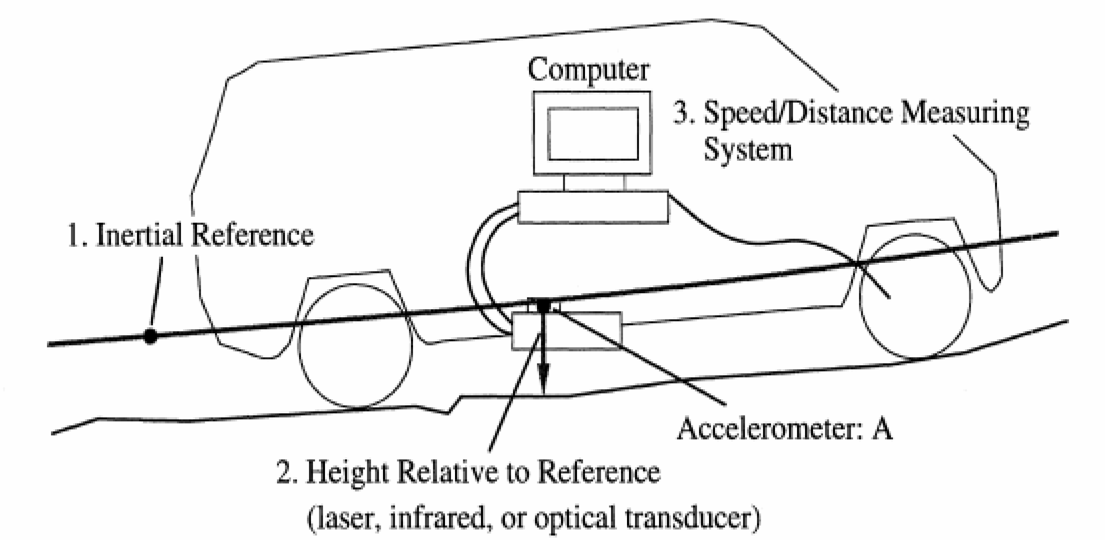

Computer
3. Speed/Distance Measuring System
1. Inertial Reference
Accelerometer: A
2. Height Relative to Reference (laser, infrared, or optical transducer)

Figure 3-2. Inertial Profiler's working principle (Perera, 1999)

Since the process is automatic operator only needs to calibrate the location and time/date at the beginning of the test. As a result, the example dataset used in this thesis includes the attributes necessary for the IRI interpretation as Longitude, Latitude, Altitude, Station ID, beginning (Starting location) of the section, End location of the section, Left wheel IRI, Center IRI, and Right Wheel IRI. Since the data is exported in an organized form the only necessary step to perform is merging the dataset, cleaning the outliers and empty values where the measurements failed.

3.2.3 Deflection – structural integrity

Deflection measurements are used to measure the response of a pavement structure to a known applied load. Technological advancements made it possible to collect even more data separately in one run including tables like Comments, Drops, Histories, Remarks, Sessions, Stations, Timing, Transducers, Version. Additionally, Modern FWD data recorders include GPS

Status, UTC (Coordinated Universal Time), Longitude, Latitude, Height, Satellite ID, Slab ID, Surface, and Air temperatures and FWD data is usually collected every 200 meters. As a result of inconsistent or faulty measurements at some points data comes noisy so that during the cleaning of the data a number of rows either need to be patched or cleaned for further processing. FWD data originally comes as .mdb Access files and later needs to be converted into acceptable .csv format.

Taking into consideration all of the recently published sources AASHTO 93 is used widely since it contains the latest reliable correction factor graphs which are shown in Figure 3-3 and Figure 3-4. "The Handbook of Highway Engineering" (Edited by T.F. Fwa, Graphs at Page 11.8 - September 2005). Mentioned graphs were converted to formulas in order to be able to proceed further on the excel dataset for the correction of the deflection value.

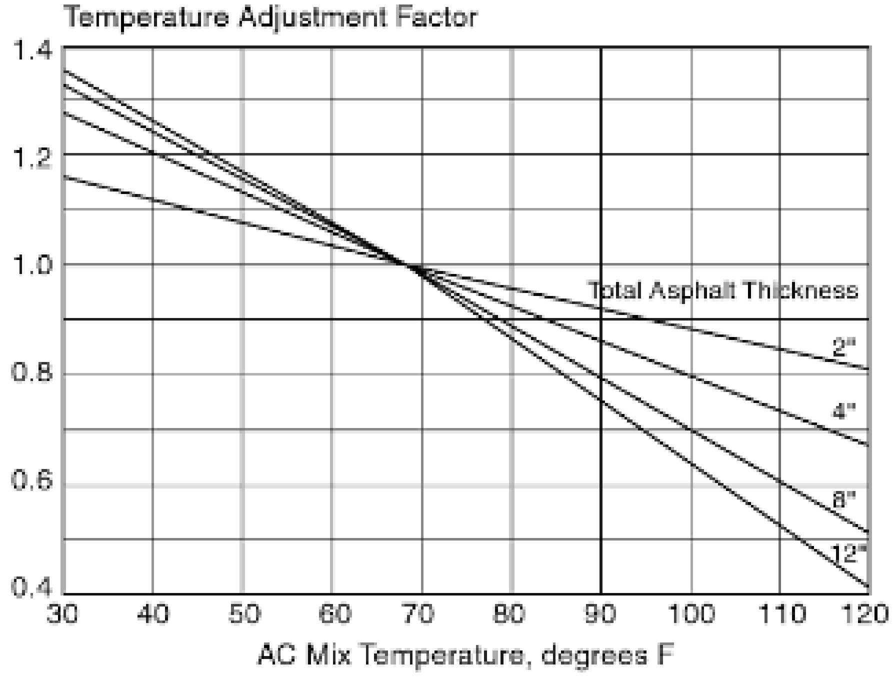


Figure 3-3. Adjustment to the deflection value - for the pavement with granular or asphalt-treated base. *(AASHTO, 1993)*

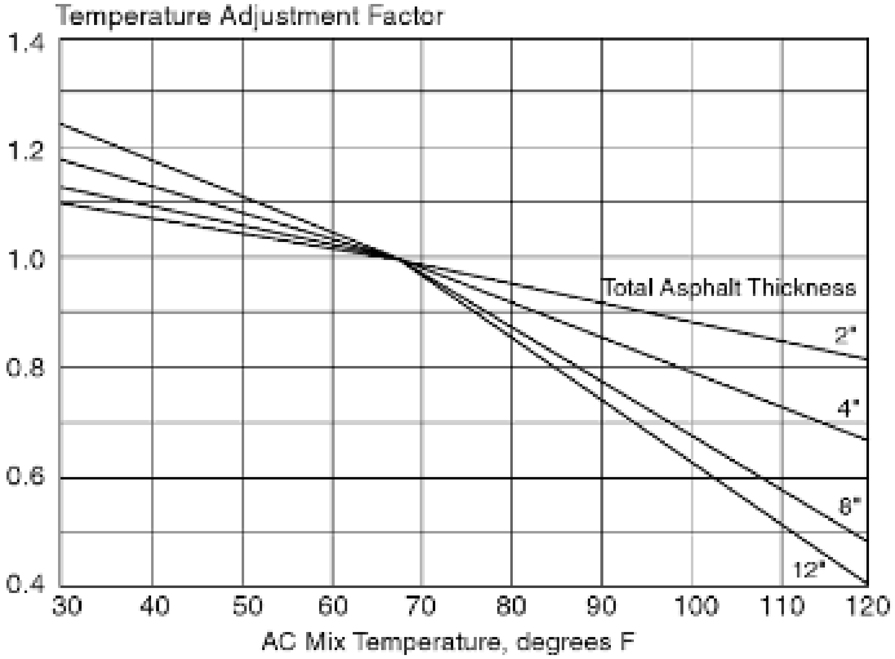


Figure 3-4. Adjustment to the deflection value - for the pavement with cement or pozzolanic-treated base. *(AASHTO, 1993)*

Equations have been developed for the soil type that uses granular or asphalt treated base shown in Table 3-2 using Figure 3-3 and Figure 3-4.

Table 3-2. Equations of a deflection correction factor based on pavement thickness

Total pavement thickness	**Correction factor equation**
12inch thickness	d_0 = -0.0105x +1.679
8inch thickness	d_0 = -0.009x + 1.599
4inch thickness	d_0 = -0.0066x + 1.473
2inch thickness	d_0 = -0.004x + 1.29

Where:

d_0 - correction factor

x – the temperature in Fahrenheit measured from the site.

A correction factor must be calculated for every possible pavement thickness. After having the correction factor calculated, maximum deflection is multiplied by the correction factor and the product is used for clustering. Corrections are usually made for the measurements having a pavement temperature of up to 40C. After the correction around 3000-6000 rows of data that goes beyond the temperature, threshold is selected and cleaned because such data creates instability while clustering and alters the results by decreasing the value of R^2 (accuracy). Then, the deflection basin area is calculated using the formula given in Equation 3-2.

$$AREA = \frac{6*(D_0+2D_1+2D_2+D_3)}{D_0} \qquad \text{(Equation 3-2)}$$

Where:

AREA - equals the FWD AREA Parameter. Expressed in units of length (usually inches or mm).

D_0 - equals surface deflection at the test load center

D_1 - equals surface deflection at 12 inches from the test load center

D_2 - equals surface deflection at 24 inches from the test load center

D_3 - equals surface deflection at 36 inches from the test load center

The deflection basin area is also considered the main factor affecting to clustering of the homogeneous groups.

3.2.4 Precipitation

Rainfall data is generally collected using electronic data loggers that measure the rainfall in 0.01- inch increments every 15 minutes using either a tipping-bucket rain gage or a collection well gage. Twenty-four-hour rainfall totals are tabulated and presented. A 24-hour period extends from just past midnight of the previous day to midnight of the current day. Resulting attributes of the dataset includes the regions of the country and yearly rainfall amount in “mm”. In order to merge the rainfall data, ArcMap is used to assign it to the regions and then to the roads having both the deflection and IRI data to avoid the inconsistencies based on their longitude and latitude coordinates.

3.3 Data pre-processing

In the beginning, datasets of all four daily factors discussed in Section 3.2 are received as distinct datasets which makes it impossible to start evaluating the data and hidden patterns among attributes. For this reason, this thesis includes several different dataset merging techniques to create a single table for further analyses. In this study, along with the command terminal and JMP (software), ArcMap (software) is one of the most essential tools used to merge and create datasets.

3.3.1 Using command terminal to merge datasets

It is often common to have hundreds of separate .csv files containing pavement data collected from the field. It is, therefore, necessary to merge all such datasets into one table using the most straightforward method. Windows' command terminal was used to merge all .csv files into one table in order to make the location-based joining easier. In order to perform the operation, it is important to navigate to the command terminal and then to the folder where the separate .csv formatted datasets exist. Following the command "/copy *.csv combined.csv" is the only line of code needed to create a new file named "combined.csv" in the specified folder.

3.3.2 Using ArcMap to merge datasets

Prior to importing the dataset, it is necessary to make sure that the Geographic coordinate systems are configured. In this thesis, data frame properties are configured with the Geographic Coordinate System of WGS 1984.

Mapping – After creating tables user must fit the longitude and latitudes of the table on the X and Y-axis. This can be done by displaying XY Data (X - Field is longitude, Y- Field is latitude). Repeating the same steps for the second dataset will result in two different maps. In order to be able to join them these tables must be saved as shapefiles. After the geodatabase (.gdb) files have been created user can match and join the attributes easily.

Spatial Join – Spatial join involves matching rows from the "Join Features" to the "Target Features" based on their relative spatial locations. Joining large datasets based on their locations is a very complicated task on other platforms, however, in ArcMap it is straightforward. "Match Options" is available depending on the data structure so that the user can select and input the necessary search radius (meters, kilometers, etc.) to start the spatial join.

3.3.3 Using JMP to merge datasets

JMP was created by Statistical Analysis Systems (SAS) in 1989 and today it is one of the most powerful statistical analysis & data mining software. It is built on JSL Scripting language which is easy to understand and implement any analysis. The latest version of JMP is capable of translating the resultant scripts into any other platform (Java, Python, R, etc.) to make it easily accessible through the formula depot. There are 3 methods to merge and split datasets:

1. By matching Columns
2. By Row number
3. Cartesian Join (1 row from the 1st data table to every other row of the 2nd data table)

In this thesis, latitude and longitude columns were matched to merge different datasets into one table in order to proceed to analysis. As a first step of the joining process, both datasets are merged and then multiple values are dropped from the final data table.

3.3.4 Dataset cleaning

The method of outlier cleaning used in this thesis is the "Multivariate Robust Fit Outliers" that is utilized to examine the relationships between multiple variables. This outlier analysis provides three different methods to calculate the distances to identify outliers.

As the first method, it calculates the Mahalanobis distances from each point to the center of the multivariate normal distribution. Mahalanobis distances method has a simple basis, the greater the distance from the center, the higher the probability that it is an outlier. This method is used to clean the data because the outliers are visually represented. The formula used to calculate Mahalanobis distance is as shown in Equation 3-3 (SAS, Distance Measures, 2019).

$$M_i = \sqrt{(Y_i - \bar{Y})'S^{-1} \times (Y_i - \bar{Y})} \qquad \text{(Equation 3-3)}$$

In which M_i is the Mahalanobis distance for the i^{th} observation, Y_i is is the data for the i^{th} row, $\bar{Y}$ is the row of means, S is the estimated covariance matrix for the data.

T^2 method is another method of describing the Mahalanobis method by simply squaring it and the formula to calculate T^2 distances is as shown in Equation 3-4 (SAS, Distance Measures, 2019).

$$UCL_{T^2} = \frac{(n-1)^2}{n} \times \beta_{[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}]} = (UCL_{Mahalanobis})^2 \quad \text{(Equation 3-4)}$$

In this formula:

n = number of observations

p = number of variables (columns)

$\beta_{[1-\alpha;\frac{p}{2};\frac{n-p-1}{2}]}$ = $(1-\alpha)^{th}$ quantile of a Beta $(\frac{p}{2};\frac{n-p-1}{2})$ distribution and the α (alpha) value is the correlation confidence intervals and it can be edited as required.

"Jackknife distance" was the main method considered in this thesis under Multivariate Robust Fit Outliers technique. It is a better form of the Mahalanobis distances. The distance for each observation is calculated with estimates of the mean, standard deviation, and correlation matrix that do not include the observation itself. Equation 3-5 illustrates the formula to calculate the Jackknife distances (SAS, Distance Measures, 2019).

$$J_i = \sqrt{\frac{(n-2)n^2}{(n-1)^3} \times \frac{M_i^2}{1-\frac{nM_i^2}{(n-1)^2}}} \qquad \text{(Equation 3-5)}$$

Where:

J_i = Jackknife distance for the ith observation

n = number of observations

M_i = Mahalanobis distance for the ith observation

As it is obvious from Figure 3-5 that two points shown as an outlier in Mahalanobis distances are actually not considered an outlier according to Jackknife distances. In this case, it is necessary to check the data with Jackknife distances before removing any rows from the data table. This limits the list of columns to only those that contain outliers. The jackknife method works by repeatedly re-computing the summary statistic leaving out one data item at a time from the dataset.

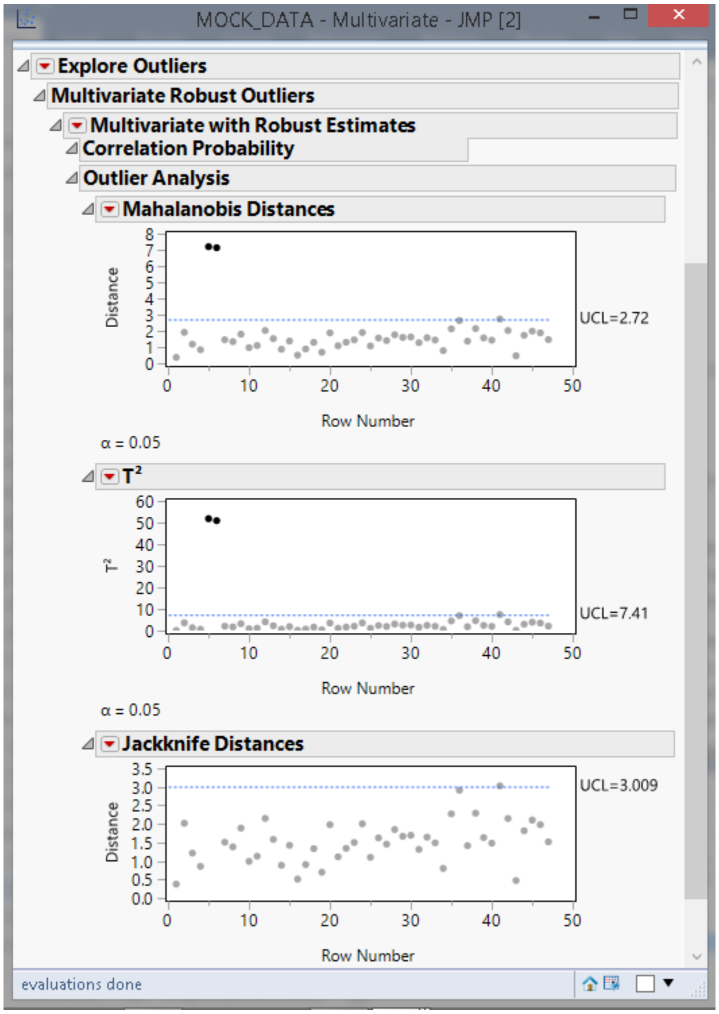


Figure 3-5. Comparison of the results of Mahalanobis, T^2 and Jackknife distances gathered from an example dataset available at *(Mukhtarli, 2019)*.

3.4 Clustering methods

3.4.1 K-means clustering

The K-means approach is a special case of a general approach called the EM (Expectation-Maximization) algorithm. The K-means method is intended to be used with larger data tables, from approximately 200 to 100,000 observations (SAS, 2020). When variables of datasets do not share a common measurement scale, to prevent one variable dominating the clustering process columns must be scaled individually. Johnson transformation is used to bring the far values closer to the clusters while scaling and spreading them around the cluster centroid. More information about the Johnson transformation can be accessed from the following source (Unistat, 2019). This is especially useful when the data is widely spread over the X and Y axes and when some data points stay outside the clusters and act as outliers. This behavior impacts the CCC (Cubic Clustering Criterion) value which is crucial in defining the optimal number of clusters.

The CCC is one of the methods to estimate the optimal number of clusters using Ward's minimum variance method. The idea of CCC is to compare the R^2 obtained from a given set of clusters with the R^2 one would get by clustering a uniformly distributed set of points in multi-dimensional space (Dickey, 2015). The performance of the CCC is evaluated by Monte Carlo methods. Empirical formula to calculate the CCC is given below in Equation 3-6 (SAS, 2019).

$$CCC = \ln\left(\frac{1-E(R^2)}{1-R^2}\right)\frac{\sqrt{\frac{np}{2}}}{(0.001+E(R^2))^{1.2}} \qquad \text{(Equation 3-6)}$$

Where: $E(R^2)$ is the expected R^2, p is the dimensionality of the between-cluster variation and n is the number of samples. More detailed information about the Cubic Clustering Criterion and its calculation can be obtained from the references of SAS (SAS, 2019).

The optimal number of clusters can be either set to a certain number or can be defined as a range to be examined. In this thesis, a range is given and CCC value is calculated for each of the options in the defined range and the optimal number of clusters is set based on CCC calculations. The higher the CCC value, the better the clustering is. Hence the problem of finding the best number of clusters can be formulated as an optimization problem for CCC. Negative values of the CCC with comparatively large absolute values, e.g. -30, may be the result of having outliers in the dataset. Outliers generally should be removed before clustering. If all values of the CCC are negative, the distribution is probably unimodal or long-tailed. If the CCC increases continually as the number of clusters increases, the distribution may be grainy, or the data may have been excessively rounded or recorded with just a few digits.

3.4.2 SOM clustering

Another clustering method used in this study to generate homogeneous groups is the SOM. Just like K-means clustering SOM also takes CCC as the criterion to choose the optimal number of clusters and higher the CCC better the result.

Starting the SOM analysis is following similar procedure as K-means where the traffic, precipitation, deflection, deflection basin area, and the IRI values are the features.

After selecting the features that generate the homogeneous groups of pavements, the SOM grid is specified. A grid is a two-dimensional plane that includes centroids called nodes and the dataset is clustered around these centroids. Moreover, Self-Organizing Map uses competitive learning as opposed to error-correction learning to adjust it weights. It means during each iteration only one node (centroid) is activated to respond to input features and calculate a new position for itself. Therefore, the system is called self-organized maps where the data points are appointed to centroids in each iteration until the best result is achieved. Since in the literature there is no particular method for identifying the grid configuration, map size is selected as 3x3 (9 nodes in total) after several trial-and-error. This configuration was identified as the best option for the currently examined datasets.

3.5 Cluster evaluation.

The prediction of the pavement condition is important in any pavement management system. Several pavement prediction models are generally available, including regression, Markov prediction models, stochastic models, etc. However, in this thesis, regression models are used as an extrinsic evaluation method to test the goodness of the clustering of homogeneous groups.

Regression analysis is a statistical method used to describe the relationship between two variables and to predict one variable from another.

3.5.1 Logistic regression model

The general goal of the logistic regression is to find the best fitting model to describe the relationship between the characteristics of interest (dependent variable) and a set of independent (predictor or explanatory) variables.

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values. In order to run the logistic regression value that is going to be predicted must be nominal (Class). The nominal value often referred to as the categorical variables contains a finite number of categories or groups. Such data may not have a logical order. Values to be predicted are also called response variables. In order to interpret the model results, user must achieve as much higher R^2 values as possible and compare with other prediction models. An interpretation of a typical logistic regression model is explained below with the given sample graph. To interpret the results, it is possible to plot the results of logistic regression in binary fitted line plots.

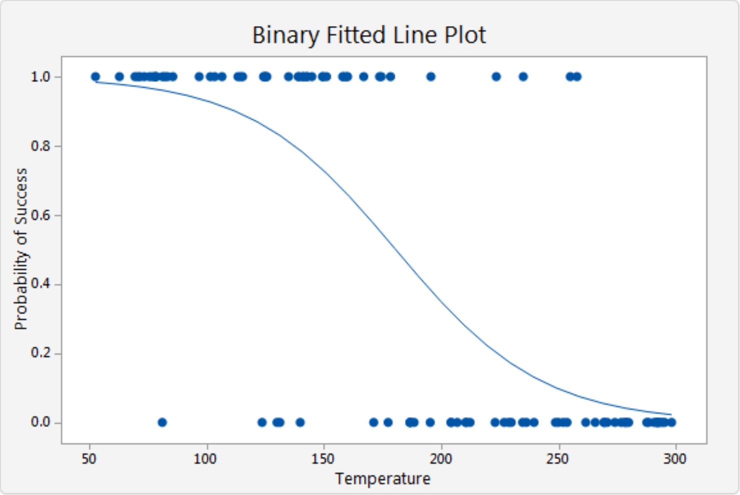


Figure 3-6. An example of binary fitted line plot gathered from a sample dataset available at *(Mukhtarli, 2019)*

In Figure 3-6, the response value of 1 on the y-axis represents successful modeling (Example dataset is available at (Mukhtarli, 2019). The plot shows that the probability of success decreases as the temperature increases. When the temperature in the data is near 50, the slope of the line is not very steep, which indicates that the probability decreases slowly as temperature increases. The line is steeper in the middle portion of the temperature data, which indicates that a change in temperature of one degree has a larger effect in this range. When the probability of success approaches zero at the highest end of the temperature range, the line flattens again.

In every model presented in this thesis the higher the R^2, the better the model fits the data. The R^2 value is always varying between 0% and 100%. In the present thesis, the goal of creating a logistic regression model is to identify the goodness of SOM and K-means clustering. It also provides an insight into whether including the previous years' data in the analysis improves the accuracy of the models or having no impact on the results.

3.5.2 Neural networks

In the current study, various combinations of neural network models are designed to accomplish the evaluation of the quality of the homogeneous groups. Various models can be specified by the network topology, node characteristics, and training or learning rules.

At the beginning of the analysis, the default number of hidden nodes is taken as three to generate a hidden layer structure. Then during the evaluation process activation functions such as hyperbolic tangent (tanH), Gaussian and Linear functions or Boosting method are selected along with several hidden layer options. In literature, there is no closed solution offered for making the selection on the structure or activation function. Thus, it is done based on a trial-and-error method.

Boosting is a method for improving the performance of the learning algorithms. The default boosting step is usually one (1) where it means the network will only be filtered once to improve the learning algorithm. Hence, better results will be achieved by higher numbers of boosting until the accuracy of the model reaches a constant value.

The tanH (hyperbolic tangent function) is a sigmoid function and the graphical representation is given in Figure 3-7. It transforms the values to stay between -1 and 1 and is the centered and scaled version of the logistic function. TanH function is calculated using the Equation 3-7 (SAS, 2019).

$$TanH = \frac{(e2x - 1)}{(e2x + 1)} \qquad \text{(Equation 3-7)}$$

Where, x is a linear combination of the X variables.

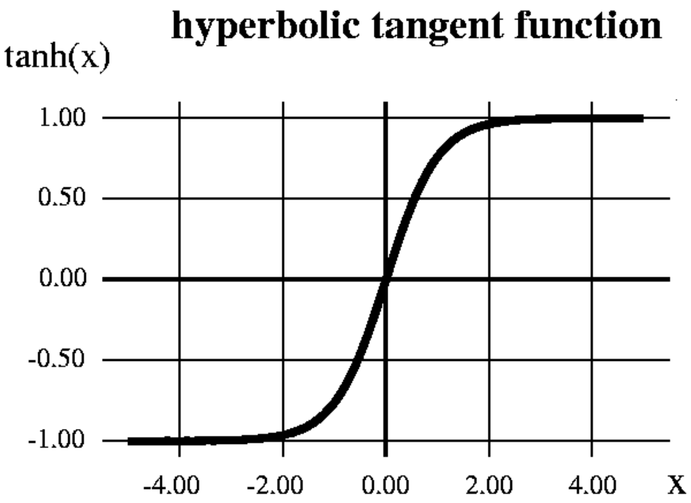


Figure 3-7. Graphical representation of the Hyperbolic Tangent Function *(SAS, 2019)*.

Besides hyperbolic tangent function, the Gaussian activation function is also used in evaluations to understand if the results improve and the equation of the function is given in Equation 3-8 (SAS, 2019).

$$e^{-x^2} \qquad \text{(Equation 3-8)}$$

where x is a linear combination of the X variables.

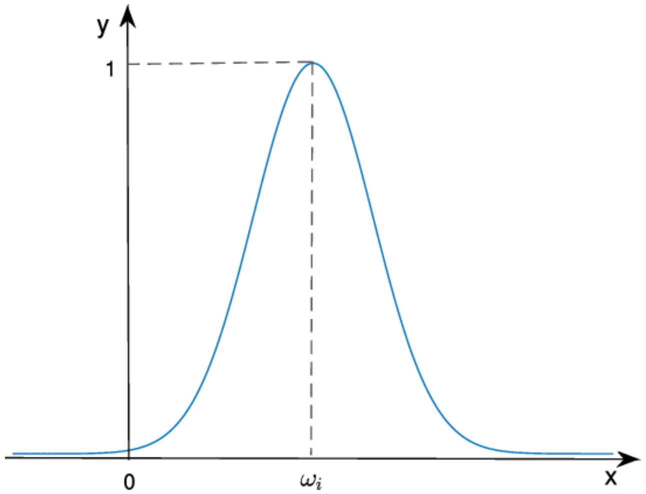


Figure 3-8. Graphical representation of the equation of the Gaussian function.

Moreover, changing the number of hidden nodes will obviously yield different results. The optimal number of hidden nodes must be calculated accordingly. Specifically, the number of neurons comprising the layer is equal to the number of features (columns) in the dataset. The optimal number of the hidden layer is usually between the size of the input and output layers (Heaton, 2019). The upper bound on the number of hidden neurons that won't result in over-fitting is calculated with the below formula (Hagan, 1999).

$$N_h = \frac{N_s}{(\alpha * (Ni + No))} \qquad \text{(Equation 3-9)}$$

Where:

Ns – is the number of samples in the training set

α – is an arbitrary scaling factor between 2-10, usually, 2 is the best for not overfitting the model.

Ni – is the number of input neurons

No – is the number of output neurons.

3.5.3 Partition modeling – decision trees

Validation to assess the credibility of a given model has become a necessary activity when one is faced with the need for making critical decisions based on the results of computer modeling and simulations. Methods like Naïve Bayes, Bootstrap Forest, Boosted Tree, K-Nearest Neighbors are only available in JMP Pro. Therefore, because of this limitation, the decision tree method is used in this thesis to test its goodness and future works must consider these methods as well. Users can split the data to achieve the best possible R^2 value as shown in Figure 3-9. One advantage of using decision trees is the ability to see how each attribute affects the outcome of the response value. As in all models, the goal of building a partition model is summarized as follows:

1. Calibrate model with a dataset

2. Compare model outcomes

3. Validate clustering goodness

For detailed information about the partition modeling algorithms, one can refer to the following source (Morrison, Bryant, Terejanu, Prudhomme, & Miki, 2013).

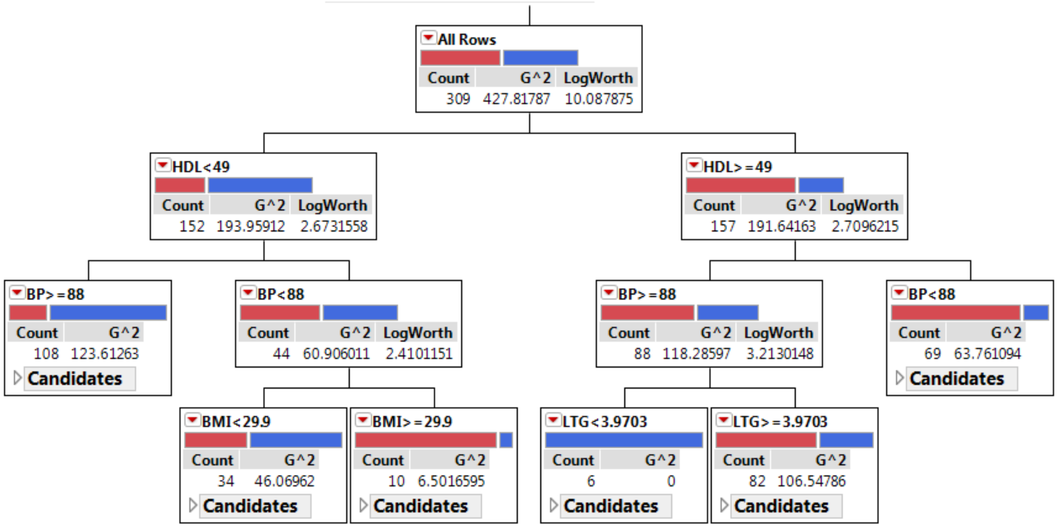


Figure 3-9. An example of a decision tree for a sample dataset available at *(Mukhtarli, 2019)*

3.6 Conclusion

In the end, the current study covers the identification of methods for creating homogeneous groups of pavements considering factors affecting their deterioration. Although, several previous research papers mentioned similar clustering operations, only the K-means approach was involved in the calculations. Because of this reason, it is necessary to carry out a research to introduce an additional clustering technique for grouping pavements. Moreover, several extrinsic and intrinsic evaluation methods are introduced to assess the goodness of the similarly acting groups.

Chapter 4

4. Case study of Costa-Rica

4.1 Introduction

According to the reports of the University of Costa Rica ‘s National Laboratory of Materials and Structural Models (LANAMME, 2019), as of 2016, the percentage of Costa Rican roads in good condition has increased over the past two years. However, the laboratory’s “Seventh Report on the Status of the National Roadway Network” also characterized the country’s roadway maintenance strategy as “reactive” and “poorly planned” (LANAMME, 2019). The National Roadway Council (CONAVI, 2019) repaired and maintained nearly half of the national roadway network during the last period spending with private participation of 663 million USD (TradingEconomics, 2019). However, CONAVI also found that the number of roads in ‘very bad’ condition has also increased. Costa Rica has 5,053 Kilometers (3,140 miles) of paved roads. Of these, 1,913 Km (1,189 miles) were in good condition in 2015 (TradingEconomics, 2019). Of the 48 percent of roads that are in good condition, 36 percent are in regular condition, while 5 percent require minor repairs. Having these results over the years requires extensive research to eradicate all these problems by the time.

The average temperature in Costa Rica is 76°F (ClimateData, 2019). Located in the tropics, Costa Rica has twelve climatic zones varying from hot and humid to cold and frosty. The east coast and the plain overlooking the Caribbean Sea are rain-soaked and receive more than 3,000 millimeters (120 inches) per year; here the climate can be defined as equatorial, that is, with no dry season. Precipitation data for Costa Rica’s different regions are given below in Table 4-1.

Table 4-1. Costa-Rica’s Average precipitation data by regions *(TradingEconomics, 2019)*

Limón - Average precipitation

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
Prec.(mm)	305	230	195	270	325	300	440	315	145	215	380	450	3565
Prec.(in)	12	9.1	7.7	10.6	12.8	11.8	17.3	12.4	5.7	8.5	15	17.7	140.4
Days	18	16	17	16	18	19	23	19	14	17	17	21	215

Puntarenas - Average precipitation

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
Prec.(mm)	6	4	5	30	205	215	175	225	295	280	130	30	1600
Prec.(in)	0.2	0.2	0.2	1.2	8.1	8.5	6.9	8.9	11.6	11	5.1	1.2	63
Days	2	2	3	8	19	21	18	22	24	25	15	6	164

Quepos - Average precipitation

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
Prec.(mm)	45	25	40	155	390	415	445	470	545	535	350	160	3570
Prec.(in)	1.8	1	1.6	6.1	15.4	16.3	17.5	18.5	21.5	21.1	13.8	6.3	140.6
Days	7	4	5	12	23	24	26	26	26	27	23	15	219

San José - Average precipitation

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
Prec.(mm)	15	5	20	45	230	240	210	240	310	305	145	40	1805
Prec.(in)	0.6	0.2	0.8	1.8	9.1	9.4	8.3	9.4	12.2	12	5.7	1.6	71.1
Days	2	0	2	5	17	20	20	21	22	22	13	4	148

4.2 Objective of the case study

The objective of this chapter is to identify homogeneous groups of roads from Costa Rican roads network based on the daily factors such as IRI; deflection; precipitation; and traffic (ESALs) that affect the pavement deterioration process. Several studies have been conducted considering the physical characteristics of the pavement however none of them carried extensive network-wide research to classify the road network of a whole country based on the pavement deterioration taking into account the factors that affect the deterioration.

4.3 Methodology of the case study

A case study of the Costa-Rica is presented in this chapter to discuss the results of the two main clustering techniques used to create homogeneous groups which included daily factors that affect the pavement deterioration and prediction models to evaluate generated clusters. To do this, statistical analysis and data mining were applied (using JMP software) with traffic loading, IRI, deflections and precipitation data collected by the University of Costa Rica (UCR). Additionally, IRI from the previous years were also available and this dataset has also been tested to investigate whether the inclusion of previous year pavement conditions will have any effect on the current and future years. To start with, only the main attributes were used in the clustering, through both methods; and then the IRI of 2015 were added. As mentioned before several data mining techniques were examined to create groups that act similarly. The techniques used include Logistic Regression, Neural Networks, Decision Tree, K-means Clustering and SOM clustering. Data were collected for the period of 2015 and 2017 separately and were merged using various methods to create the final dataset. Modeling and Clustering results were compared at the end to identify the best results. As has been mentioned in the literature review, all the prediction models were compared using R^2 value, and according to recent studies, a range of values between 0.68 and 0.80 was considered as an acceptable accuracy (Abudinen, Carvajal-Muñoz, & Fuentes, 2016).

Creating homogeneous groups to assess the pavement deterioration requires a reliable clustering method which must be tested considering all of the possible options. Despite there are many clustering techniques existing nowadays the most well-known and developed methods used in pavement management systems were tested in this thesis. Thus, two main clustering methods - K-means and Self Organizing Maps- were given the priority for the case study of Costa-Rica. Previously SOM clustering was mainly used in pavement design for crack identification. IRI data for Costa Rica was provided by Lanamme UCR. Dataset included centerline, right and left IRI measurements and Longitude and Latitude. In the beginning, only the main attributes (traffic, precipitation, deflection value and deflection basin area) were used in the clustering through both methods; and then the 2015 IRI data were added.

4.4 Clustering methods

4.4.1 K-means clustering

Case I – Generating clusters neglecting the IRI of the previous year.

In the first case, the predictors of K-means analysis included – IRI 2017, precipitation (mm/year), corrected center deflection value, deflection basin area, traffic (ESALs). A range of the number of clusters between 3 and 50 was examined and the optimal clusters have resulted with 7 clusters based on Cubic Clustering Criterion (CCC).

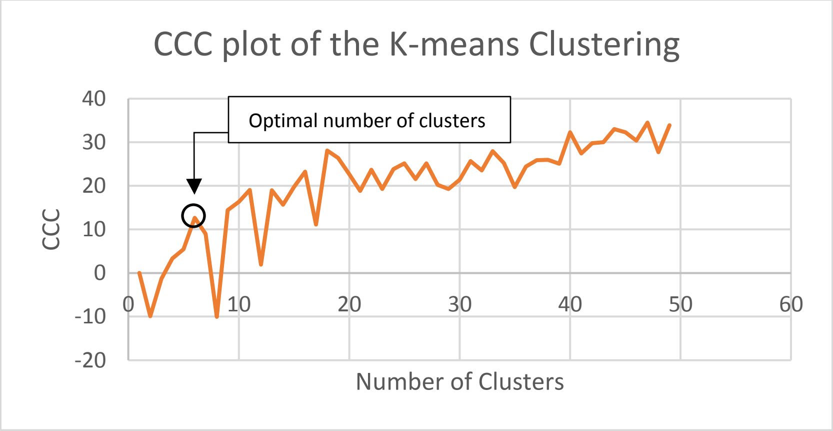


Figure 4-1. CCC plot of the first case of the K-means clustering (IRI of the previous year is neglected)

As Figure 4-1 describes the corresponding CCC value for a different number of clusters, and as discussed in the literature review the highest point before the first dramatic fall to a negative value in the trend is selected as the optimal number of clusters. However, if the CCC calculation is not interrupted with the negative numbers then the highest value of the coefficient is selected to be the optimal number of clusters. Values after the drop are not being considered because the clustering adds up noise and will create the risk of 'overfitting'. Centroids for the optimal clustering are shown in Figure 4-3.

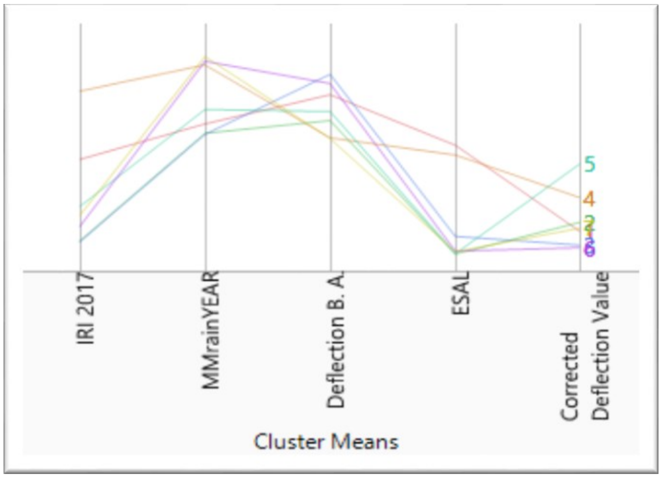


Figure 4-2. Parallel coordinate plots of the first case of the K-means clustering (IRI of the previous year is neglected)

It can be concluded from the parallel coordinate plots shown in Figure 4-2 that exposure to high traffic load inevitably led to the highest IRI values. Additionally, the same cluster parallels crossed the precipitation and deflection basin area lines in higher percentages which indicates that those 3 factors mainly affected the clustering.

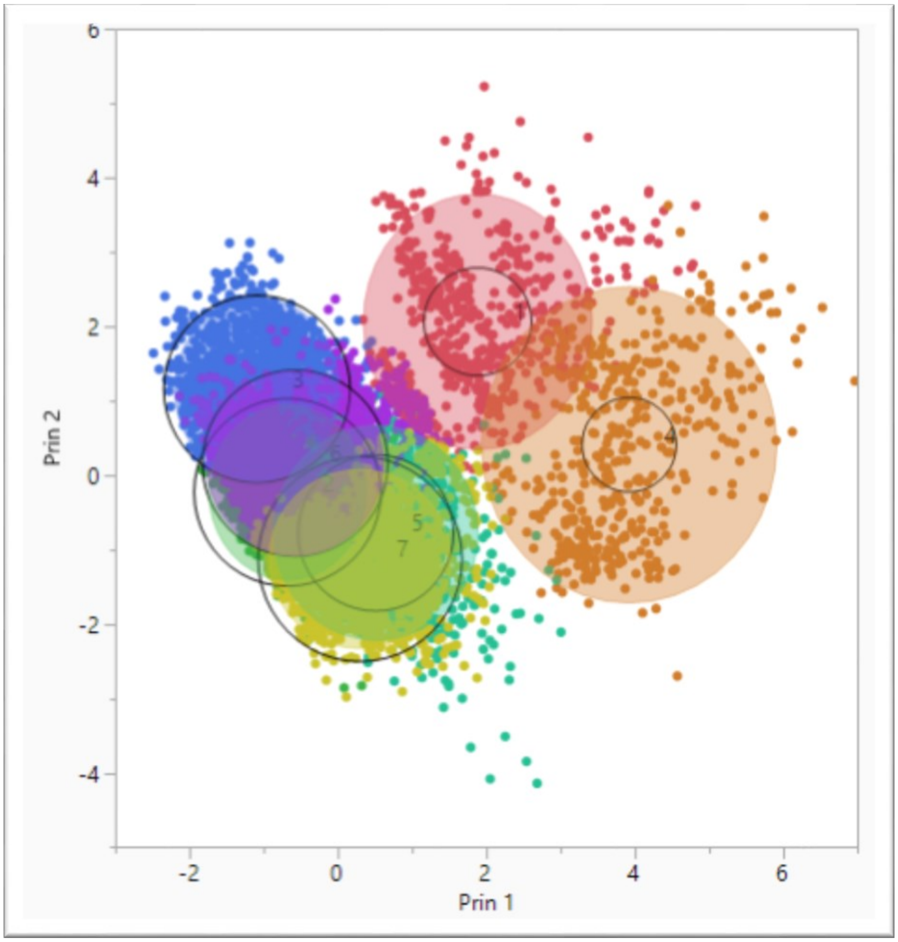


Figure 4-3. Bi-plot of the first case of the K-means clustering (IRI of the previous year is neglected)

In Figure 4-3 K-means bi-plot illustrates the visual distribution of the 7 optimal clusters and clustering seems successful on the XY plane. Figure includes only two principal X and Y components that can be displayed in a plot. However, when five components or features are suggested like in this thesis, to obtain a better view user must show 5 dimensional biplot which is not possible. Therefore, it is obvious that the clusters are separated into two generalized groups. One containing five clusters and the second group having two clusters. This separation can also be observed in Figure 4-2 where two clusters cross the traffic and the precipitation lines at the highest points. This phenomenon has already been explained above in the interpretation of Figure 4-2. The same limitation is the case for all the biplots illustrated in further subsections.

Case II – Generating clusters considering the IRI of the previous year.

In the second case of the K-means analysis predictor parameters included – IRI of 2015, IRI of 2017, Precipitation (mm/year), Corrected Center Deflection value, Deflection Basin Area, Traffic (ESALs). The number of clusters was set to range from 3 to 50 clusters. As has been explained in the previous analysis CCC plot failed to reach the negative breaking point before identifying the optimal number of clusters. This means, increasing the range of clusters from 50 to any further number will result in generating even more inefficient clusters. Therefore, the number of groups, in this case, was found to be 44 as it is the highest value, and the trend is shown in Figure 4-4.

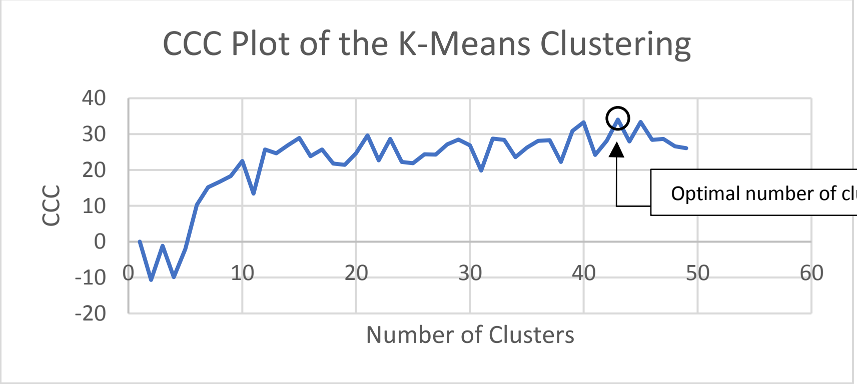


Figure 4-4. CCC plot of the second case of the K-means clustering (IRI of the previous year is included)

This significant change creates the first impression that the K-means clustering is not suitable for this analysis because of the overlapping values. Since K-means analysis is using a "hard-clustering" method it becomes less reliable when additional data is included in the analysis. It is because "hard clustering" struggles when deciding whether to include the overlapping values in the same cluster or separate them into two distinct groups. This results in the shifting of the optimal number of clusters beyond the meaningful number.

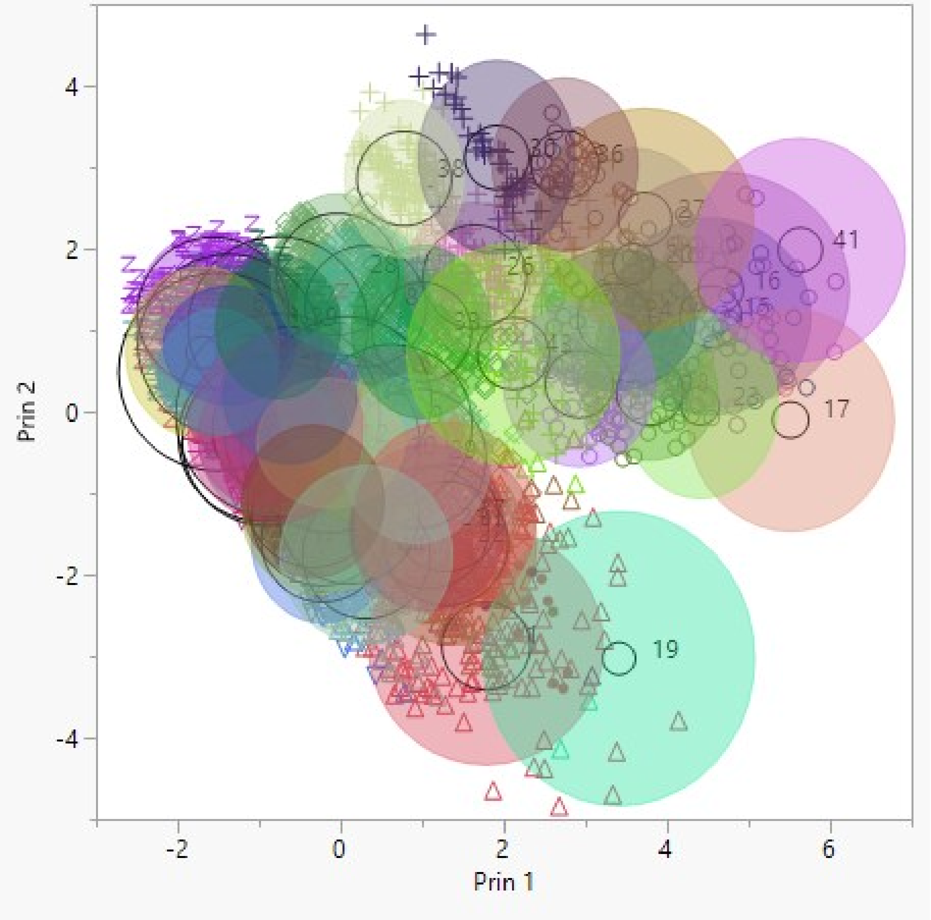


Figure 4-5. Bi-plot of the second case of the K-means clustering (IRI of the previous year is included)

Obviously, from Figure 4-5 and Table 4-2 it can be concluded that the clustering failed since some clusters had less than 100 elements.

Table 4-2. Summary of the second case of the K-means clustering (IRI of the previous year is included)

Cluster No	Elements count	Cluster No	Elements count
1	112	22	220
2	280	23	21
3	215	24	76
4	290	25	475
5	405	26	153
6	329	27	41
7	5	28	213
8	210	29	278
9	373	30	60
10	302	31	226
11	594	32	459
12	397	33	141
13	278	34	562
14	402	35	66
15	54	36	67
16	29	37	164
17	19	38	128
18	65	39	462
19	3	40	275
20	24	41	28
21	213	42	248
		43	75
		44	233

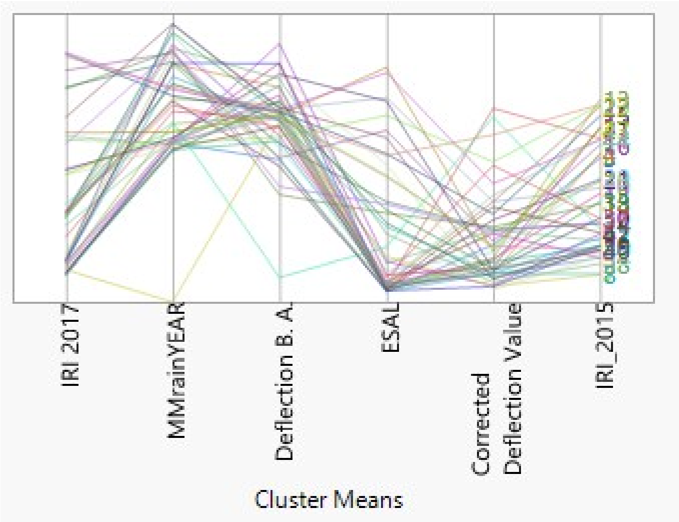


Figure 4-6. Parallel coordinate plots of the second case of the K-means clustering (IRI of the previous year is included)

It can be observed from Figure 4-6 that the clusters variation over the lines of the predictors is very high which means that clustering is not stable and reliable.

Thus, following the above results it is possible to conclude that there are several numbers of reasons why K-means clustering is not an ideal option for this dataset:

1. Adding new attributes (features) to the data frame will result in an unbounded increase in the optimal number of clusters. I.e. when additional features are introduced to the analysis, the value of K representing optimal number of homogeneous groups skyrockets. Thus, K-means clustering is vulnerable to overfitting and makes the automated clustering impossible.
2. The K-means clustering method cannot decide on overlapping data points.
3. Some of the clusters do not include any data points or only include less than 100 members (<1% of the data).

4.4.2 SOM clustering

Case I – Generating SOM clusters neglecting the IRI of the previous year.

The same as before, the clustering analysis using the SOM technique was performed with and without the previous year's IRI results. For the case without the previous year results, the range for the number of clusters examined was between 3 and 50 and the attributes selected for clustering were IRI (2017), traffic, precipitation, deflection, deflection basin area. According to the same principle explained in K-means clustering, the optimal number of clusters was obtained as 8 based on CCC measure and the graph illustrating the CCC versus the number of clusters is given below in Figure 4-7.

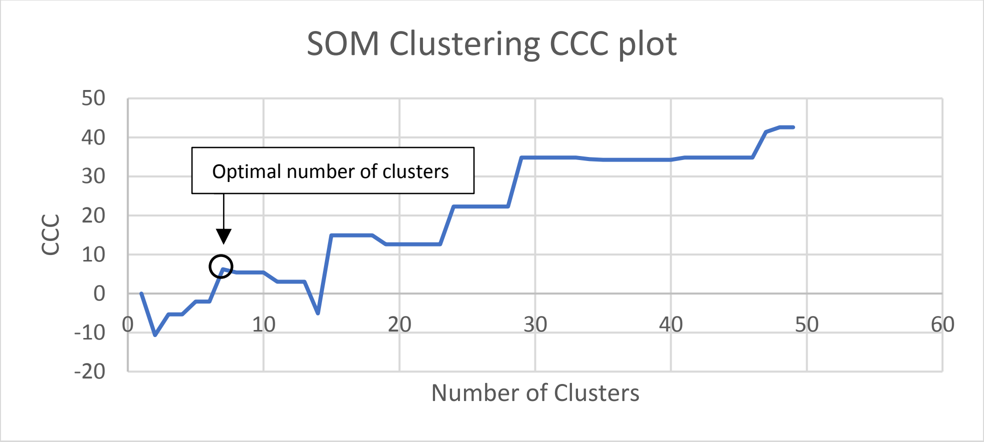


Figure 4-7. CCC plot of the first case of the SOM Clustering (IRI of the previous year is neglected)

The summary of the results of SOM clustering is given in Table 4-3. The numbers given in the table explain the average values and standard deviations of each predictor in each cluster. It can be observed that the average of the IRI value is ranging from 5.24 to 13.39 and these numbers fall into the “fair” and “poor” category according to Section 2.3.3. Furthermore, the numbers given in Table 4-3 clearly indicates that the average precipitation and traffic load has the most impact on the deterioration of the pavement since this combination led to the highest average IRI among other clusters.

Table 4-3. SOM clustering results (IRI of the previous year is neglected)

Cluster Means					
Cluster	IRI of 2017	(mm)rain/year	Deflection B. A.	ESAL	Corrected Deflection Value
1	13.3990	3412.37305	22.8068395	2452668.83	874.608145
2	10.1526	2517.48927	27.6730749	2744797.34	494.78778
3	7.7874	3464.55215	26.8846906	1429306.23	381.751792
4	6.5875	3538.77018	22.8584873	384614711	531.426401
5	6.8889	2765.3743	25.7714747	377787.984	1062.02768
5	5.7136	3284.12416	28.4385489	263578.365	381.306466
7	5.1718	2315.59893	25.8948152	457420.196	485.446582
8	5.2413	2268.41203	30.7461475	760555.478	314.489189
Cluster Standard Deviations					
Cluster	IRI of 2017	(mm)rain/year	Deflection B. A.	ESAL	Corrected Deflection Value
1	1.7591	482.823344	3.6067309	867399.711	383.565138
2	1.0662	440.755642	2.95733963	765799.693	307.374212
3	0.5868	429.907024	2.45631157	467498.284	212.00332
4	1.3226	373.936467	2.12877183	272093.516	198.274428
5	1.4018	448.55	2.55816678	360472.957	391.694528
6	1.0301	386.086451	2.41021845	168614.447	183.063429
7	0.8968	265.363256	2.29446753	452109.767	205.79006
8	0.9315	223.694718	2.05790135	537517.351	187.491375

Table 4-4. SOM cluster summary (IRI of the previous year is neglected)

Cluster Summary	
Cluster	Number of Elements
1	388
2	494
3	625
4	1675
5	1448
6	1524
7	1747
8	1369

Moreover, when comparing the first case of the K-means (Figure 4-1) and SOM (Figure 4-7) analysis it becomes obvious that the results are very similar to each other where both techniques derive the optimal number of clusters around 7 or 8.

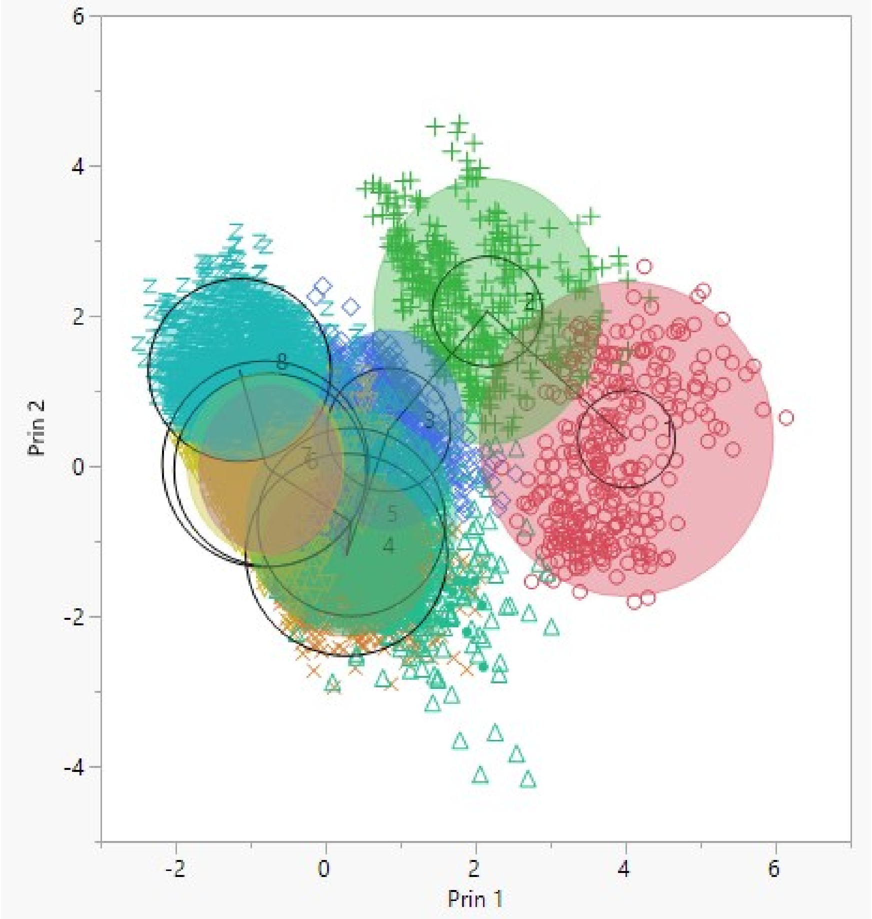


Figure 4-8. Bi-plot of the first case of the SOM Clustering (IRI of the previous year is neglected)

Case II – Generating SOM clusters considering the IRI of the previous year.

In the second case of the SOM analysis predictors included – IRI of 2015, IRI of 2017, Precipitation (mm/year), Corrected Center Deflection value, Deflection Basin Area, Traffic (ESALs), and the range of the number of the clusters to be tested based on their CCC value was between 3 and 50. As a result, from Figure 4-9 it is obvious that the number of the optimal clusters still remained as eight (8) despite the addition of new data points. This clearly indicates that SOM analysis is advantageous when handling overlapping values. Since SOM is using a soft-clustering method it becomes more reliable as overlapping values can be included in more than one cluster.

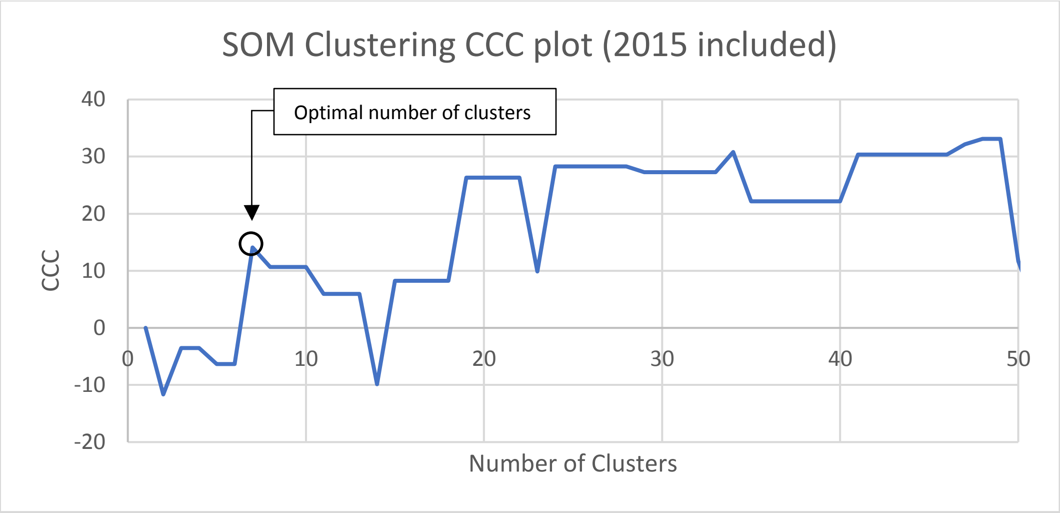


Figure 4-9. CCC plot of the second case of the SOM clustering (IRI of the previous year is included)

Once again cluster summary given in Table 4-5 is describing how well the data is distributed in each cluster for both cases of SOM grouping. Means and standard deviations of the data of 8 clusters are given in Table 4-5 and the average IRI value is ranging from 5 to 13. The mean IRI value of the data collected in previous years is shown to be scaled between 3 and 7 when the numbers are rounded up. This indicates that the pavement has been subject to deterioration in the past 4 years, mainly affected by the traffic load and precipitation as it was concluded in the first case of the SOM analysis.

Table 4-5. SOM clustering results (IRI of the previous year is included)

Cluster Means						
Cluster	IRI of 2017	(mm)rain/year	Deflection B. A	ESAL	IRI of 2015	Corrected Deflection Value
1	13.2904979	2880.5974	25.2883125	3232489.68	6.75187447	951.321709
2	10.2152481	2903.92921	25.8686959	2279669.64	3.29075069	562.3357
3	7.36816934	2816.16433	25.3147521	510992.758	5.18521744	1177.84138
4	6.5436017	3512.95312	23.2885009	387024.52	3.63897771	534.090024
5	5.77480474	2761.48583	26.2805692	337224.802	6.45521095	539.118268
5	5.42765579	2267.94055	25.7472936	477721.522	3.44828755	579.713559
7	5.26442258	2336.34726	29.8590004	753269233	3.1014458	312.872333
8	6.09647863	3525.07176	29.2988813	501523.532	2.90649185	263.572251
Cluster Standard Deviations						
Cluster	IRI of 2017	(mm)rain/year	Deflection B. A	ESAL	IRI of 2015	Corrected Deflection Value
1	2.19270592	551.802309	3.350565	917683.223	1.50047204	426.750274
2	1.84449953	706.395168	4.084587	628299.902	1.37637832	274.932459
3	1.40038954	467.656247	2.727507	391966.249	1.80450426	399.807037
4	1.33764907	370.798907	2.293022	339526.692	1.44468097	206.123715
5	1.23264677	493.085999	3.033251	338980.514	1.35937414	218.977024
6	1.1157487	276.125739	2.322553	424641.376	1.05018202	243.902217
7	0.92729304	232.753835	2.328297	555917.307	1.1776459	171.916791
8	1.21267621	366.934554	2.309911	510058.194	1.17231347	153.313842

Table 4-6. SOM Cluster Summary (IRI of the previous year is included)

Cluster Summary	
Cluster	Number of Elements
1	280
2	740
3	1.023
4	1795
5	1267
6	1304
7	1442
8	1359

When comparing the group summaries of the first and the second case of the SOM analysis obvious similarities appear in the numbers. It means that the addition of the data did not complicate the assignment of the data points to the clusters. Considering summaries of the second case of both K-means and SOM clustering given in Table 4-2 and Table 4-6, groups generated by SOM had an apparent advantage of handling the additional data.

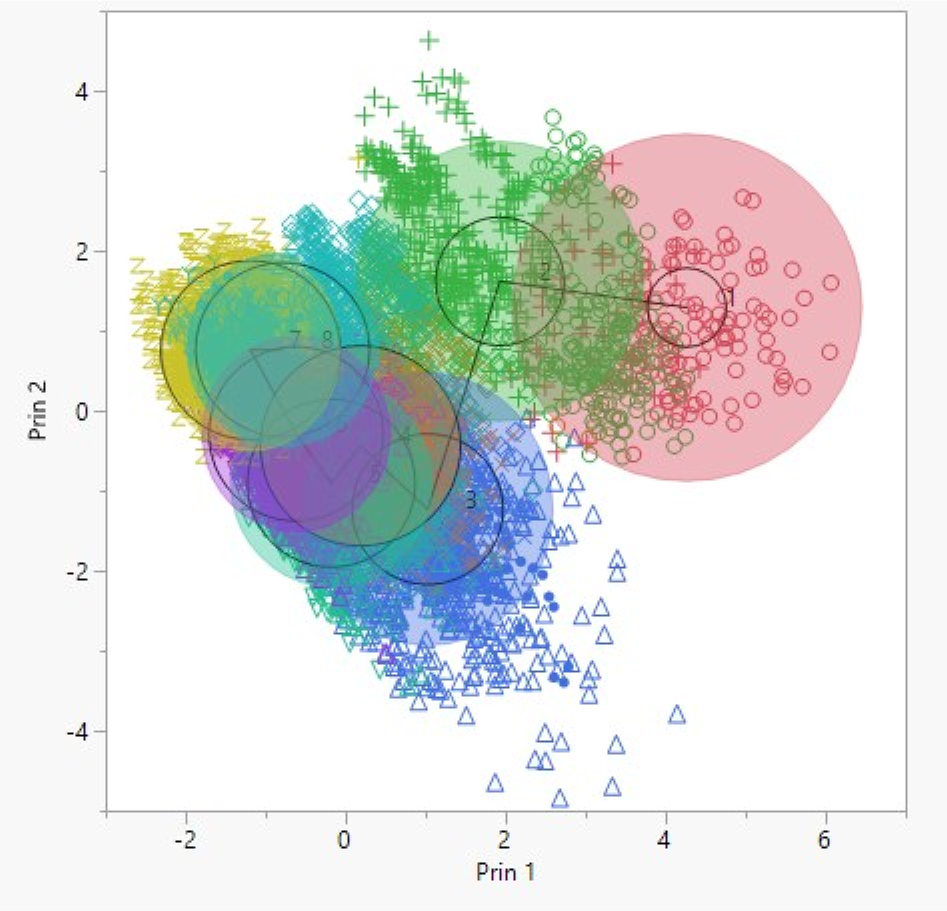


Figure 4-10. Bi-plot of the second case of the SOM clustering (IRI of the previous year is included)

In conclusion, as a result of 4 distinct cases of two different clustering methods the following statement can be hypothesized: Considering the similarities in bi-plots of the first case of K-means and both cases of SOM clustering, it is safe to mention that the clustering is successful and addition of a new dataset can be handled by Self Organizing Maps.

4.5 Cluster evaluation

Clusters that were created are homogeneous groups which are the main purpose of this thesis since it'll help us to understand the relationship between the daily factors and the pavement deterioration process.

A perfect cluster exists in a dataset only if the data can be divided into groups where all of the data elements within a group show identical characteristics over the range of attributes. It implies the necessity of finding the correct optimal number of clusters. Therefore, it is important to verify the quality of the generated homogeneous groups. There are two widely used methods to check the goodness of clusters.

1. Extrinsic Evaluation
2. Intrinsic Evaluation

Extrinsic (External) evaluation of a cluster is important to check the goodness of the homogeneous groups. An external evaluation is a measure of agreement between two partitions where the first partition is the a priori known clustering structure, and the second results from the clustering procedure (Dudoit, 2002). In order to run the extrinsic evaluation for the generated clusters, three different prediction models will be built. These prediction models are Logistic Regression, Neural Networks, and Partitioning. In total, thirty-four (34) Neural Network, six (6) Decision Tree and six (6) Logistic regression models were trained and discussed in the next subsections with all the possible combinations and the best one was selected with the help of the model comparison tool discussed at the end of this section.

4.5.1 Extrinsic cluster evaluation

4.5.1.1 Logistic regression model

To start with, six different combinations (Table 4-7) were created within the Logistic Regression model where clustering results and the data from the previous years were also considered.

Table 4-7. Combinations of the logistic regression models.

Logistic Regression Models						
Analysis	1	2	3	4	5	6
Combinations	Evaluation (w/o previous year data)	Evaluation (w/ previous year data)	Evaluation including SOM Clustering (w/o previous year data)	Evaluation including K-means Clustering (w/o previous year data)	Evaluation including SOM Clustering (w/ previous year data)	Evaluation including K-means Clustering (w/ previous year data)

There are three types of modeling variables: continuous, ordinal and nominal and Logistic regression is a model where the response variable is selected to be a class (nominal variable). For the first analysis given in Table 4-7, only 4 main predictors (casual factors) were selected as independent variables. Then the same evaluation procedure was repeated until all of the six possible analysis options shown in Table 4-7 are tested.

The accuracy of logistic regression models was considerably lower than the results of the neural network and partitioning models. As for the first analysis, R^2 value of the IRI prediction was 0.68 (68% accuracy) where the lowest results from the Neural Network and Partitioning models derived the R^2 of 0.70 and 0.85 given in Section 4.5.1.2 and 4.5.1.3. Despite the lower values, the analysis run until the end and below results were obtained as shown in Table 4-8.

Table 4-8. Logistic Regression results.

Analysis No	Generalized R^2	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate
1	0.6908	0.1933	0.2307	0.1018	0.064
2	0.6864	0.1956	0.2327	0.1029	0.0649
3	0.7956	0.137	0.1896	0.0708	0.0409
4	0.7723	0.1505	0.201	0.0808	0.0511
5	0.7933	0.1383	0.1906	0.0717	0.0423
6	0.7754	0.1487	0.1989	0.0793	0.0479

Considering the results given in Table 4-8, it can be assumed that logistic regression is not the most suitable prediction model for this case study. However, when comparing and evaluating the clustering methods (through extrinsic performance measure), the outcome of the analysis involving SOM clustering is still better than the results of the options that include K-means clusters. Moreover, it is obvious from Table 4-8 that the result derived the accuracy of 79% (SOM) and 77%(K-means). Additionally, it is concluded from the results that the least effective model deriving the lowest R^2 and the highest RASE value was the one where the analysis considered the previous year's data without involving the homogeneous groups. On the contrary, the model involving SOM groups with no previous year's data performed as the best model with the highest accuracy of 79% and the lowest value of RASE.

In short, although the logistic regression demonstrated lower results among the rest of the models, the best prediction was noted in the 3rd analysis where the SOM clustering played a major role.

4.5.1.2 Neural networks

In this section as an extrinsic evaluation method neural network is used to determine the goodness of SOM and K-means clusters by predicting IRI value. Different ANN combinations shown in Table 4-9 were trained and tested to achieve the best results to evaluate the generated clusters. In the beginning, these models were tested with a hyperbolic tangent activation function

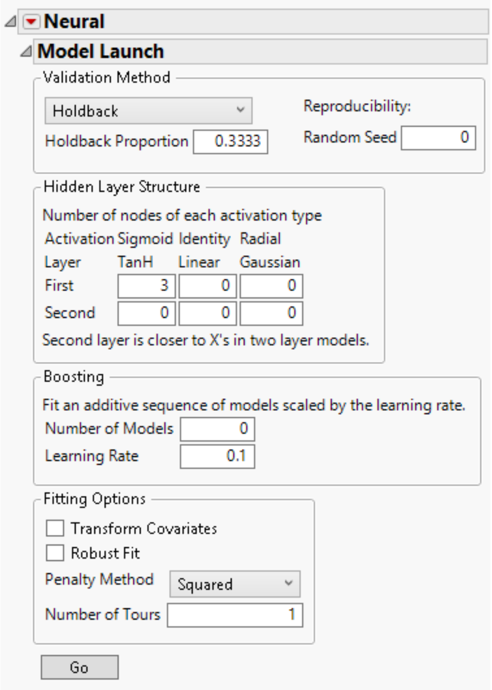
(Section 3.5.2) and the main predictors remained the same which included traffic, precipitation, deflection, and deflection basin area. Then Gaussian and Linear activation functions, and 10 step boosting was applied to the combination which derived the highest R^2 value from the tanH model. Boosting is an additional function where the neural network model trains the same dataset several numbers of times identified by the user. Above mentioned boosting number is called a 'step' and it has been set to 10 in this case study because boosting more than 10 steps didn't increase the accuracy of the prediction. Besides various functions, several different combinations of hidden layers and hidden nodes were also tested.

Table 4-9. Combinations of neural network models built for the extrinsic evaluation of clusters.

Combinations of hidden layers	Neural Network Models						
	Evaluation (w/o previous year data)	Evaluation (w/ previous year data)	Evaluation including SOM Clustering (w/o previous year data)	Evaluation including K-means Clustering (w/o previous year data)	Evaluation including SOM Clustering (w/ previous year data)	Evaluation including K-means Clustering (w/ previous year data)	Activation Function
	3 [1]	3	3	3	3	3	tanH
	20	20	20	20	20	20	
	25	25	25	25	25	25	
	23-3 [2]	23-3	23-3	23-3	23-3	23-3	
	25-3	25-3	25-3	25-3	25-3	25-3	
	N/A	N/A	25-3 (Best Model [3])	N/A	N/A	N/A	Gaussian
							Linear
							Boosting

[1] Number of hidden nodes in the first layer of Neural Networks.
[2] Number of hidden nodes in the first and second layers of Neural Networks.
[3] The model that provided the best R^2 value from tanH activation functions.

Firstly, a single hidden layer was tested with three (3) hidden nodes using tanH function without considering clustering and the previous year's data in the analysis. In the beginning, IRI data from 2017 was identified as a response value and 4 main factors (precipitation, traffic, deflection, and deflection basin area) affecting the deterioration of the pavement were selected as features. Then using the following window shown in Figure 4-11 the number of hidden layers, hidden nodes, activation function, and Holdback value was configured. Holdback value was set to 0.333 which means 33.3 % of the data will be automatically considered as a validation dataset.



Neural

Model Launch

Validation Method

Holdback

Holdback Proportion 0.3333

Reproducibility:

Random Seed 0

Hidden Layer Structure

Number of nodes of each activation type

Activation	Sigmoid	Identity	Radial
Layer	TanH	Linear	Gaussian
First	3	0	0
Second	0	0	0

Second layer is closer to X's in two layer models.

Boosting

Fit an additive sequence of models scaled by the learning rate.

Number of Models 0

Learning Rate 0.1

Fitting Options

□ Transform Covariates

□ Robust Fit

Penalty Method Squared

Number of Tours 1

Go

Figure 4-11. Neural networks model launch – defining the number of hidden nodes, layers and activation function.

The purpose of the validation set is to separate the certain percentage of the data from the main dataset and run the analysis separately. At the end of the analysis, the results of both validation and the training sets are shown in Table 4-10. It is important to note that when comparing these results closer R^2 values of training and validation sets indicate better results.

Table 4-10. Final neural networks result with just 3 hidden nodes (results of the first model).

Neural Network Model tanH 1-1(3)			
Training Set		Validation Set	
Measures	Value	Measures	Value
R^2	0.7096304	R^2	0.7188083
RASE	1.1790047	RASE	1154.9021
Mean Abs Dev	0.93914182	Mean Abs Dev	0.92041339
– LogLikelihood	9785.7045	– LogLikelihood	4856.1686
SSE	8590.5218	SSE	4193.1206
Sum Freq	6180	Sum Freq	3090

It is described in Table 4-10 that the R^2 value of the validation set is obtained as 0.718. Thus, the accuracy of the prediction of two sets illustrates significant similarities. This, in fact, means that the model is still successful, however, it is important to run the analysis with other options for better results. Therefore, the procedure explained above will be repeated for all the combinations shown in Table 4-9.

It is obvious from the results shown in Table 4-11, that the best model configuration neglected the IRI of the previous years and included ten (10) step Boosting with two hidden layers -involving twenty-three (23) and three (3) hidden nodes respectively in each layer- and SOM clustering. The R^2 value of the given model was noted as 0.8866 which means the results were 88.6 % accurate when predicting the IRI value. It also proves that without considering the data from the previous years SOM Clustering brings up better results than K-means clustering because all the models that included K-means analysis were having a lower accuracy. The goodness of SOM clustering is not only interpreted with the R^2 value. It is also necessary to consider RASE and AAE (Section 2.5.2) values when comparing the results of the combinations of two different clustering methods. The combination which included SOM clustering had the lowest Average Absolute Error and Root Average Squared Error values of 0.5936 and 0.7411 respectively, where models including K-means clustering results were having a minimum of 0.6183 and 0.7698 AAE and RASE values accordingly.

Additionally, it can be derived from Table 4-11 that the worst combination neglected the clustering. The results were obtained from an ANN model which included a single layer having three hidden nodes with the lowest R^2 and the highest RASE and AAE values. Thus, it implies that the clustering is successful and it increases the accuracy of the prediction.

Table 4-11. Final results of all neural network models.

Measures of Fit for IRI of 2017 (Neural Networks)			
Predictor	RSquare	RASE	AAE
2017 -2015EXCLUDED - KMEANS - 25 - 3	0.8665	0.8005	0.6354
2017 -2015INCLUDED - 3	0.7059	1.1881	0.9351
2017 -2015INCLUDED - 20	0.8457	0.8606	0.6744
2017 -2015INCLUDED - 25	0.8479	0.8544	0.6748
2017 -2015INCLUDED - 23 - 3	0.8629	0.8113	0.638
2017 -2015INCLUDED - 25 - 3	0.8623	0.813	0.6383
2017 -2015EXCLUDED - 3	0.7318	1.1348	0.9137
2017 -2015EXCLUDED - 20	0.8403	0.8757	0.6855
2017 -2015EXCLUDED - 25	0.8433	0.8673	0.6808
2017 -2015EXCLUDED - 23 - 3	0.8617	0.8147	0.644
2017 -2015EXCLUDED - 25 - 3	0.8695	0.7914	0.6255
2017 -2015INCLUDED - SOM - 3	0.8176	0.9399	0.7445
2017 -2015INCLUDED - SOM - 20	0.8761	0.7747	0.6209
2017 -2015INCLUDED - SOM - 25	0.8739	0.7814	0.6226
2017 -2015INCLUDED - SOM - 25 - 3	0.8773	0.7707	0.6141
2017 -2015INCLUDED - KMEANS - 3	0.8065	0.9681	0.7703
2017 -2015INCLUDED - KMEANS - 20	0.8735	0.7828	0.626
2017 -2015INCLUDED - KMEANS - 25	0.8734	0.7831	0.6286
2017 -2015INCLUDED - KMEANS - 23 - 3	0.8606	0.8217	0.6543
2017 -2015EXCLUDED - SOM - 3	0.8243	0.9224	0.7284
2017 -2015EXCLUDED - SOM - 20	0.873	0.7841	0.629
2017 -2015EXCLUDED - SOM - 25	0.879	0.7654	0.6113
2017 -2015EXCLUDED - SOM - 23 - 3	0.8842	0.7487	0.5987
2017 -2015EXCLUDED - SOM - 23 - 3 Boost10	0.8866	0.7411	0.5936
2017 -2015EXCLUDED - SOM - 23 - 3 Gaussian	0.8752	0.7775	0.6198
2017 -2015EXCLUDED - SOM - 23 - 3 Linear	0.7634	1.0703	0.8431
2017 -2015EXCLUDED - SOM - 25 - 3	0.8823	0.755	0.5999
2017 -2015EXCLUDED - KMEANS - 3	0.7983	0.984	0.78
2017 -2015EXCLUDED - KMEANS - 20	0.8596	0.821	0.6485
2017 -2015EXCLUDED - KMEANS - 25	0.8702	0.7894	0.6302
2017 -2015EXCLUDED - KMEANS - 23 - 3	0.8711	0.7856	0.622
2017 -2015 INCLUDED - KMEANS - 25 - 3	0.8776	0.7698	0.6183

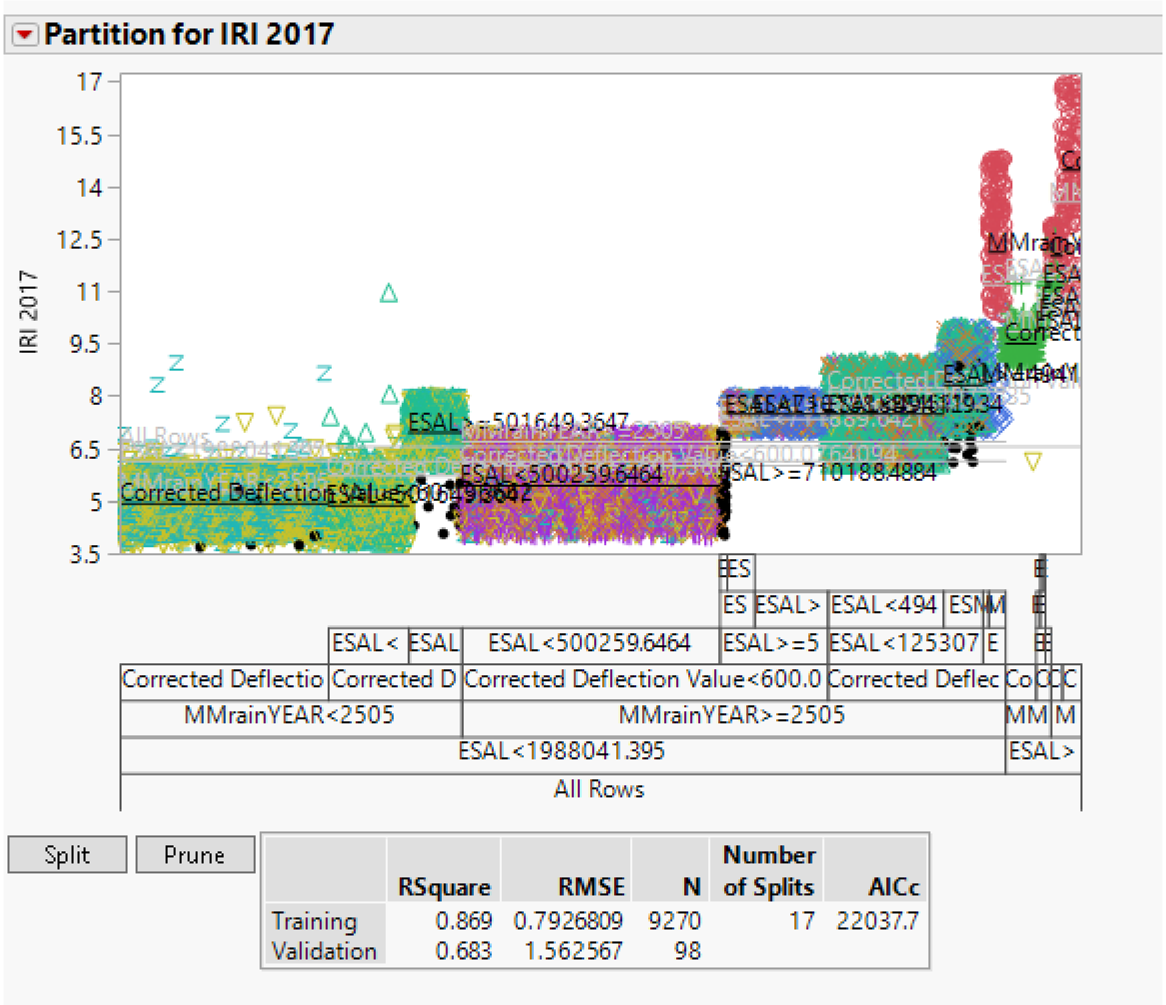
4.5.1.3 Partition modeling

As in all above-mentioned prediction models, six different options were tested in partition modeling (Decision tree) where clustering results and the data from the previous years were considered in some of the analyses.

Table 4-12. Combinations of partitioning models built for the extrinsic evaluation of clusters.

Partitioning Models (Decision Tree)						
Analysis	1	2	3	4	5	6
Combinations	Evaluation (w/o previous year data)	Evaluation (w/ previous year data)	Evaluation including SOM Clustering (w/o previous year data)	Evaluation including K-means Clustering (w/o previous year data)	Evaluation including SOM Clustering (w/ previous year data)	Evaluation including K-means Clustering (w/ previous year data)

In the first analysis, the model only included the predictors that contribute to the deterioration of the pavement. At the beginning of the analysis, the decision tree was split (Figure 4-12) until the best result achieved. Splitting must stop when the R^2 value is constant. R^2 value remains constant when splitting the data into partitions doesn't contribute to the results. As described in Figure 4-12 the data has been split 17 times to achieve the highest possible R^2 value (0.869) before the value kept constant. However, if we look at the accuracy of both validation and training sets it can be concluded that there is a huge difference in numbers. Thus, it is assumed that the first analysis given in Table 4-12 is not successful and the rest of the options must be calculated to achieve higher accuracy. Therefore, the same procedure later applied to all of the possible partitioning cases shown in Table 4-12. The results of the partitioning shown below in Table 4-13 includes all of the six possible options.



	RSquare	RMSE	N	Number of Splits	AICc
Training	0.869	0.7926809	9270	17	22037.7
Validation	0.683	1.562567	98		

Figure 4-12. Partitioning process - splitting the decision tree until the best result is achieved.

Table 4-13. Final results of all partitioning models.

Partitioning Model Type	R^2	RASE	AAE
IRI of 2017 Predictor - 2015 Inc	0.869	0.7926	0.613
IRI of 2017 Predictor - 2015 Exc	0.8711	0.7867	0.619
IRI of 2017 Predictor - 2015 Inc - SOM	0.8726	0.7822	0.6161
IRI of 2017 Predictor - 2015 Exc - SOM	0.8776	0.7665	0.6098
IRI of 2017 Predictor - 2015 Inc - KMEANS	0.852	0.8429	0.6501
IRI of 2017 Predictor - 2015 Exc - KMEANS	0.8673	0.7981	0.6251

The final outcome of the analysis indicates that the best results having an accuracy of 87% were achieved where the data from the previous years was neglected and SOM clustering was included. The results shown in Table 4-13 indicates that the lowest AAE value is observed in the analysis that included SOM clustering. Tn the contrary, the lowest accuracy, and the highest AAE value were achieved in models involving K-means.

In conclusion, the results of all three extrinsic cluster evaluation methods indicate that the modeling options involving SOM clusters and neglecting the IRI of the previous years were the best responding models.

4.5.2 Intrinsic cluster evaluation

Intrinsic evaluation is another method of determining the goodness of clustering. Unlike extrinsic evaluation, the intrinsic assessment of clusters is easier and helps understand the structure of the data. Intrinsic clustering evaluation is carried out for the SOM since it yielded the best results. WEKA's intrinsic "cluster-to-cluster" evaluation method is used in this analysis. The formula shown in Equation 4-1 was to evaluate the goodness of the clustering based on Table 4-14.

Table 4-14. WEKA's intrinsic "cluster-to-cluster" evaluation method.

Practice	Reference							
	R1	R2	R3	R4	R5	R6	R7	R8
C1	0	0	0	0	0	0	0	0
C2	0	0	0	0	0	0	0	0
C3	0	0	0	0	0	170	0	0
C4	0	0	9	0	0	6	0	1
C5	0	0	5	0	0	0	0	0
C6	0	0	9	0	119	11	0	0
C7	0	0	10	0	3	0	0	0
C8	0	1	43	0	0	0	0	171
C9	0	0	0	0	0	0	1	0
Cumulative	0	1	76	0	122	187	1	172

Columns given in Table 4-14 (values indicated with "R") represent the original reference clusters (generated in SOM with the original performance indicator -IRI- measured in field) and the rows (values indicated with "C") illustrate the each cluster derived from the SOM clustering where the performance indicator -IRI- is predicted by the best responding ANN model (discussed in Section 4.5.1.2). Thus, the numbers filled in the table represent the difference in number of data points between "R – C". Determining the difference of these values is very straightforward when both clusters are graphed on the same X and Y plane as described in Figure 4-13.

To identify the goodness of clustering, the sum of the highest value from each reference cluster given in Table 4-14 must be divided by the sum of the cumulative of reference clusters.

$$\mathrm{G} = \frac{(\mathrm{R1+R2+R3+R4+R5+R6+R7+R8})}{\sum \mathrm{C}} \times 100 \qquad \text{Equation 4-1}$$

Where:

G – the goodness of clustering in percentage

R – the number of data points in each original reference cluster

C – the cumulative sum of each reference cluster

As a result of the calculation of intrinsic analysis, the goodness of SOM clustering is obtained as 90%.

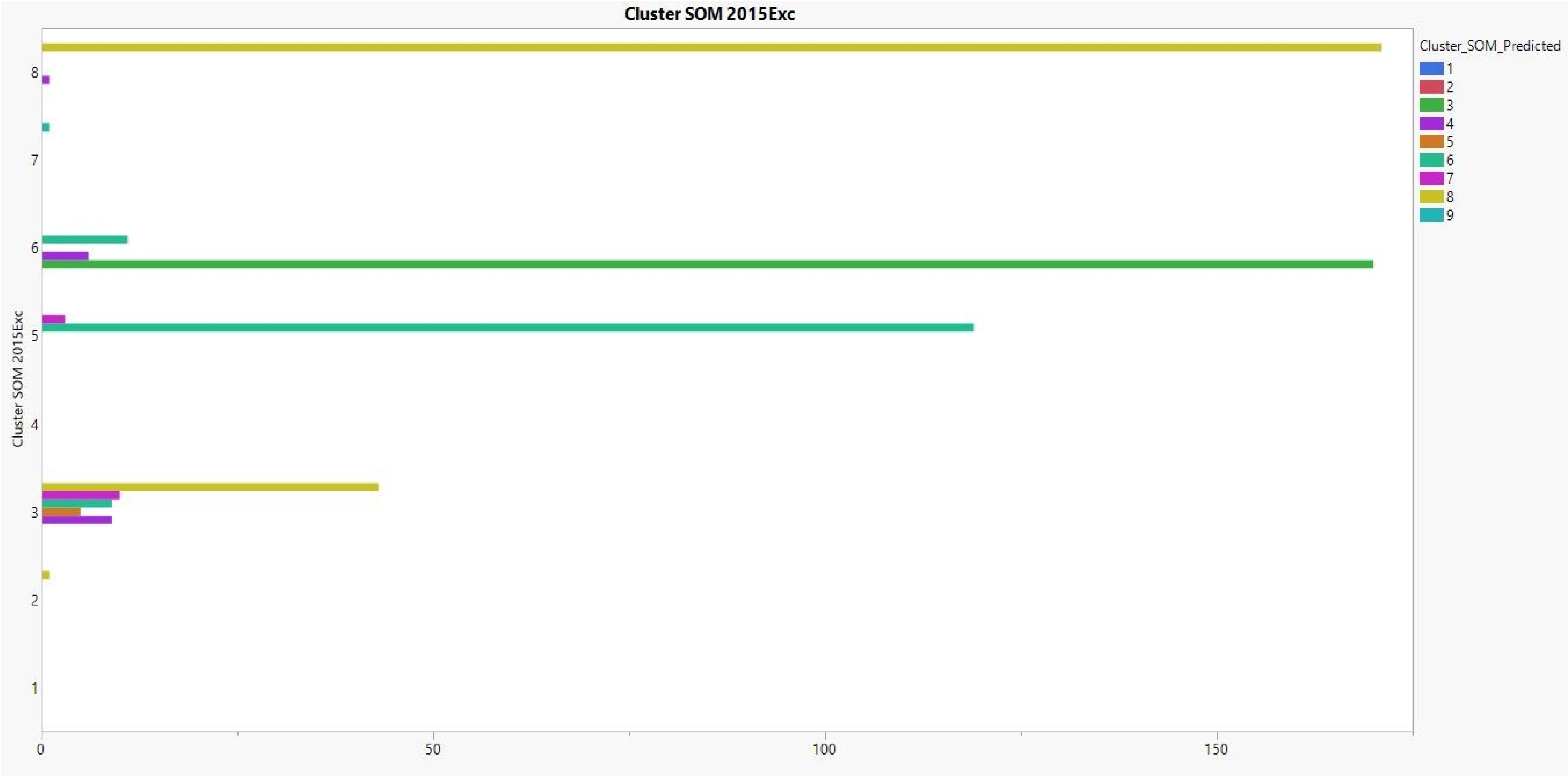


Figure 4-13. Difference in the number of data points in each group of both original and predicted clusters.

4.6 Results and discussions

The Summary of the results of both clustering techniques is given in Table 4-15.

Table 4-15. Summary of all analyzed cluster cases.

Clustering Technique	Clustering Case	Number of Optimal Clusters	Extrinsic Evaluation Results		
			Highest R^2 Value	RASE Value	AAE Value
K-means	Case I*	7	0.8735	0.7828	0.626
	Case II*	44	0.8776	0.7698	0.6183
Self Organizing Maps	Case I	8	0.8773	0.7707	0.6141
	Case II	8	0.8866	0.7411	0.5936

*Case I: IRI of the previous year is included.
*Case II: IRI of the previous year is neglected.

After comparing all the models considered in this study, it can be concluded from Table 4-15 that for the current case study, SOM clustering provides the best clustering method to create homogeneous groups. This thesis also considered the IRI data from previous years to reveal the relationships or changes in results.

The results of K-means clustering were controversial when considering the IRI of the previous year. In the first cases of clustering hiding the 2015 IRI as an additional feature allowed to identify the resilience of clustering techniques. Moreover, it became clear that additional feature did not alter the results of prediction models, however, in previous Markovian PMS prediction models previous years features had strong weight. The summary and comparison of the results of clustering methods are briefly discussed below. Thus, in the first case, the configuration of K-means clustering was as follows:

- Clustering range to identify the optima: 3-50
- Attributes were scaled individually.

- Attributes Selected: 2017 IRI Data; Traffic (ESALS); Precipitation (mm/year); Deflection Basin Area; Center Deflection value

As a result, the optimal number of clusters was seven (7) with the CCC (Cubic Clustering Criterion) of 12.7. The Parallel coordinates plot of this clustering is shown in Figure 4-14.

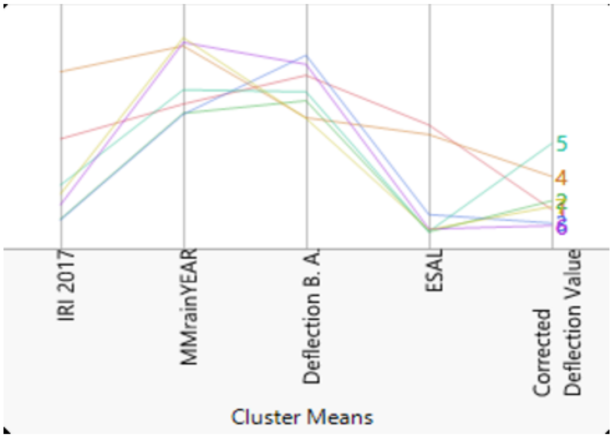


Figure 4-14. The parallel coordinate plot of the first case of K-means Clustering.

When the IRI data from previous years is included as a predictor, the number of optimal clusters increases to 44 groups with the CCC (Cubic Clustering Criterion) of 33.67 (results are shown in Figure 4-15). This clearly indicates that K-means clustering is unable to group the elements having the same properties as it is using the hard clustering method.

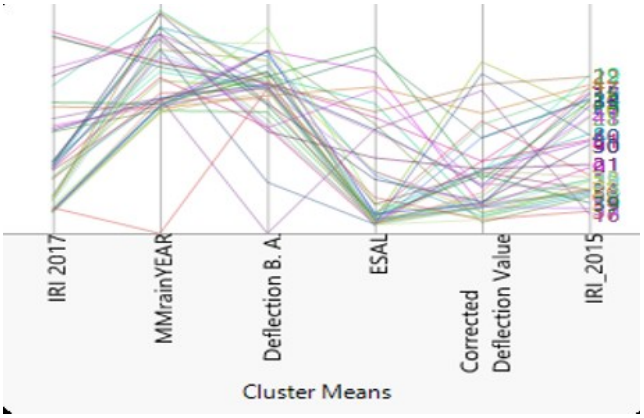


Figure 4-15. The parallel coordinate plot of the second case of K-means clustering with 2015 IRI data.

Figure 4-16 depicts the comparison of biplots for both clustering cases and clearly indicates the inadequacy of K-means clustering in this case study. Biplots of clustering were varying significantly because K-means is a hard-clustering method. Therefore, K-means clustering fails to decide which cluster the elements of groups must be included when the values share the exact same properties. However, the soft clustering method calculates the distance of the data point to the cluster centroids based on the percentage and can even include the exact same values in both clusters. Thus, when considering the jump in the optimal number of clusters from 7 to 44 in K-means analysis, the resilience of soft clustering against the addition of a feature marginally outperforms the K-means clustering.

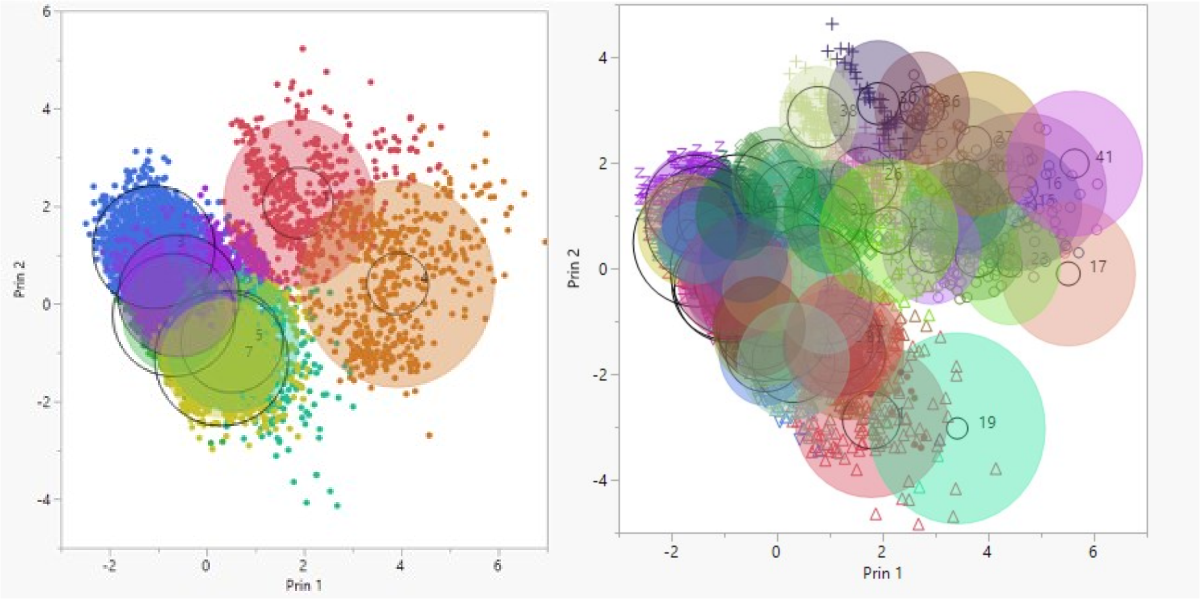


Figure 4-16. Bi-plot comparison of the first (left) and the second (right) cases of K-means clustering.

On the other hand, SOM clustering is using soft clustering as mentioned in the previous sections and it can better cluster the data on an imaginary plane than other clustering methods. Clustering configurations for the first case of the SOM used in this study is given below:

- Clustering range: 3-50

- Attributes were scaled individually.
- 3x3 SOM grid is utilized in both cases.
- Attributes involved:
  - 2017 IRI Data
  - Traffic (ESALS)
  - Precipitation (mm/year)
  - Deflection Basin Area
  - Center Deflection value

As a result, in the first case, the optimum number of clusters was eight (8) with the CCC value of 6.47.

In the second case, IRI IRI of the previous year was included as an additional attribute and the optimal number of clusters was still calculated as eight (8) with the CCC value of 3.97. The results of the parallel coordinate plots of both cases were similar to each other and graphs are shown in Figure 4-17.

Thus, it indicates that the SOM clustering is yielding more stable and reliable homogeneous groups for this specific case study.

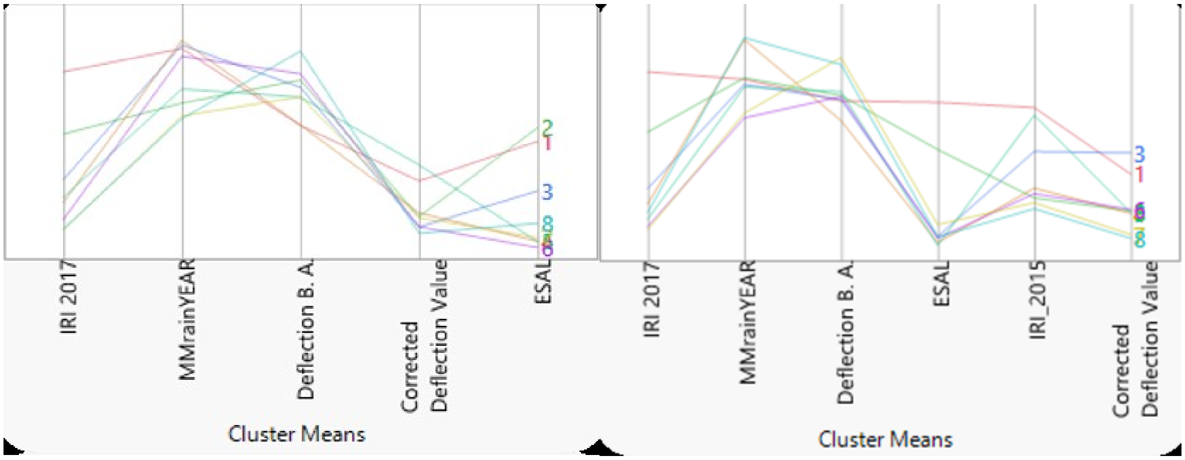
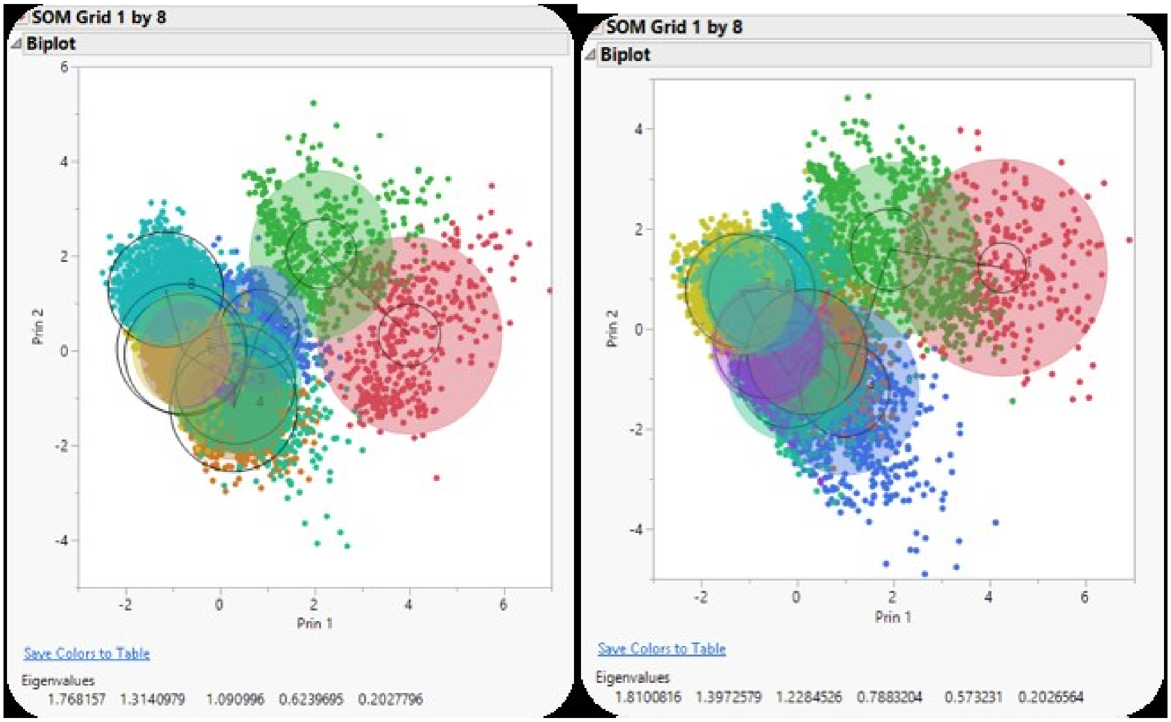


Figure 4-17. Comparison of parallel coordinate plots of the first (left) and the second (right) cases of SOM clustering.

In the end, from Figure 4-17 and Figure 4-18 it is obvious that SOM clustering is yielding more stable and reliable homogeneous groups. The number of clusters, distribution of points, behavior of centroids and the range of the clusters were very similar. Thus, it can be concluded from this comparison that SOM performs clustering more properly in this particular study.



```
Eigenvalues
   1.768157   1.3140979   1.090996   0.6239695   0.2027796
```

```
Eigenvalues
   1.8100816   1.3972579   1.2284526   0.7883204   0.573231   0.2026564
```

Figure 4-18. Comparison of bi-plots of the first (left) and the second (right) cases of the SOM clustering.

Moreover, the outcomes of the extrinsic analysis measuring the goodness of the clustering using three different prediction models (Neural Networks, Decision tree, and Logistic Regression) were indeed able to highlight in Figure 4-19 that neglecting the previous year's IRI data in SOM Clustering affected the results in a positive way while generating homogeneous groups. As a result of the clustering, this study was able to emphasize the impact of each attribute (traffic, rain, deflection and deflection basin area) on homogeneous grouping. Each prediction model had its strengths and depending on the purpose of the analysis each model had unique outcomes. However, the most reliable model for this particular case study was the neural networks model where 10 step boosting was used with two hidden layers shown in Figure 4-19. The comparison of the R^2 values of all models used in the extrinsic evaluation is illustrated in Figure 4-19 as well. It is possible to conclude from this table that the best option to create the homogeneous groups is excluding the previous year's data from the analysis and using the SOM Clustering technique, whose soft-clustering approach helps to distribute the points better on the map.

Following the above results, SOM clustering is considered as a base clustering method to create homogeneous groups and the whole Costa-Rican roads network (the current case study) was mapped illustrating road sections and corresponding groups in Figure 4-20.

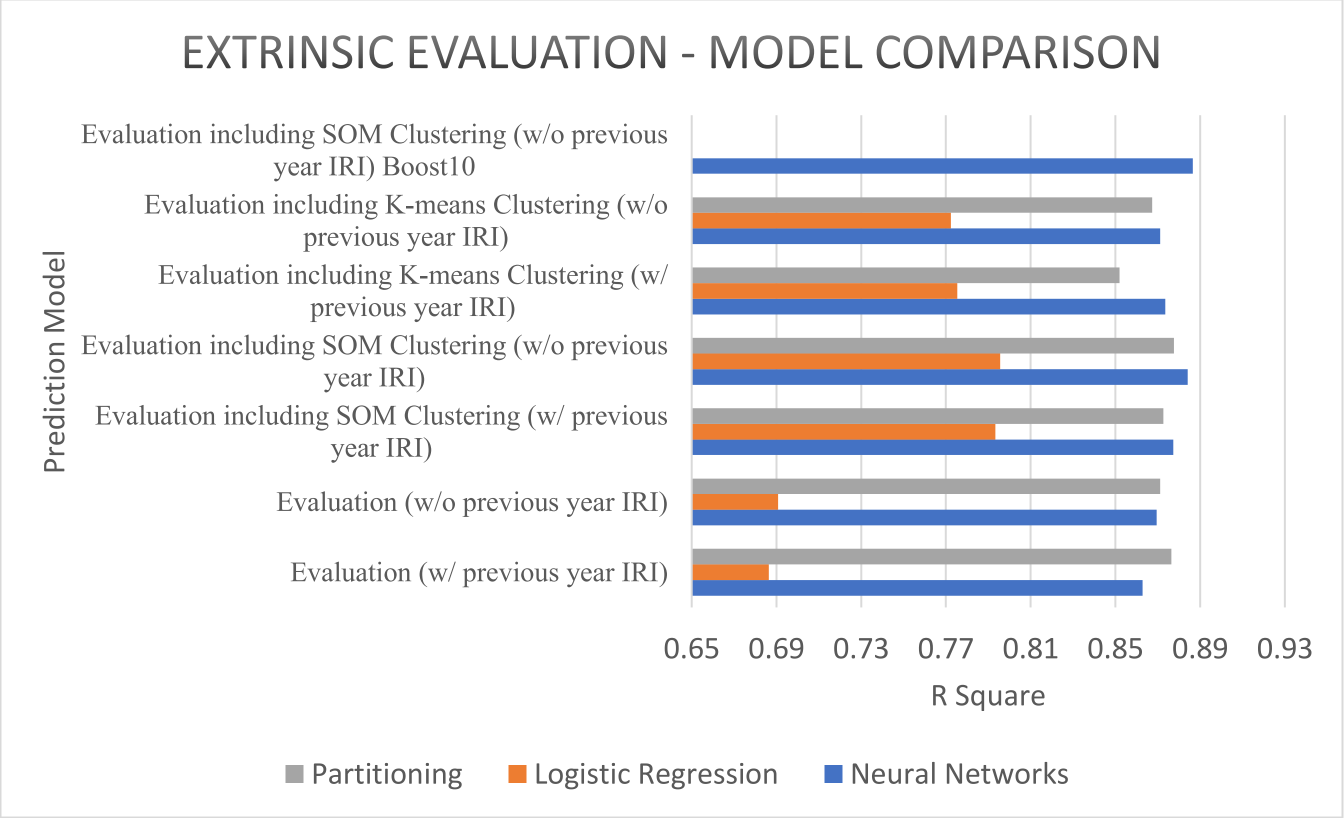


Figure 4-19. Comparison of all extrinsic evaluation models.

Table 4-16. Confidence interval of the best prediction models.

Confidence Interval	Logistic Regression	Partitioning	Neural Networks
99% confidence interval:	0.78587 ≤ R2 ≤ 0.80533	0.87022 ≤ R2 ≤ 0.88258	0.87839 ≤ R2 ≤ 0.89001
95% confidence interval:	0.78819 ≤ R2 ≤ 0.80301	0.87170 ≤ R2 ≤ 0.88110	0.87978 ≤ R2 ≤ 0.88862
90% confidence interval:	0.78939 ≤ R2 ≤ 0.80181	0.87246 ≤ R2 ≤ 0.88034	0.88049 ≤ R2 ≤ 0.88791

In addition to the comparison of the extrinsic evaluation models confidence intervals for three prediction models are given in Table 4-16. It is obvious from the table that the prediction model built using the artificial neural networks is the best responding model within 99% confidence interval with the R^2 range of 0.87-0.89.

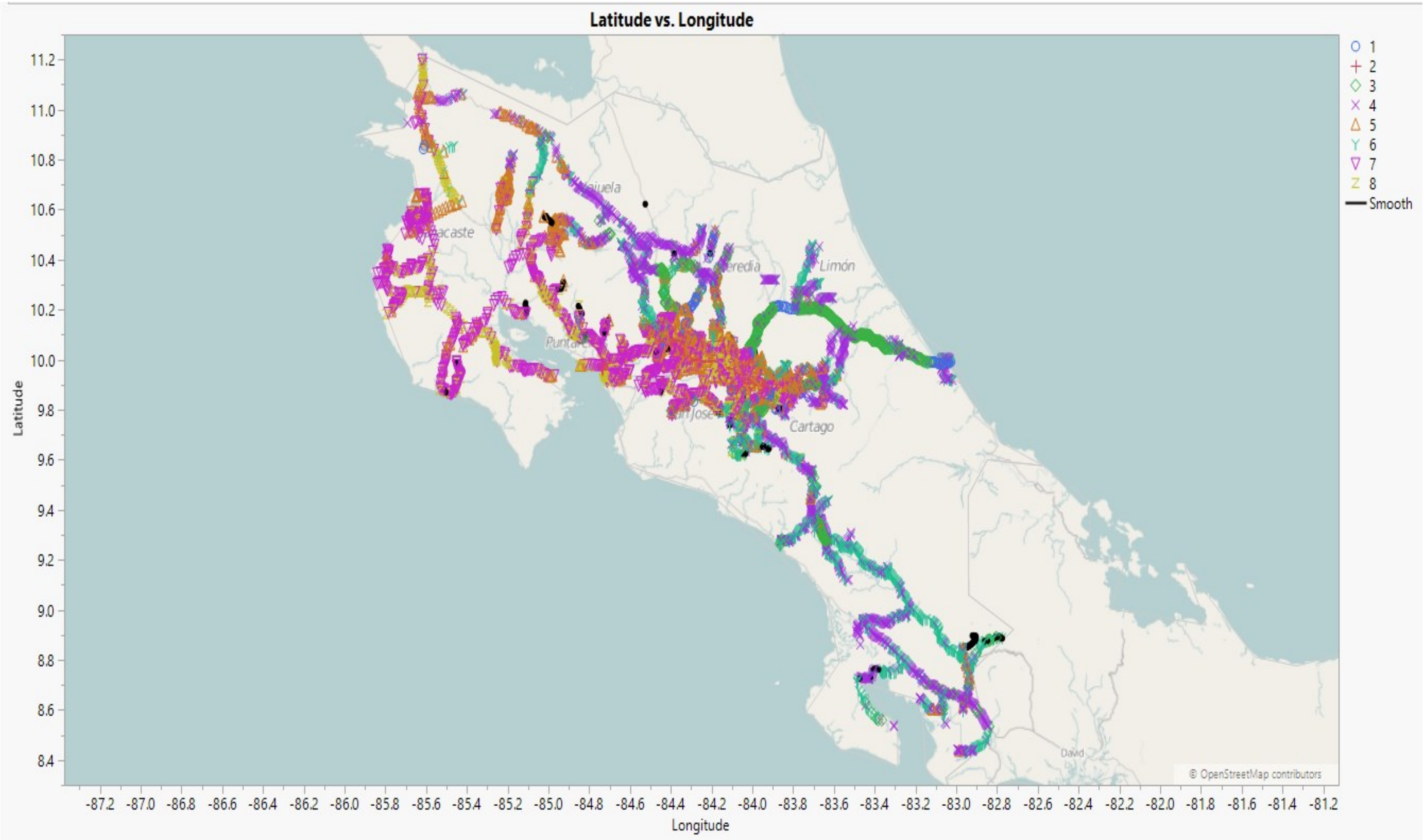
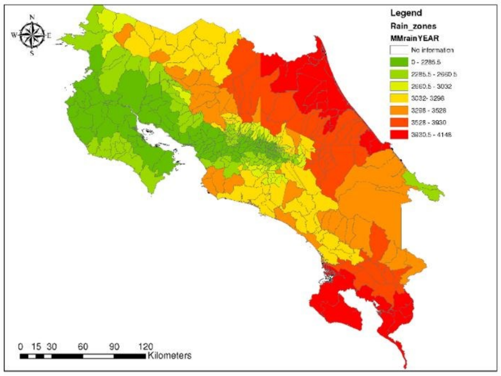
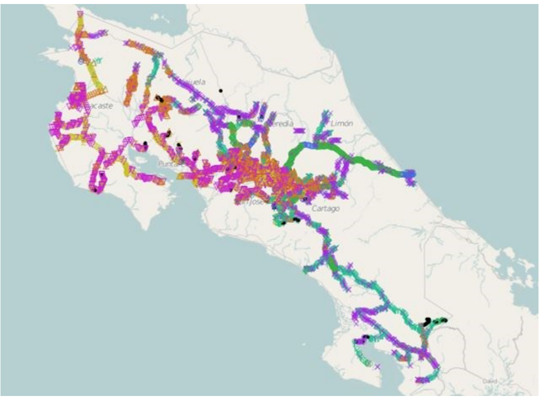
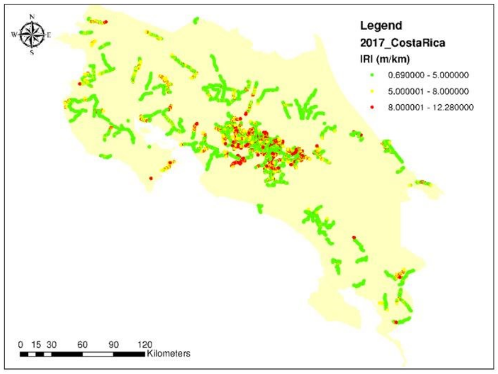


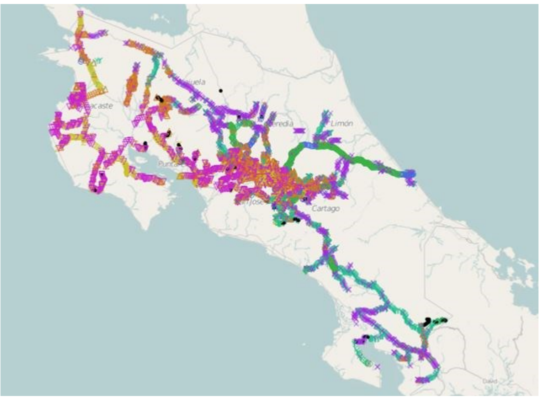
Figure 4-20. Map of the Costa-Rica illustrating similarly acting homogeneous groups of pavements.

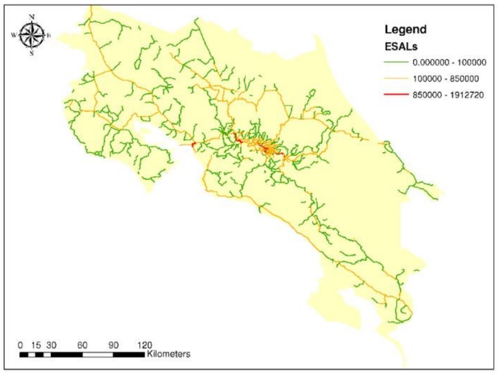
In conclusion, it is possible to emphasize that in order to create homogeneous groups based on pavement deterioration considering the daily factors affecting its conditions, the outcome of SOM clustering is satisfactory and this method can be taken as a basis for further future researches to polish and build a new decision making-system for the pavement management systems.

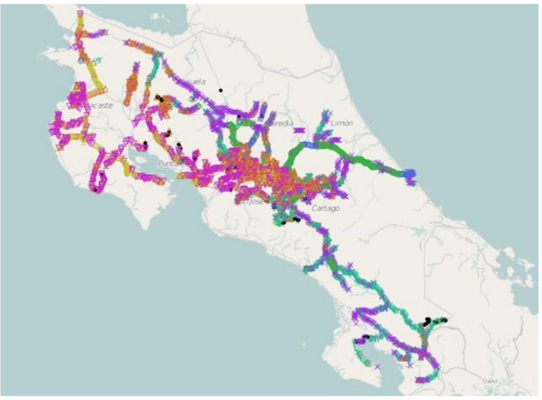












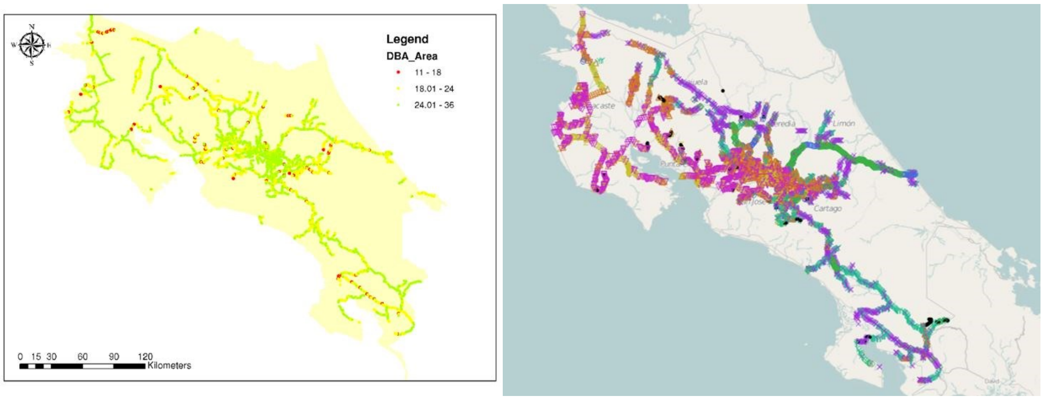


Figure 4-21. Correlation of attributes with clustering.

Additionally, it can be concluded from the visual correlation of an each attribute with the clustering shown in Figure 4-21 that the precipitation has the most impact on homogeneous groups.

Chapter 5

5. Conclusion and future research

The overall goal of this research was to develop an approach that is capable of clustering given datasets of deteriorating daily factors such as climate, traffic loading and the current condition of pavement based on its performance into various homogenous groups. This in return allows understanding the correlation between daily factors and their effects on the pavement performance.

As a result of the case study, it is possible to conclude that daily factors like precipitation, deflection and traffic load have a significant impact on pavement deterioration and creating homogeneous groups provides reasonable results. Thus, the latest rendered map given in Figure 4-18 (created based on the SOM clusters) illustrates the successful clustering of the road condition based on traffic and weather conditions.

The case study reported in this thesis proposes two different methods to analyze the impact of daily factors on pavement deterioration. One of the main methods implemented here was SOM clustering which provided better results than the K-means clustering method. One of the most influential factors affecting the deterioration was traffic loading. Overall, the deflection was the second most influential factor whereas the precipitation was third on the list. Here, it is proved that weather is one of the factors affecting the deterioration of the pavement where road maintenance has not been done for more than 20 years. In fact, around the capital higher values of AADT and precipitation have put the whole area into one cluster which required even further research in the future. The second method (K-means) which uses hard clustering was rather weaker than SOM clustering because the amount of the data collected couldn’t be processed as some points were overlapping.

5.1 List of findings

The major findings of this thesis can be summarized as follows.

- K-means clustering failed to identify optimal clusters when an additional feature is introduced into the data frame.
- Self-organizing maps were instrumental to cluster pavements into homogeneous groups.
- Addition of previous year’s pavement condition as an attribute didn’t alter the results of prediction models despite previous Markovian models claiming the opposite.
- Spatial visualization suggested that precipitation plays an important role in this case study.

5.2 Major contribution of this thesis

The main contributions of this thesis include:

- “Hard clustering” method such as k-means algorithm is vulnerable to feature addition.
- IRI predictors were trained through three techniques.
- Performance (goodness) of two clustering methods was tested through intrinsic and extrinsic evaluations.
- Final interactive database was generated. Thus, model will automatically generate homogeneous groups when new data is inputted.

5.3 Significance and impact

Previous studies did not develop a method to detect homogeneous groups considering pavement deterioration. The method suggested by this thesis will help to easily decide on what group or category the road section falls into by inputting the measured data. Homogeneous groups can help to decide on what type of rehabilitation method must be selected for a section of the road at first glance.

The LANAMME study concluded that the government lacks a coherent policy for the maintenance or improvement of the country's roadway network. According to the government reports during the past two years, CONAVI allocated $24.6 million for repairs along 907 Km (564 miles) of roads "without achieving the expected results" (CONAVI, 2019). That figure represents 11.5 percent of the $213.7 million invested in maintaining all the country's paved roads for a year. Considering these issues creating a method to evaluate the pavement systems becomes important.

5.4 Limitations of this thesis work

As in any data-driven research, the dataset is one of the most important aspects to consider. In this thesis integrity of datasets was one of the main challenges. All datasets were collected from different sources closed to the public in different file formats. In order to be able to cluster dataset into homogeneous groups, every affecting factor must be collected from the same location. Converting separate data files and merging them all in one file was challenging, however, cleaning the dataset was even more problematic because different indicators like IRI or FWD had different measuring distances as one of them collected the data on the roads every 100 meters and the other every 200 meters which led to cleaning of almost half of the final database. Thus, the lack of pavement condition indicator and affecting factor's measurements in some parts of the country and/or roads was one of the main limitations of the thesis work.

Moreover, besides the size of the datasets and difficulties in merging them, necessary attributes affecting the deterioration of the pavement such as the age of the pavement, rutting data, drainage data were missing.

5.5 Future research

Further research can take this topic one step further by creating a decision-making software based on homogeneous groups. This can assist engineers and consulting companies in choosing the most reasonable reconstruction and rehabilitation methods on certain sections of the roads that correspond to a certain homogeneous group. However, it is also necessary to include other factors like the age of the pavement, condition indicator, drainage, etc. in future research papers because these features will alter the results of the clustering. Additionally, some anomalies in homogeneous groups can be observed in Figure 4-20. These inconsistencies occur because different pavement structures exist in the same zones. Thus the material, structure and thickness of the pavement must be included in future researches to eradicate these anomalies.

References

AASHTO. (1993). Guide for Design of Pavement Structures, Volume I. Washington DC.

Amandeep, K. M., & Navneet, K. M. (2013). Review Paper On Clustering Techniquesî. Global Journal Of Computer Science And Technology Software & Data Engineering, VOL 13, 201.

Arat, M. M. (2019, September 1). Implementing K-means Clustering from Scratch - in Python. Retrieved from mmuratarat.github.io: https://mmuratarat.github.io/2019-07-23/kmeans_from_scratch

ASHGHAL, Q. i. (2016). iRAP Coding Manual - Qatar. iRAP.

Attoh-Okine, N. (1994). Predicting roughness progression in flexible pavements using artificial neural networks.

Barai, S. K. (2003). Data Mining applications in transportation engineering. Kharagpur: Indian Institute of Technology .

Ceylan, H., Bayrak, M. B., & Gopalakrishnan, K. (2014). Neural Networks Applications in Pavement Engineering: A Recent Survey.

Charles, V. M. (2017). Investigation Of Deflection Basins To Identify Structural Distresses Within Flexible Pavements. Auburn: Auburn University.

Chatti, K., Kutay, M. E., Lajnef, N., Zaabar, I., Varma, S., & Lee, H. S. (2017). Enhanced Analysis of Falling Weight Deflectometer Data for Use With Mechanistic-Empirical Flexible Pavement Design and Analysis and Recommendations for Improvements to Falling Weight Deflectometers. FHWA.

Choudhury, A. (2019, December 15). Self Organising Maps on IMBD Movie Covers. Retrieved from Medium: https://blog.usejournal.com/self-organising-maps-on-imbd-movie-covers-a2e666a06052

ClimateData. (2019, December 14). Costa Rica Climate. Retrieved from climate-data.org: https://en.climate-data.org/south-america/bolivia/beni/costa-rica-999068/

CONAVI. (2019, December 14). Conavi Vialidad. Retrieved from Conavi Vialidad: https://conavi.go.cr/wps/portal/CONAVI

Dae, Y. K., Chi, S., & Kim, J. (2018). Selecting Network-Level Project Sections for Sustainable Pavement Management in Texas. Busan: Pusan National University.

Dickey. (2015, December 15). Analytics, Datamine, CCC. Retrieved from Extra Notes: https://www4.stat.ncsu.edu/~dickey/Analytics/Datamine/Extra_Notes/CCC.pdf

Dudoit, S. &. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology, 1-21.

Fagerland, M. &. (2013). A goodness-of-fit test for the proportional odds regression model. Statistics in medicine. 32.10.1002/ sim.5645.

Fayyad, U. &. (1996). Data Mining and Knowledge Discovery in Databases. Communications of the ACM, 24-26.

FEMA. (2003). Guidelines and Specifications for Flood Hazard Mapping Partners.

FHWA. (2018). Validation of Pavement Performance Measures Using LTPP Data: Final Report - FHWA-HRT-17-089 .

Gryp, v. d., Bredenhann, A., Henderson, S., & Rohde, M. (1998). Determining the visual condition index of flexible pavements using artificial neural networks.

Guide, A. (1993). AASHTO Guide. AASHTO.

Haas. (1997). Pavement design and management guide. Ottawa: Transportation Association of Canada.

Hagan, D. B. (1999). Neural Network Design. Campus Pub. Service.

Hajek, J. J., Selezneva, O. I., Mladenovic, G., & Jiang, Y. J. (2005). Estimating Cumulative Traffic Loads,Volume II: Traffic Data Assessment and Axle Load Projection for the Sites with Acceptable Axle Weight Data, Final Report for Phase 2. McLean: FHWA.

Heaton, J. (2019, December 15). The Number of Hidden Layers. Retrieved from Heaton Research: https://www.heatonresearch.com/2017/06/01/hidden-layers.html

Heidari, M. J., Najafi, A., & Alavi, S. (2018). Pavement Deterioration Modeling for Forest Roads Based on Logistic Regression and Artificial Neural Networks. Croat. j. for. eng., 271–287.

Holdaway, K. R. (2014). Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data-Driven Models. Haboken: John Wiley & Sons, Inc.

Hu, J. (2013). Research on comfort and safety Treshold of Pavement Roughness.

Hussein, A. S. (2015). The Effect of Pre-processing Techniques and Optimal Parameters on BPNN for Data Classification.

JMP. (2019, 02 01). The Model Comparison Report. Retrieved from JMP Statistical Discovery (From SAS): https://www.jmp.com/support/help/14-2/the-model-comparison-report.shtml

Johnson, N. (1949). Systems of frequency curves generated by methods of translation. Biometrika, 149–176.

King, M. B.-A. (2014). The Effect of Road Roughness on Traffic Speed and Road Safety.

Kohavi, R. (2000). Data Mining and Visualization.

Kohonen, T. (1990). The Self-Organizing Map. IEE.

Kulkarni, R. B. (2003). Pavement Management Systems: Past, Present, and Future. Journal of the Transportation Research Board, no. 1853, 65–71. .

LANAMME. (2019, December 14). Laboratorio Nacional de Materiales Y Modelos Estructurales. Retrieved from Universidad de Costa Rica: https://www.lanamme.ucr.ac.cr/

Lea, J. (2004). Data Mining of the Caltrans Pavement Management System (PMS) Database.

Lee, H. N. (1996). Development of geographic information system-based pavement management system for Salt Lake City. Transp. Res. Rec., , 1524, 16–24.

Li, Q., Qiao, F., & Yu, L. (2016). Clustering Pavement Roughness Based On the Impacts on Vehicle Emissions and Public Health. Journal of Ergonomics.

Lin, J. Y. (2003). Corrolation analysis between international roughness index (IRI) and pavement distress by neural network. .

Madanat, S., Nakat, Z., & Sathaye, N. (2005). Development of Empirical-Mechanistic Pavement Performance Models using Data from the Washington State PMS Database.

Mathavan, S. (2014). Use of a Self-Organizing Map for Crack Detection in Highly Textured Pavement Images. Journal of Infrastructure Systems.

MDOT, C. (2016). Performance-Based Planning And Programming For Pavement Management. Ann Arbor: MDOT.

Medina, A. G. (1999). "Geographic Information Systems-Based Pavement Management System: A Case Study,". Washington D.C: National Research Council.

MOPT. (2019, December 15). MOPT - Pagina Principal. Retrieved from MOPT: https://www.mopt.go.cr/wps/portal/Home/inicio/!ut/p/z1/hY7LDoIwEEW_hQVbZihojLuSKFFJfLAQuyFgasEUSkqF35fgCoNxdvfmnNwBBgmwOutKkZlS1Zkc8o0t090pRPfgYxS6WxfpmWIcb4iHxIfrCOCPowjsn89G5NsLLiTwEMMjmQUmE3tgQqr88y6tc28lgGn-4Jpr56WHujCmadc22tj3vSOUEpI7d1XZOKcUqjWQTElo

Morrison, R. E., Bryant, C. M., Terejanu, G., Prudhomme, S., & Miki, K. (2013). Data partition methodology for validation of predictive models.

Mukhtarli, K. (2019, December 15). Gofile. Retrieved from Gofile: https://gofile.io/?c=6jymkR

Neelamegam, S., & Dr.Ramaraj, E. (2013). Classification algorithm in Data mining: An Overview. Karaikudi: IJPTT.

Özkan, C. &. (2003). The comparison of activation functions for multispectral Landsat TM image classification. Photogrammetric Engineering & Remote Sensing, 1225-1234.

Panerati, J., Schnellmann, M. A., Patience, C., Beltrame, G., & Patience, G. S. (2019). Experimental methods in chemical engineering: Artificial neural networks–ANNs. The Canadian Journal of Chemical Engineering, https://doi.org/10.1002/cjce.23507.

Perera, S. M. (1999). Guidelines for Longitudinal Pavement Profile Measurement . University of Michigan.

Pierce, L. M. (2014). Quality Management For Pavement Condition Data Collection.

Powers, D., & Xie, Y. (2008). Statistical Methods for Categorical Data Analysis, 2nd ed. In D. Powers, & Y. Xie, Statistical Methods for Categorical Data Analysis, 2nd ed. (pp. 31–66). Bingley, UK: Emerald Group.

Rifai, A. I., Sigit, P. H., Correia, A. G., & Pereira, P. (2015). The Data Mining Applied For The Prediction Of Highway Roughness Due To Overloaded Trucks. International Journal of Technology.

SAS. (2019, December 15). Cubic Clustering Criterion. Retrieved from SAS Enterprise Miner 14.3: https://documentation.sas.com/?docsetId=emref&docsetTarget=n1dm4owbc3ka5jn11yjkod7ov1va.htm&docsetVersion=14.3&locale=en

SAS. (2019, December 15). Distance Measures. Retrieved from JMP Statistical Discovery TM, From SAS: https://www.jmp.com/support/help/14-2/distance-measures.shtml#230090

SAS. (2019, December 15). Hidden Layer Structure. Retrieved from JMP Statistical Discovery From SAS: https://www.jmp.com/support/help/14-2/hidden-layer-structure.shtml

Sayers, M. K. (1995). The Little Book of Profiling. UMTRI. In M. Sayers, On the Calculation of IRI from Longitudinal Road Profile.

Sibi, M. P. (2005). Enantioselective addition of nitrones to activated cyclopropanes. Journal of the American Chemical Society, , 5764-5765.

SS. (2019, January 1). What is Logistic Regression? Retrieved from Statistics Solutions: https://www.statisticssolutions.com/what-is-logistic-regression/

Tan, F., Bao, H., & Dong, K. (2007). Dynamic response of concrete pavement structure with asphalt isolating layer under moving loads. Journal of China and Foreign Highway, 202-205.

Taylor, J. (n.d.). Introduction to Regression and Analysis of Variance. Retrieved from http://statweb.stanford.edu: http://statweb.stanford.edu/~jtaylo/courses/stats203/notes/robust.pdf

Ting-Wu Ho, C.-C. C.-T.-D. (2010). Pavement distress image recognition using k-means and classification algorithms. Proceedings of the International Conference on Computing in Civil and Environmental Engineering. Nottingham: Nothingham University Press.

TowardsDataScience. (2019, December 15). activation-functions-and-its-types. Retrieved from towardsdatascience.com: https://towardsdatascience.com/activation-functions-and-its-types-which-is-better-a9a5310cc8f

TradingEconomics. (2019, December 15). Costa Rica Average Precipitation. Retrieved from tradingeconomics.com: https://tradingeconomics.com/costa-rica/precipitation

Tuamsuk, K., & Silwattananusarn, T. (2012). Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012. International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol 2.

UCR. (2019, December 14). Universidad de Costa Rica. Retrieved from Universidad de Costa Rica: https://www.ucr.ac.cr/

Unistat. (2019, December 16). Quality control data transformation. Retrieved from UNISTAT Statistical Software: https://www.unistat.com/guide/quality-control-data-transformation/#u9371

V. Sunitha, A. V. (2012). Cluster-Based Pavement Deterioration Models for Low-Volume Rural Roads.

Van Dam, T. J., Harvey, J. T., Muench, S. T., Smith, K. D., Snyder, M. B., Al-Qadi, I. L., . . . Kendall, A. (2015). Towards Sustainable Pavement Systems: A Reference Document. Urbana: FHWA.

Wang, K. &. (2010, 01). Gray Clustering-Based Pavement Performance Evaluation. Journal of Transportation Engineering-asce - J TRANSP ENG-ASCE, 136, . doi:10.1061/(ASCE)0733-947X(2010)136:1(38)

Wang, M., & Rennolls, K. (2005). Tree diameter distribution modelling: Introducing the logit logistic distribution. Can. J. For. Res, 1305–1313.

Witten, I. H., & Frank, E. (2000). Chapter 8: Tutorial. In E. F. Ian H. Witten, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers.

Xu, G., Bai, L., & Sun, Z. (2014). Pavement Deterioration Modeling and Prediction for Kentucky Interstate and Highways. Louisville: University of Louisville.

Zaniewski, A. S., Hossain, M. M., & John, P. (1991). Characterization of Falling Weight Deflectometer Deflection Basin.

Zhou, G. (2011). Co-Location Decision Tree for Enhancing Decision-Making of Pavement Maintenance and Rehabilitation. Blacksburg: Virginia Polytechnic Institute and State University.