5-1-2017

Generalized Clusterwise Regression for Simultaneous Estimation of Optimal Pavement Clusters and Performance Models

Mukesh Khadka
University of Nevada, Las Vegas, imukeshkhadka@gmail.com

GENERALIZED CLUSTERWISE REGRESSION FOR SIMULTANEOUS ESTIMATION OF OPTIMAL PAVEMENT CLUSTERS AND PERFORMANCE MODELS

By

Mukesh Khadka

Bachelor of Engineering in Civil Engineering,
Institute of Engineering, Tribhuvan University, Nepal
2007

Master of Engineering in Construction, Engineering and Infrastructure Management
Asian Institute of Technology, Thailand
2011

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy - Civil and Environmental Engineering

Department of Civil and Environmental Engineering and Construction
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
May 2017

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

January 18 ,2017

This dissertation prepared by

Mukesh Khadka

entitled

Generalized Clusterwise Regression for Simultaneous Estimation of Optimal Pavement Clusters and Performance Models

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Civil and Environmental Engineering
Department of Civil and Environmental Engineering and Construction

Alexander Paz, Ph.D.
Examination Committee Chair

Mohamed Kaseko, Ph.D.
Examination Committee Member

Moses Karakouzian, Ph.D.
Examination Committee Member

Pramen P. Shrestha, Ph.D.
Examination Committee Member

Ashok K. Singh, Ph.D.
Graduate College Faculty Representative

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Interim Dean

ABSTRACT

Generalized Clusterwise Regression for Simultaneous Estimation of Optimal Pavement Clusters and Performance Models

by

Mukesh Khadka

Dr. Alexander Paz, Examination Committee Chair
Associate Professor, Civil and Environmental Engineering and Construction
University of Nevada, Las Vegas

The existing state-of-the-art approach of Clusterwise Regression (CR) to estimate pavement performance models (PPMs) pre-specifies explanatory variables without testing their significance; as an input, this approach requires the number of clusters for a given data set. Time-consuming ‘trial and error’ methods are required to determine the optimal number of clusters. A common objective function is the minimization of the total sum of squared errors (SSE). Given that SSE decreases monotonically as a function of the number of clusters, the optimal number of clusters with minimum SSE always is the total number of data points. Hence, the minimization of SSE is not the best objective function to seek for an optimal number of clusters.

In previous studies, the PPMs were restricted to be either linear or nonlinear, irrespective of which functional form provided the best results. The existing mathematical programming formulations did not include constraints that ensured the minimum number of observations required in each cluster to achieve statistical significance. In addition, a pavement sample could be associated with multiple performance models. Hence, additional modeling was required to combine the results from multiple models.

To address all these limitations, this research proposes a generalized CR that simultaneously 1) finds the optimal number of pavement clusters, 2) assigns pavement samples

into clusters, 3) estimates the coefficients of cluster-specific explanatory variables, and 4) determines the best functional form between linear and nonlinear models. Linear and nonlinear functional forms were investigated to select the best model specification. A mixed-integer nonlinear mathematical program was formulated with the Bayesian Information Criteria (BIC) as the objective function. The advantage of using BIC is that it penalizes for including additional parameters (i.e., number of clusters and/or explanatory variables). Hence, the optimal CR models provided a balance between goodness of fit and model complexity. In addition, the search process for the best model specification using BIC has the property of consistency, which asymptotically selects this model with a probability of '1'.

Comprehensive solution algorithms – Simulated Annealing coupled with Ordinary Least Squares for linear models and All Subsets Regression for nonlinear models – were implemented to solve the proposed mathematical problem. The algorithms selected the best model specification for each cluster after exploring all possible combinations of potentially significant explanatory variables. Potential multicollinearity issues were investigated and addressed as required.

Variables identified as significant explanatory variables were average daily traffic, pavement age, rut depth along the pavement, annual average precipitation and minimum temperature, road functional class, prioritization category, and the number of lanes. All these variables were considered in the literature as the most critical factors for pavement deterioration.

In addition, the predictive capability of the estimated models was investigated. The results showed that the models were robust without any overfitting issues, and provided small prediction errors. The models developed using the proposed approach provided superior explanatory power compared to those that were developed using the existing state-of-the-art approach of clusterwise regression. In particular, for the data set used in this research, nonlinear models provided better explanatory power than did the linear models. As expected, the results illustrated that different clusters might require different explanatory variables and associated coefficients. Similarly, determining the optimal number of clusters while estimating the corresponding PPMs contributed significantly to reduce the estimation error.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Alexander Paz for his guidance and encouragement through my study. His technical and editorial advice was essential for this research. This dissertation would never have been possible without his supervision.

I would like to extend my sincere gratitude and appreciation to my dissertation committee members, Dr. Mohamed Kaseko, Dr. Moses Karakouzian, Dr. Pramen P. Shrestha, and Dr. Ashok Singh, for their constructive and invaluable suggestions to finalize my work.

Many thanks to the Nevada Department of Transportation for providing the research data. Special thanks to Julie Longo, Technical Writer at UNLV's Howard R. Hughes College of Engineering, her help reviewing this manuscript. Many thanks to Cristian, Antoine, and Fabien for their help with the implementation of the solution algorithms. I sincerely thank my colleagues in the Transportation Research Center, Naveen, Victor, and Daniel for their help and friendship. It was pleasure working with you all. Thanks to the Nepalese community in Las Vegas for making me feel at home.

Finally, I would like to thank all my family members for their love and blessings. No words can express my gratitude to my parents, parents-in-laws, and brother for their continuous support and inspiration to work harder. I would also like to express my deep appreciation and gratitude to my wife, Dr. Sichu Shrestha, for being patient and supportive throughout this journey. To my daughter, Aadya, thank you for cheering me up even after a long work with your sweet smile. To my nephew, Aryav, thank you for your sweet word "kaaku". Last but not least, I would like to remember Dr. Sinora Shrestha and Dr. Biswas Joshi, who were the first persons I met in USA, for your warm and dreamful welcome for my bright and successful career.

DEDICATION

To my parents for their dream and unconditional love.

To my wife and daughter for their unconditional love, motivation, support, and continuous encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

In practice, it is very important to achieve a balance among the number of Pavement Performance Models (PPMs); the number of explanatory variables; the resources required to develop, maintain, and use these models; and the associated explanatory power. To seek this balance, PPMs typically are developed using clusters of pavement segments. A few predefined explanatory variables are used to assign pavement segments into clusters, instead of estimating the cluster memberships using statistical methods and testing for significance. In terms of performance or deterioration, the clusters thus formed are likely to include heterogeneous pavement segments.

The existing state-of-the-art proposes Clusterwise Linear Regression (CLR) to determine, simultaneously, pavement clusters and associated PPMs using a single objective function. In CLR, different clusters are formed such that segments assigned within a cluster are homogenous in terms of the effects of the explanatory variables on pavement performance (Park et al. 2015). That is, the homogeneity of pavement segments in a cluster is defined by similarities of the observed values of both dependent and explanatory variables and largely by the proximity of segments with respect to an underlying PPM (Preda and Saporta 2007). Hence, observations of all pavement segments assigned to a cluster fit the same PPM with a minimum estimation error.

In the field of pavement management, to the best knowledge of the author, only three studies (Luo and Chou 2006, Luo and Yin 2008, Zhang and Durango-Cohen 2014) have used CLR. These studies have revealed the advantages of pavement performance modelling when using the CLR framework over other available methods. However, the existing literature suffer from a few serious limitations.

- First, the number of clusters is a pre-specified input for CLR. However, it is impossible to know *a priori* the optimal number of clusters that minimizes the estimation error. Time-consuming ‘trial and error’ methods as well as an extensive sensitivity analysis are required to determine this number.
- Second, previous mathematical programs used the minimization of the sum of squared errors (SSE) as the objective function. By increasing the number of clusters, the number of parameters is increased, which translates into a smaller SSE.
- Third, the existing literature assumes that all explanatory variables are significant. Assignment of pavement segments into clusters using predefined and fixed explanatory variables introduces bias into the statistical analysis (Gupta and Ibrahim 2007). In addition, clustering using explanatory variables that do not provide any information about the underlying clustering structure does not reveal the true cluster assignments. This illustrates the negative consequences of assuming the significance of variables.
- Fourth, previous mathematical programs did not include constraints to restrict a minimum number of observations required in a cluster to achieve statistical significance.
- Fifth, in previous studies, a pavement sample could be associated with multiple performance models. Hence, an additional modeling was required to combine the results from multiple models.
- Sixth, the existing literature assumes either linear or nonlinear functional forms to estimate PPMs. Potential functional forms were not investigated to select the one with less estimation errors.

1.2 Contributions of the Dissertation

To address all the above limitations, this dissertation proposed a generalized mathematical programing formulation and a solution algorithm to determine simultaneously 1) the optimal number of clusters and the associated cluster members (i.e., pavement samples), 2) the cluster-specific significant explanatory variables and the associated coefficients, and 3) the best functional form between linear and nonlinear models. The proposed mathematical program used the Bayesian Information Criteria (BIC), which penalizes the inclusion of additional model parameter, as the objective function. Thus, the optimal number of models with minimum BIC provided balance between the goodness of fit and model complexity. To achieve statistical significance, the mathematical program included constraints that ensured the minimum number of observations required in each cluster. In addition, the mathematical program was formulated such that all observations of a sample were assigned to the same cluster exclusively. Both, linear and nonlinear functional forms were estimated within the proposed clusterwise regression framework; the resultant models were then compared to select the one with less estimation errors for the data used in this dissertation.

The mathematical program included two constraints to prevent a search beyond a feasible number of clusters. A solution algorithm was proposed and implemented to determine the upper bound (maximum) of the potential number of clusters. The minimization of BIC was used as the objective function. Minimizing BIC reduces unexplained variation in the dependent variable. The search process for the best model specification using BIC has the property of consistency, which asymptotically selects this model, with probability one. The mathematical program and solution algorithm explored all possible combinations of explanatory variables to seek for the best model specification that provides the explanatory power and is free of potential multicollinearity issues.

To avoid issues about samples associated with multiple clusters, the proposed mathematical program included constraints to ensure a pavement sample and all associated observations were assigned to one cluster exclusively.

To investigate the appropriateness of model functional form, both, linear and nonlinear functional forms were estimated within the proposed CR framework. The prediction accuracies of the resultant models were then compared to select the best functional form for the data used in this dissertation.

To summarize, this dissertation seeks to develop a generalized CR framework that determines simultaneously 1) the optimum number of pavement clusters, 2) cluster memberships of pavement samples, 3) clusters-specific significant explanatory variables, 4) estimated regression coefficients for PPMs, and 5) the best functional form between linear and nonlinear models. Minimization of Bayesian Information Criteria was used as the objective function; the optimal number of models with minimum BIC provided balance between goodness of fit and model complexity. The proposed mathematical program included constraints that ensured the minimum number of observations required in each cluster so as to achieve statistical significance. In addition, the mathematical program was formulated such that all observations of a pavement sample were assigned to the same cluster exclusively. Comprehensive algorithms – that integrated Simulated Annealing and Ordinary Least Square for the linear models, and All Subset Regression for the nonlinear models – were proposed to solve the proposed mathematical problems. To estimate the nonlinear model parameters, a logarithmic transformation was performed to linearize the adopted model; the estimated model parameters were then transformed back to the original scale by taking exponential.

Minimum effort was required to determine estimation parameters required to run the optimization experiments. Researchers, practitioners, departments of transportation, and other transportation agencies, including Federal Highway Administration can benefit from this study.

1.3 Objectives of the Dissertation

The primary objective of this dissertation was to propose a comprehensive framework to estimate accurate PPMs that minimize the overall estimation error. The models thus estimated must represent historical performance and can be used to predict conditions of pavement segments. The proposed methodology is motivated by the need to use the regression effects of explanatory variables on the pavement performance while simultaneously clustering the pavement segments and estimating PPMs so as to achieve balance between goodness of fit and model complexity. The specific problems addressed to achieve the above objective are:

(i) Develop a comprehensive CR framework that determines simultaneously 1) the optimal number of clusters, 2) the assignment of pavement samples into clusters, and 3) the cluster-specific significant explanatory variables and associated coefficients.

(ii) Use Bayesian Information Criteria as the objective function so as to penalize for the inclusion of additional parameters, the number of clusters and explanatory variables. The estimated models provide a balance between goodness of fit and model complexity.

(iii) Utilize a variable selection procedure to identify the best model specification. This procedure must distinguish between relevant and irrelevant variables, thus providing true underlying clusters and regression models. In addition, potential multicollinearity issues should be addressed.

(iv) Propose an algorithm to find a feasible range of the potential number of clusters that could be formed using the available data. The proposed algorithm must determine the upper bound (maximum) of the potential number of clusters to prevent a search beyond the feasible number of clusters for the given data.

(v) Propose a comprehensive solution algorithm to solve the proposed mathematical program. Determine the optimization parameters required for the given data set.

(vi) Investigate the appropriateness of linear and nonlinear pavement performance models within the proposed CR framework. It is intuitive that estimation errors would be high if an incorrect functional form is used to estimate model parameters. As pavement performance is a complex process and governed by many factors, several functional forms have been used in the literature. The best functional form is data-specific and depends on multiple aspects, such as number of observations and explanatory variables. Hence, various functional forms are required to be investigated to select the best to minimize the overall estimation errors.

1.4 Organization of the Dissertation

This dissertation is divided into five chapters. Chapter 2 proposes a CR framework to determine the optimal number of PPMs and a procedure to estimate a feasible range of potential number of clusters. In addition, the advantages of using BIC as the objective function are illustrated. A detailed solution algorithm to solve the proposed mathematical problem is presented along with the required estimation parameters. The models developed using the proposed framework are compared with the ones developed using the existing state-of-the-art method.

Chapter 3 extended the mathematical program proposed in Chapter 2 to propose a comprehensive CR framework to test the significance of explanatory variables. A variable selection procedure is proposed to select the best model specification that is free from the multicollinearity issues and provides balance between goodness of fit and model complexity.

Chapter 4 investigates the appropriateness of using a nonlinear functional form within the CR framework. A power functional form for PPM estimation is tested. The performances of the nonlinear and linear models are compared. Results show that the nonlinear models are superior for the data set used in this research.

Chapter 5 summarizes the overall insights from this research and discusses the significant contributions. Potential future works are also discussed in this chapter.

CHAPTER 2

SIMULTANEOUS GENERATION OF OPTIMUM PAVEMENT CLUSTERS AND ASSOCIATED PERFORMANCE MODELS

2.1 Introduction

Pavement deterioration is a complex process, involving both observed and unobserved factors (Hong and Prozzi 2010). Simple pavement performance models (PPMs) typically are developed using only a few critical factors, such as pavement type, age, and traffic volume. These models assume that all critical factors have significant effects on pavement performance (Hong and Prozzi 2015). This assumption is not necessarily correct in all cases. In addition, such factors as timing, scope of maintenance, rehabilitation, and reconstruction (MR&R) treatments, and construction methods and techniques significantly affect pavement performance; often, however, they are ignored (Prozzi and Madanat 2004). Consequently, errors are introduced to the estimated PPMs, leading to inaccurate forecasts (Hsiao 2003).

From a practical perspective, a single PPM could be ideal; however, it is associated with large prediction errors because it is very difficult to capture the heterogeneity in the entire roadway network with only one model (Shahin 1994). In contrast, prediction errors can be reduced significantly by developing an individual model for each pavement segment. In practice, this often is unfeasible because 1) there are not sufficient data for each pavement segment, 2) development costs are high, and 3) maintaining a system that includes thousands of models is impractical.

Typically, PPMs are developed using a Two-Step approach. First, pavement segments with similar characteristics are grouped together to form a cluster. The objective of clustering is to group the pavement segments that perform similarly over time The deterioration pattern can be determined by tracking a pavement performance measure, such as Present Serviceability Index (PSI). In practice, the performances of pavement segments within a cluster differ significantly because clusters are formed using only a few critical factors (Luo and Chou 2006, Luo and Yin 2008).

In the second step, the corresponding PPMs are developed using statistical techniques. A major challenge is to select characteristics that define clusters and the corresponding segments associated with them (Steinbach et al. 2003). If inappropriate characteristics are used, clusters may include homogeneous segments with different performance behaviors or heterogeneous segments with similar performance behaviors (Pulugurta 2007). The prediction accuracy of PPMs can be improved by subdividing the pavement segments into more uniform clusters. However, this subdivision is not always possible due to limited information (Luo and Chou 2006).

Figure 2.1a and 2.1b provide an example of heterogeneous performance behavior for two segments, each grouped within the same cluster (the Prioritization Category), using the Two-Step approach and actual data from Pavement Management System (PMS) of the Nevada Department of Transportation (NDOT). Segments ‘SR445 (SB), MP: 40-39’ and ‘SR445 (NB), MP: 36-37’ were assigned into one cluster, Prioritization Category 4. Segments ‘SR156 (WB), MP: 3-2’ and ‘SR892 (SB), MP: 25-24’ were assigned into Prioritization Category 5.

In contrast, Figure 2.1c illustrates that segments ‘SR445 (SB), MP: 40-39’ and ‘SR156 (WB), MP: 3-2’ had more homogeneous performance behavior. Similarly, segments ‘SR892 (SB), MP: 25-24’ and ‘SR445 (NB), MP: 36-37’ showed more consistent behavior as shown in Figure 2.1d. This suggests that factors other than the Prioritization Category are critical in causing the differences in performance behaviors. Influencing factors could include subgrade type, traffic loading characteristics, or any hidden factors.

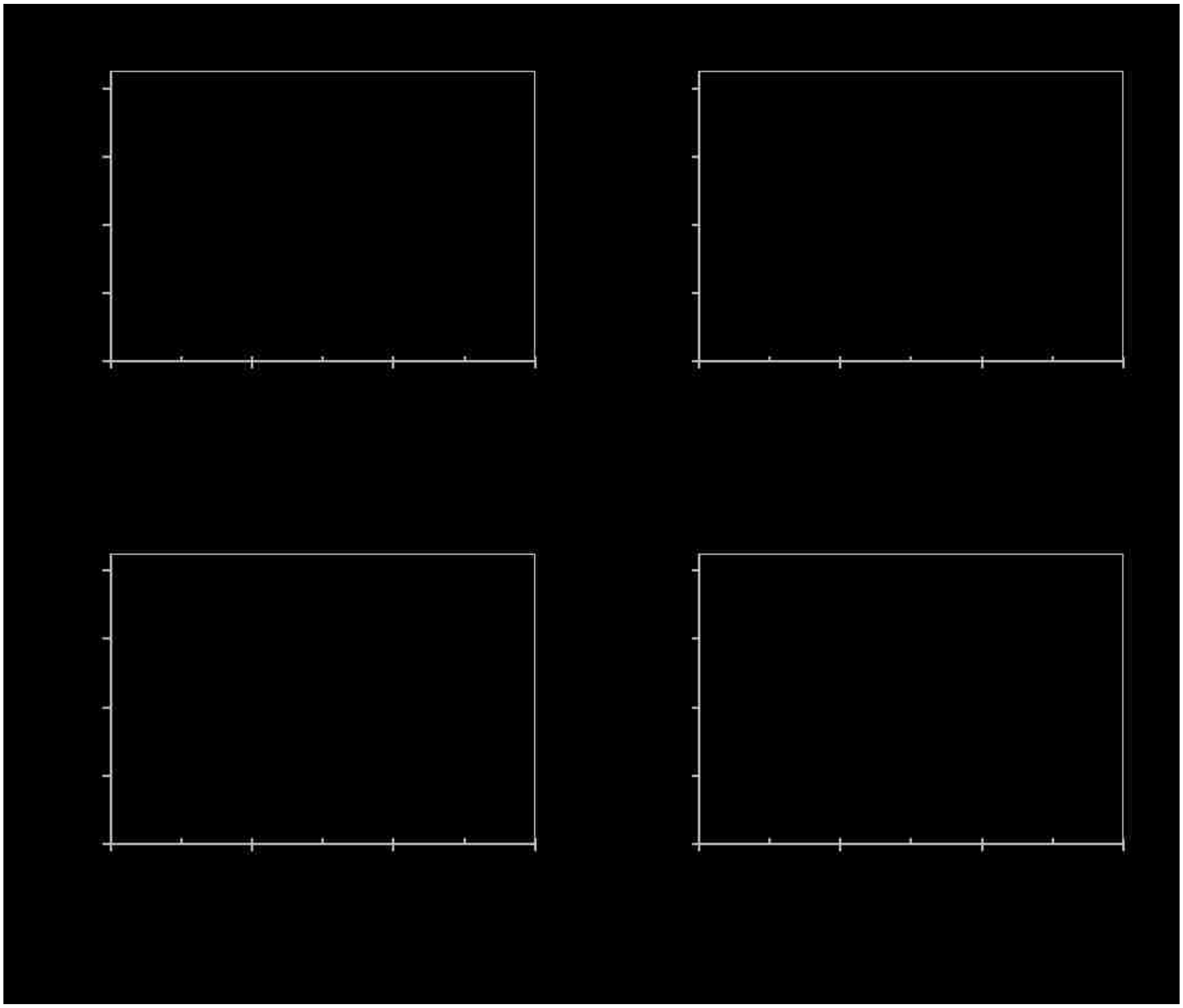

Figure 2.1 Heterogeneous performance behavior of pavement segments from the same Prioritization Category (a and b), and potential clusters with pavement segments with homogeneous performance behavior (c and d).

To address the limitations of the Two-Step approach, the existing literature proposed using Clusterwise Linear Regression (CLR) to determine clusters and associated regression models simultaneously. CLR generates clusters according to the effects that the explanatory variables have on the response variable of the regression models (Park et al. 2015). Pavement segments with similar regression effects are assigned into clusters such that the overall sum of squared errors within the clusters is minimal. CLR minimizes the overall prediction error by simultaneously determining the explanatory variables' coefficients, and by assigning each pavement segment into an appropriate cluster.

Spath (1979) introduced CLR with the exchange algorithm. Since then, the approach has been developed further, and implemented in many fields (DeSarbo et al. 1989, Wedel and SteenKamp 1989, Lau et al. 1999, Carbonnea et al. 2011, Schlittgen 2011, Zhen et al. 2012, Tan et al. 2013, Lu et al. 2014). In the context of pavement management, CLR first was used by Luo and Chou (2006) to model the deterioration of pavement conditions. First, the pavement network was clustered by using a few critical pavement characteristics; then, CLR was used to divide the data set further into several homogeneous clusters. The subdivision was performed at the data-point level; that is, data points collected over various years for a pavement segment could be assigned to multiple clusters; hence, there was a chance of pavement segments being associated with multiple performance models. An additional step was proposed to predict performance using the results from multiple models. Later, Luo and Yin (2008) expanded their research, using CLR to formulate the development of pavement distresses. Both studies (Luo and Chou 2006, Luo and Yin 2008) used pavement data collected by the Ohio Department of Transportation, and pavement age was the only explanatory variable.

Zhang and Durango-Cohen (2014) used CLR to account for heterogeneity in pavement deterioration. A potential overfitting issue was diagnosed using a methodology proposed by Brusco et al. (2008); this study included multiple explanatory variables to address some of the limitations present in the studies completed by Luo and Chou (2006) and Luo and Yin (2008). However, data used in the study were collected during the AASHO Road Test (Highway

Research Board 1962), conducted in the late 1950s in Ottawa, Illinois. The data set was collected for a single site in a relatively controlled environment; clearly, the site characteristics were not representative of all other locations. For example, the data set did not represent the varieties of soil and climatic conditions. Construction techniques and materials have changed substantially since this test was conducted. In addition, the optimum number of clusters was determined manually using the ‘elbow’ criterion. Experiments were run only for number of clusters equal to 1, 5, 10, 15, 20, and 25. One possible reason of not examining all potential number of clusters might have been large amount of computational time needed, given that the exchange algorithm was utilized with 100 instances with various initial assignments.

The existing state-of-the-art approach for CLR does not test the explanatory power of variables used in both clustering and regression analyses. That is, all the explanatory variables used in regression models are assumed to be significant. The effects of any insignificant explanatory variables on the dependent variable are accounted during assignment of pavement segments into clusters and estimation of regression coefficients. The presence of any insignificant explanatory variables might distort the underlying regression effects of other significant explanatory variables. This may lead to the incorrect assignment of pavement segments and estimation of PPMs.

Another key limitation of standard CLR is a need to pre-specify the number of clusters. In order to avoid pre-specifying the number of required clusters, this study proposes a mathematical program to simultaneously determine an optimal number of clusters, assignment of segments into clusters, and regression coefficients for all pre-specified explanatory variables. This mathematical program is flexible enough to handle multiple explanatory variables, multiple observations per pavement segments, and user-defined constraints on cluster characteristics.

In previous studies using CLR for pavement management (Luo and Chou 2006, Luo and Yin 2008, Zhang and Durango-Cohen 2014), the objective function was to minimize the sum of squared errors of prediction (SSE), which decreases monotonically as the number of clusters increases. That is, for a given data set, the optimum number of clusters is always the total number of data points (Kodinariya and Makwana 2013). The optimum number of clusters is equal to the total number of pavement segments, and each pavement segment is the sole member of its own cluster. Such clustering structure is unlikely to provide statistical reliable models. In addition, SSE always decreases when a new explanatory variable is added to the model (Wooldridge 2006). Usually, this leads to an overfitting problem (Wu 2014). Therefore, SSE is not the best objective function to use for searching the optimal number of clusters.

Even though SSE decreases as the number of clusters increases, the decrease rate also decreases significantly after a particular number of clusters. Going forward from this number, known as the *elbow point*, improvement in SSE is very small for each additional cluster. This elbow point indicates an optimal number of clusters for the given data set (Tibshirani et al. 2001). However, the SSE versus the number of clusters curve might not exhibit this elbow point distinctly in all the cases. Hence, it could be very challenging to choose the right number of clusters.

To address these limitations, this study proposes as an objective function the Bayesian Information Criteria (BIC) (Schwarz 1978), which penalizes the inclusion of additional parameters. An optimal number of clusters provides a balance between model complexity and goodness of fit. Given that BIC is an increasing function of SSE and free parameters (the number of clusters and regression coefficients) to be estimated, a clustering with the lowest BIC is preferred. Minimizing BIC reduces unexplained variation in the dependent variable (Schwarz

1978). The search process for the best model specification using BIC has the property of consistency, which asymptotically selects this model with probability of ‘1’ (Rao and Wu 1989, Yang 2005, Maydeu-Olivares and García-Forero 2010, Vrieze 2012).

In this study, the data limitations in the existing literature were addressed using actual field data collected across a variety of environmental, traffic, design, construction, and maintenance conditions. Pavement data were used that was collected for the past 12 years over the entire State of Nevada were used. This data included significant variations across a large range of characteristics, e.g., pavement segments exposed to either extreme desert heat or to very low winter temperatures in the mountains.

2.2 Methodology

2.2.1 Problem Formulation

This section includes the definition of a pavement sample, notation and definitions of terms used, proposed mathematical program, and a procedure to find upper bound of the range of the feasible number of clusters.

Definition of a Pavement Sample

The condition of a pavement segment improves when intervention occurs by applying an MR&R treatment. Such intervention alters the physical characteristics of the pavement. Hence, the performance of a pavement before and after the intervention differs, even though all other contributing factors remain constant. In this circumstance, the same pavement segment before and after intervention should be treated as two different samples. Given that the physical location of a pavement segment is the same, a different identifier is required to distinguish the set of consecutive observations before and after the intervention. In this study, the term *pavement sample* is used as an identifier to uniquely represent the set of consecutive observations that accounts for the historical interventions made on a pavement segment. Figure 2.2 provides a simplified depiction of how consecutive observations of a pavement segment are divided into two pavement samples.

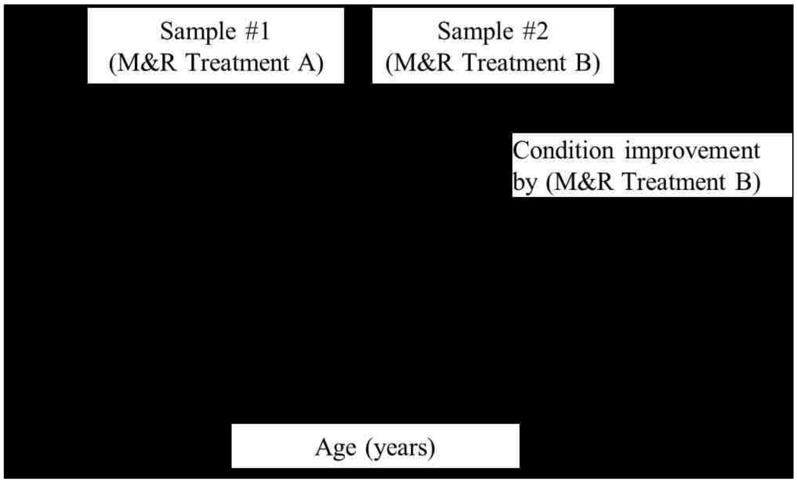

Figure 2.2 A typical pavement performance curve and a simplified depiction of how the observations of a pavement segments are divided into two samples.

In this study, a pavement sample – instead of pavement segment – was considered as a single entity during cluster analysis. Hence, if a pavement segment consists of two or more pavement samples, these samples could be assigned to different clusters.

Notation and Definition of Terms

The following notation and definitions were used to describe the proposed model:

I = Number of pavement samples in the network

i = Subscript for a pavement sample in the network, $i \in I$

T_i = Number of observation periods for a pavement sample i

t = Subscript for an observation period for a pavement sample, $t \in T_i$

O = Total number of observations = $\sum_i^I T_i \ \forall \ i \in I$

J = Number of explanatory variables

j = Subscript for an explanatory variable including an intercept, $j = 0,\dots,J$

x_{ijt}^{k} = Measurement of an explanatory variable j for a sample i at observation period t that is assigned to a cluster k $\forall$ $i \in I, j \in J, t \in T_i$

y_{it}^{k}= Measurement of dependent variable for a sample i at observation period t that is assigned to a cluster k $\forall$ $i \in I, t \in T_i$

K = Optimum number of clusters $(1 \leq k \leq K_{max})$

k = Subscript for a clusters, $k \in K$

K_{max} = Maximum number of potential clusters that could be formed

n = Minimum number of observations required in a cluster

C_K = Set of pavement samples that are assigned to cluster $k \; \forall \; k \in K$

p_{ik} = Cluster membership of a pavement sample i to a cluster k, $\forall$ $i \in I, k \in K$

β_{jk} = Estimated regression coefficient for an explanatory variable j including an intercept in cluster $k \; \forall \; j = 0,\dots,J$, $k \in K$;

Mathematical Program

In this study, PSI was chosen as the dependent variable, y. PSI serves as a unified standard, and is widely accepted for evaluating pavement performance, especially in terms of ride quality (Shoukry et al. 1997, Terzi 2006, Attoh-Okine and Adarkwa 2013). In addition, PSI reflects human rider's response, and is understood by highway users and legislators (Hudson et al. 2015). The adopted functional form for the regression model is expressed as:

$$y_{it}^{k} = \beta_{0k} + \sum_{j=1}^{J} \beta_{jk} * x_{ijt}^{k} \tag{2.1}$$

This study proposes a mixed-integer, nonlinear mathematical program to formulate a CLR. This mathematical program can determine an optimal number of clusters, assignment of segments into clusters, and regression coefficients for all pre-specified explanatory variables. The problem was defined by the optimum number of clusters, *K;* the number of predefined explanatory variables, *J;* the number of pavement samples to be clustered, *I*; and the number of observation periods, T_i associated with each pavement sample. The formulation partitioned pavement samples into an optimum number of clusters, with a PPM model fit to each cluster.

The objective function involved minimization of the BIC across *K* clusters. Decision variables to be determined were the optimum number of clusters, *K*; coefficients for all the pre-specified explanatory variables, β_{0k} and β_{jk}; and the cluster membership, p_{ik}.

Objective Function

$$Min.\ BIC = O+O*ln(2\pi)+O*ln\left(\frac{SSE}{O}\right)+(JK+2K\text{-}1)*ln(O) \qquad (2.2)$$

Subject to:

Sum of squared errors:

$$SSE= \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{t=1}^{T_i}\left[\beta_{0k}+\sum_{j=1}^{J}\beta_{jk}*x_{ijt}^{k}\text{ - }y_{it}^{k}\right]^2 *p_{ik}\ \forall\ i\in I,\ t\in T_i,\ j\in J,\ k\in K \qquad (2.3)$$

Membership constraints:

$$\sum_{k}p_{ik}=1\ \forall i\in I,\ k\in K \qquad (2.4)$$

$$p_{ik}=\begin{cases}1, & if\ \ sample\ i\ is\ assigned\ to\ cluster\ k;\\ 0, & Otherwise\end{cases}\quad \forall\ i\in I,\ k\in K \qquad (2.5)$$

Constraints for feasible partitions:

$$C_k=\left\{i\middle|p_{ik}=1\forall i\in I,\ k\in K\right\} \qquad (2.6)$$

$$C_{k'} \cap C_{k''} = \text{null} \;\; \forall k' \neq k'', k' \text{ and } k'' \in K \tag{2.7}$$

$$\cup_{k \in K} |C_k| = I \tag{2.8}$$

$$\sum_{i \in C_k} T_i \geq n \; \forall \, C_k \tag{2.9}$$

Constraints for range of clusters:

$$1 \leq k \leq K_{max} \tag{2.10}$$

$$K_{max} = F(I, T_i, n) \tag{2.11}$$

The objective (Equation 2.2) was to minimize BIC as a function of the number of clusters, the assignment of segments into clusters, and the coefficients to be determined. The constraint (2.3) calculated total SSE, which evaluates goodness of fit of the models. Each cluster was associated with a linear regression model with predefined explanatory variables. Deviations of predicted data from actual data were calculated separately for each cluster, and summed to obtain the total SSE.

The constraints (2.4 and 2.5) ensured that each pavement sample was assigned to exactly one cluster. The indicator p_{ik} equalled 1 if and only if a pavement sample i belonged to cluster k. Otherwise, it took a value of zero.

The constraints (2.6 - 2.9) determined feasibility of partition. These constraints ensured that pavement samples were partitioned exclusively into K clusters. The minimum-size constraint (2.9) was imposed to ensure sufficient observations in each cluster for statistically reliable estimation of coefficients. The total observations in a cluster were required to be no less than the minimum number of observations, n.

The constraints (2.10 and 2.11) were imposed to determine the maximum number of potential clusters. If a pavement sample had more than *n* observations, regression over these observations could generate statistically reliable estimates of coefficients. Hence, a cluster could be formed with only one sample that had more than *n*. If all pavement samples had more than *n* observations, each pavement sample could form a cluster. In this case, the maximum number of clusters would be the total number of pavement samples, *I*. However, in reality, it is possible that none of the pavement samples would have more than *n* observations. In this case, samples would be grouped to form a cluster at the sample level, but not at the observation level; that is, observations of a sample must not be assigned to more than one cluster. The constraint (2.11) determined the maximum number of potential clusters. This number if a function of *I*, T_i, and *n*, and represent by F. The procedure to calculate K_{max} is illustrated in the flow-chart as shown in Figure 2.3 and described in the following paragraphs:

Step 1. If the total number of observations, *O*, is less than the minimum number of observations required to form a cluster, *n*, then set K_{max} = zero and go to Step 6. Otherwise, create a matrix, **M** of size (τ_{max} x 2) with the following elements, where τ_{max} is maximum number of observations of a pavement segment(s) in the data set:

- The first column of **M** includes all integer values from 1 to τ_{max}.
- The second column includes the number of segments associated with the number of observations.
- If no segments have a particular number of observations in the data set, then set the second column of the matrix to zero.

Step 2. If any segment has a number of observations greater than or equal to *n* ($m_{\tau,1} \geq n$), then

a. Calculate $K_{max} = \sum_{\tau \geq n} m_{\tau,2}$.

b. Update $m_{\tau,2}$ with 0 for $\tau \geq n$.

Otherwise, go to Step 3 to find the maximum number of clusters that could be formed.

Step 3. If the matrix **M** has all zeros in its second column $\left(\sum_{\tau} m_{\tau,2} = 0\right)$, then return $K_{max} = K_{max}$ and go to Step 6. Otherwise:

a. Update **M** by removing all rows that have number of segments equal to zero ($m_{\tau,2} = 0$).

b. Initialize two indices as: $\omega = \vartheta =$ number of rows in **M**.

c. Make a copy of **M** and let it represent by **M'**.

d. If the remaining total number of observations ($\sum_{\tau=1}^{\omega} m_{\tau,1} * m_{\tau,2}$) is less than n, then, $K_{max} = K_{max}$ and go to Step 6. Otherwise, initialize S with the value of $m_{\omega,1}$ and $m_{\omega,2} = m_{\omega,2} - 1$.

Step 4. Repeat the following steps until $S = n$.

Step 4.1. If ($m_{\vartheta,2} = 0$), then $\vartheta = \vartheta - 1$. Otherwise, go to Step 4.3.

Step 4.2. If ($\vartheta = 0$), then set:

- **M** = **M'**,
- $\vartheta =$ number of rows of **M**,
- $n = n + 1$, $S = 0$, and
- Go to Step 4.6.

Otherwise, go to Step 4.3.

Step 4.3. If ($S > n$), then set:

- $S = S - m_{\vartheta,1}$,

- $m_{\vartheta,2} = m_{\vartheta,2} + 1$, and
- $\vartheta = \vartheta - 1$.

  Otherwise, go to Step 4.6.

Step 4.4. If ($\vartheta = 0$), then set:

- $\mathbf{M} = \mathbf{M}'$,
- ϑ = number of rows of **M**,
- $\omega = \omega - 1$, and
- $S = 0$.

  Otherwise, go to Step 4.6.

Step 4.5. If ($\omega = 0$), then

- Update both indices ω and ϑ with the number of rows of **M**
- Set $n = n + 1$,
- Go to Step 4.6.

  Otherwise:

- Set $S = m_{\omega,1}$, and $m_{\omega,2} = m_{\omega,2} - 1$
- Go to Step 4.6.

Step 4.6 Update S with $\left(S + m_{\vartheta,1}\right)$ and $m_{\vartheta,2}$ with $\left(m_{\vartheta,2} - 1\right)$.

Step 5. Set $K_{max} = K_{max} + 1$, and go to Step 3.

Step 6. Return the current value of K_{max} and stop.

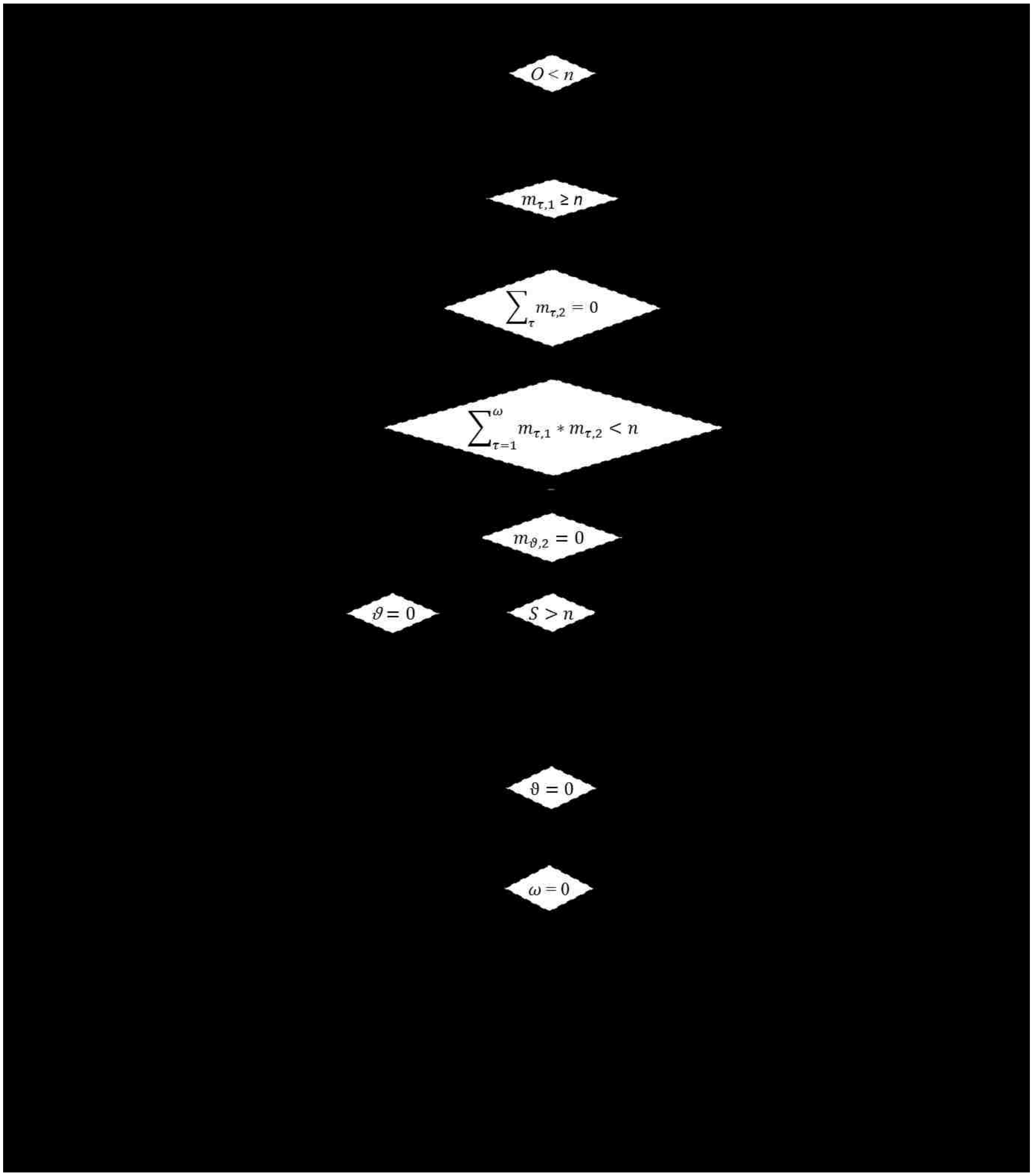

Figure 2.3 Algorithm utilized to calculate the maximum number of potential clusters.

2.2.2 Solution to the Mathematical Program

Simulated annealing (SA) coupled with an ordinary least square (OLS) algorithm was implemented to solve the above mathematical program. SA was used to cluster the data set; that is, estimate, p_{ik}. For each accepted neighborhood clusters, OLS was utilized to estimate the regression coefficients, β_{0k} and β_{jk}. The fitting linear models (lm) function, available in the statistical software, R, was used to estimate these coefficients (R Core Team 2015). DeSarbo et al. (1989) successfully implemented such an algorithm to solve the CLR problem. The algorithm utilized to solve the clusterwise multiple linear regression is described as follows and illustrated in Figure 2.4.

Step 1. Initialization:

Step 1.1. Set $K = 2$, and BIC_{min}=infinity.

Step 1.2. Set values of initial temperature (θ_0), final minimum temperature (θ_{min}), cooling rate (λ), and the maximum number of neighbors to be generated (N_{max}) at each temperature level. Set the iterator $N = 1$.

Step 2. Maximum number of potential clusters:

Calculate K_{max} utilizing function F as described above, as part of the Constraint (2.11).

Step 3. Initial estimation of regression coefficients:

Step 3.1. For a given number of clusters, K, randomly assign cluster memberships to all pavement samples.

Step 3.2. Count the number of observations of all pavement samples assigned to each cluster. If all clusters have at least n observations, then go to Step 4; otherwise, reassign the cluster memberships until all clusters have at least n observations. Let C_K^N be the valid initial clusters.

Step 3.3. Estimate β_{0k} and β_{jk} for all K clusters using OLS.

Step 4. Evaluate the objective function, BIC_K^N using Equation 2.2.

Step 5. Generate a set of neighborhood clusters near to the previous one, using the following steps:

Step 5.1. Randomly select a pre-specified number of pavement samples (N_{ps}) to change their memberships.

Step 5.2. For each of the sample selected, assign a new membership by generating a random number $u \sim U(1, K)$. If the new membership is same as the previous one, regenerate a random number $u' \sim U(1, K)$ until it is different. Repeat this process until the memberships of all the selected pavement samples are different from those that were previously assigned.

Step 5.3. Count the total number of observations of all pavement samples assigned to each cluster.

Step 5.4. If all clusters have at least *n* observations, go to Step 6; otherwise, repeat Steps 5.1., 5.2., and 5.3. until all clusters have at least *n* observations. Let C_K^{N+1} be the new set of valid neighborhood clusters.

Step 6. Search of a solution:

Step 6.1. For C_{K+1}^{N}, estimate new β_{0k} and β_{jk} for all *K* clusters using OLS.

Step 6.2. Evaluate BIC_K^{N+1} using Equation 2.1.

Step 6.3. Calculate $\Delta BIC = BIC_K^{N+1} - BIC_K^{N}$.

Step 6.4. Check the following two conditions:

a. If $\Delta BIC < 0$, accept the current set of clusters, C_{N+1}^{K}, and corresponding β_{0k} and β_{jk}. Go to Step 7; otherwise, go to Step b.

b. Generate a random number $u"{\sim}U(0,1)$. Calculate the acceptance probability, $p_{accept} = exp\left(\frac{-\Delta BIC}{B*T}\right)$; where B is a Boltzmann's constant. If $p_{accept} > u"$, accept the current set of clusters, C_{K+1}^{N}, and corresponding δ_k and β_{jk}. Go to Step 7; otherwise, return to Step 5.

Step 7. Counter and temperature update:

Step 7.1. Repeat Steps 5 and 6 for N_{max} times.

Step 7.2. If $\theta < \theta_{min}$, stop the algorithm. Otherwise, reduce temperature by multiplying the current temperature by the pre-specified cooling rate, λ, set N =1, and go to Step 5.

Step 8. Stopping criteria:

Step 8.1. Update BIC_{min} with the smallest between the one obtained in Step 7 and the current BIC_{min}. Set $K_{optimal}$ equal to K.

Step 8.2. Repeat Steps 3 to 7 for $K_{max} - 1$ times.

This algorithm seeks solutions using a probabilistic approach. The algorithm starts with a high temperature, θ, and a high probability of accepting a worse solution, p_{accept}. This enables occasional 'uphill' moves, which help escape from the local minima. The algorithm builds up a rough view of the search space by moving with large step lengths. As θ drops, p_{accept} decreases to behave more closely as a greedy algorithm, with small step lengths slowly focusing on the most promising solution space. Theoretical studies have shown that with infinitely slow cooling, the algorithm converges to a global minimum (Román-Román et al. 2012).

Figure 2.4 Algorithm utilized to solve the clusterwise multiple linear regression.

2.3 Numerical Experiment and Results

2.3.1 Experimental Research Data

Data used in this study were extracted from the PMS database of NDOT. The data consisted of various classes – location data, segment data, contract data, environmental data, traffic data, and condition data – collected throughout the entire State of Nevada. Various environmental factors were tested as explanatory variables including elevation, annual precipitation, average minimum and maximum temperatures, the number of wet days, and the freeze and thaw cycles. Provided that the minimum data requirements were met, other variables (e.g., economic and social factors) could be included. Potential explanatory variables used in this study could be divided as follows:

1. Continuous explanatory variables:
    - *age* - pavement age since the last M&R treatment;
    - *adt* - average daily traffic in one direction;
    - *trucks* - average daily trucks in one direction;
    - *elevation* - midpoint elevation of a segment;
    - *precip* - average annual precipitation (cm/yr);
    - *min_temp* - minimum average annual temperature (^{0}C);
    - *max_temp* - maximum average annual temperature (^{0}C);
    - *wet_days* - total number of wet days in a year;
    - *freeze_thaw* - total number of freeze-thaw cycles that a pavement experienced in a year;
    - *rut_depth* - average ride rut depth (cm);
2. Categorical explanatory variables:
    - Two dummy variables, *lane=2* and *lane≥3*, were encoded to represent if a segment

had two lanes and three or more lanes, respectively.

- NDOT classifies pavement segments under (i) the Interstate Route (IR), (ii) the National Highway System (NHS), or (iii) the Surface Transportation Program (STP). Two dummy variables, *sys_id=2* and *sys_id=3*, were encoded to represent if a segment belonged to the NHS or STP classification system, respectively.
- NDOT grouped its roadway network into five prioritization categories, Category 1 through Category 5, using such factors as highway classification and traffic volumes (NDOT, 2011). The type and frequency of M&R activities vary among these prioritization categories. Four dummy variables – *category=2, category=3, category=4*, and *category=5* – were encoded to represent if a segment was grouped in Prioritization Categories 2, 3, 4, or 5, respectively.
- Six dummy variables – *f_class=2, f_class=3, f_class=4, f_class= 5, f_class= 6* and *f_class= 7* – were encoded to represent if a segment is classified as Functional Class 2, 3, 4, 5, 6, or 7, respectively.

2.3.2 Data Preparation

In practice, missing and inconsistent data are commonly encountered in pavement condition data sets (Buchheit et al. 2005, Farhan and Fwa 2015). When using inconsistent data, development of accurate pavement performance models is difficult (Pierce et al. 2013, Tan and Cheng 2014). A detailed data analysis was performed to check for inconsistent and missing information in the data set, and some were found. Some of the missing data were synthesized based on associated information available in the data set. In preparation of the PMS data, the following filters were applied:

- Only one-mile segments were selected for consistency.

- Only pavement segments with the most recent maintenance contracts awarded in 2001 or later were used in the study.
- PSI of a pavement should deteriorate over time if no MR&R treatment occurred. If PSI of a segment in any year increased by 0.1 or more points from the previous year without any MR&R treatment, all observations for that year were excluded from the analysis. However, if an increase in PSI in any year was less than 0.1 from the previous year, it was assumed to be a random error during the process of pavement evaluation or data processing. Therefore, those observations were included in the analysis.
- If PSI of any year decreased by one or more points from the previous year, all observations for that year were excluded from the analysis.
- In practice, the PSI range is between 4.5 and 1.5. Therefore, if a pavement segment had a PSI beyond these limits in any year, it was considered an outlier, and all observations for that year were excluded.
- Only PSI values used were within the interval of mean minus three standard deviations to mean plus three standard deviations.
- Pavement samples that did not consist of data regarding condition for at least two consecutive years were excluded.
- Data analysis showed that the improvement of PSI was seen one or two years after the contract award date. Hence, the age of the pavement sample was set to 0 when the actual improvement occurred rather than when the contract was awarded.

After data preparation was completed, 4,138 flexible pavement samples with 17,642 observations were available. For CLR modelling, 14,637 observations were collected from 2001 to 2010; the remaining 3,005 observations, collected in 2011 and 2012 – about 17% of the total number – were used as test data set to check the accuracy of the CLR models. Descriptive statistics of the continuous explanatory variables and the dependent variable are illustrated in Table 2.1. The descriptive statistics of the categorical variables are illustrated in Table 2.2.

Table 2.1 Descriptive Statistics for the Continuous Variables

Variable	Description	Min.	Max.	Mean	Std. Dev.
psi	Present serviceability index	1.60	4.57	4.01	0.41
age	Age of the last pavement maintenance treatment	0.00	8.00	2.24	2.01
adt	Average daily traffic (single bound)	20.00	132000.00	4844.45	9812.57
trucks	Average daily trucks (single bound)	1.00	7731.00	862.29	1082.20
elevation	Midpoint elevation (m)	228.60	2667.00	1368.25	415.19
precip	Average annual precipitation (cm/year)	3.94	89.28	19.33	10.10
min_temp	Annual average minimum temperature (^{0}C)	-6.67	13.33	3.20	4.00
max_temp	Annual average maximum temperature (^{0}C)	7.78	31.67	20.31	4.22
wet_days	Number of wet days in a year	11.00	81.00	42.14	15.67
freeze_thaw	Number of freeze-thaw cycles in a year	0.00	230.00	136.75	51.51
rut_depth	Average ride rut depth (cm)	0.00	1.60	0.14	0.14

Table 2.2 Descriptive Statistics for the Categorical Variables

Variable	Category	Dummy Variable	Number of Obs.	Percent
System ID	IR	-	5,165	29.3
	NHS	*sys_id=2*	6,281	35.6
	STP	*sys_id=3*	6,196	35.1
Number of Lanes	1	-	10,438	59.2
	2	*lane=2*	6,494	36.8
	≥ 3	*lane≥3*	710	4.0
Prioritization Category	1	-	5,643	32.0
	2	*category=2*	4,017	22.8
	3	*category=3*	3,778	21.4
	4	*category=4*	1,872	10.6
	5	*category=5*	2,332	13.2
Functional Class	1	-	5,265	29.8
	2	*f_class=2*	134	0.8
	3	*f_class=3*	6,326	35.9
	4	*f_class=4*	3,294	18.7
	5	*f_class=5*	2,216	12.6
	6	*f_class=6*	354	2.0
	7	*f_class=7*	53	0.3

Note: *IR - Interstate Route, NHS - National Highway System, STP - Surface Transportation Program*

2.3.3 Estimation Parameters

Performance of the SA algorithm generally depends on the values of optimization parameters utilized for a given problem. To ensure proper initialization and search for optimal solutions, selection of the most appropriate parameter values is critical (Park and Kim 1998, Roshan et al. 2013). A body of literature exists regarding various methodologies for finding the most appropriate values for annealing parameters in SA (Park and Kim 1998, Kirkpatrick et al. 1983, Collins et al. 1988, Rose et al. 1990, Selim, and Alsultan 1991, Guo and Zheng 2005).

If an SA algorithm is allowed to run for a sufficiently long time by setting a high initial temperature with a slow cooling rate, the algorithm performs well, as shown in the study performed by Anily and Federgruen (1987). In such a cooling scheme, the selection of the most appropriate parameter values may not be critical. However, computation time cannot always be ignored. Hence, the algorithm has to find a good solution in a reasonable time (Kirkpatrick et al. 1983).

Effective values to be assigned to the optimization parameters depend of the type and complexity of the problem. These values may not be obvious to determine, but rather might be determined by ‘trial and error’ methods for a given problem (Collins et al. 1988). In this study, values assigned to the optimization parameters were determined using experience gained from previous research (Paz et al. 2015a and b, Khadka and Paz 2017, Paz and Khadka 2017) that involved SA and other comparable algorithms. Table 2.3 lists the parameter values used in this study.

Table 2.3 Setup Parameters for Implementation of the Proposed Algorithm

Parameter	Value	Remarks
θ_0	10	Initial temperature
θ_{min}	10e-17	Minimum temperature
B	3000	Boltzmann constant
λ	0.97	Cooling rate
N_{max}	10	Number of neighborhood solutions generated at each temperature level
n	800	Minimum number of observations required in a cluster
N_{ps}	25	Number of pavement samples, which memberships were changed to generate a neighborhood cluster

2.3.4 Results and Discussion

Given the constraints for feasible partitions defined in the problem formulation and the minimum number of observations required in a cluster, n = 800, the proposed algorithm determined 16 as the maximum number of potential clusters. The algorithm searched for the optimum number of clusters from 2 to 16. Seven-cluster CLR models provided the optimum solution with the lowest BIC. The estimated regression coefficients for the CLR models are presented in Table 2.4.

Figure 2.5a shows the smallest BIC for each of the clusters (K = 2 to16) considered in this experiment. Figure 2.5b shows the trajectory of the objective function, BIC, when the CLR models were used. The initial value of BIC was 8,502. After 1,360 iterations, the BIC decreased to 3,008. This change was equivalent to an improvement of 65%.

It was observed that not all coefficients had associated p-value less than 0.05. In this study, the significance level was considered to be 5%. As expected, coefficients differed in magnitude and sign across the clusters, which indicated that the deterioration patterns of pavement samples varied among the clusters. However, seven explanatory variables had the same sign across all clusters.

Table 2.4 Coefficients Obtained Using the Proposed CLR Approach

Parameters	C_1 (2,279)	C_2 (1,959)	C_3 (2,169)	C_4 (2,094)	C_5 (1,883)	C_6 (1,936)	C_7 (2,317)
intercept	4.1450	6.2420	2.7910	6.7280	7.7810	12.1400	3.8730
age	-0.0347	-0.0400	-0.0350	-0.0327	-0.0464	-0.0392	-0.0498
adt†	-0.0059	-0.0035	-0.0262	-0.0028	-0.0078	-0.0053	-0.0334
trucks†	0.0002‡	0.0205‡	0.0190‡	-0.0151‡	0.0306	0.0557	0.0752
elevation†	0.0066‡	0.0182‡	-0.0352‡	-0.1079	-0.1418	-0.0131‡	0.1060
precip	-0.0037‡	-0.0118‡	-0.0037‡	-0.0248	0.0094‡	-0.0518	-0.0252
min_temp	-0.0301	0.0129‡	-0.0092‡	0.0497	-0.0321	-0.0447	0.0532
max_temp	0.0252	-0.0297	0.0249	-0.0568	-0.0124‡	-0.0554	-0.0311
wet_days	0.0049	-0.0104	0.0115	0.0031‡	-0.0028‡	0.0004‡	-0.0061
freeze_thaw†	-1.6970	1.5130	-0.2029‡	1.8550	-1.4100	-13.7900	4.2370
rut_depth	-0.6316	-1.0020	-1.1060	-0.5614	-0.8408	-0.3999	-0.9900
lane=2	-0.3710	-0.1663	0.0121‡	-0.1216	-0.5574	-0.2745	0.0258‡
lane≥3	-0.3248	-0.1950	-0.1713	-0.3215	-0.3974	-0.2511	0.0391‡
sys_id=2	-0.4417	-0.4070	0.7454	-0.2811	-0.2639	-0.4300	1.4320
sys_id=3	-0.8024	-0.4868	0.6121	-0.3452	-0.1346‡	-0.3379	1.1350
f_class=2	0.5210	0.0940‡	-0.9383	0.4118	0.7050	0.9474	-1.5090
f_class=3	0.3819	0.3377‡	-0.8551	0.3270	0.3107	0.4025	-1.3770
f_class=4	0.6185	0.2964‡	-0.6925	0.2945	0.0189‡	0.5468	-1.0330
f_class=5	0.6183	-0.0703‡	-0.6471	-0.7763	-0.6250	0.3015	-1.5590
f_class=6	0.4724	-0.4187	-1.2170	-0.6566	-0.7955	0.0582‡	-1.7880
f_class=7	0.5543	0.4159	-1.3840	-0.1375‡	NA	0.7614	-1.3640
category=2	-0.3059	-0.0444‡	0.0762‡	-0.1300	-0.6077	-0.1175‡	0.0246‡
category=3	-0.3191	-0.0156‡	-0.0421‡	-0.2457	-0.6000	-0.2317	-0.0080‡
category=4	-0.4683	-0.2976	-0.3665	-0.1535	-0.6331	-0.5050	-0.3982
category=5	-0.4634	-0.3555	-0.7165	-0.1899	-0.8446	-0.6001	-0.4145
BIC	136	338	238	216	496	857	271

Note: The quantity included in parentheses represents the total number of observations in a cluster.
† *variable value in thousands,* ‡ *coefficient with p-value > 0.05,* and *NA = Not applicable*

Different clusters had different number of significant explanatory variables. For example, Cluster #2 had #10 variables with insignificant regression coefficients. In addition, among all seven clusters, five variables – *age*, *adt*, *rut_depth*, *category=4*, and *category=5* – were significant. However, four variables – *trucks*, *elevation*, *precip*, and *category=2* – were insignificant in four different clusters.

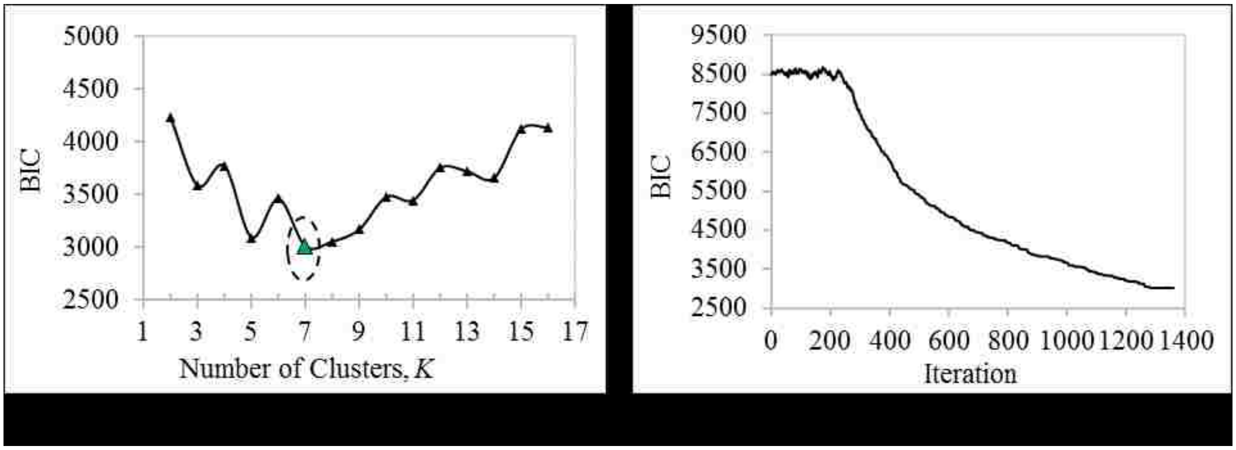

Figure 2.5 BIC trend over the number of clusters (a), and trajectory of the BIC during optimization for 11-cluster models (b).

The performance of the proposed CLR approach was compared with that of the existing CLR for pavement management. Experiments using the existing CLR approach were run for all feasible clusters (*K* = 2 to 16). Figure 2.6a shows the smallest SEE for each of these clusters. As expected, SSE decreased with an increasing number of clusters, but at a very small rate after *K*=11. In this case, Figure 2.6a does not exhibit a clear elbow point. Hence, an optimum number of clusters needs to be decided by visual inspection while considering the trade-off between goodness of fit and model complexity (i.e., of the number of models and explanatory variables). This inherent subjectivity when choosing an optimum number of clusters is a major drawback for the existing state-of-the-art CLR approach.

After careful assessment, 11-cluster CLR models were selected as the optimum solution. Figure 2.6b shows the trajectory of SSE, when the 11-cluster CLR models were used, and Table 2.5 provides the corresponding regression coefficients. Similar to the results obtained from the proposed CLR approach, the coefficients differed in terms of magnitude and sign. In addition, some coefficients had p-values larger than 0.05.

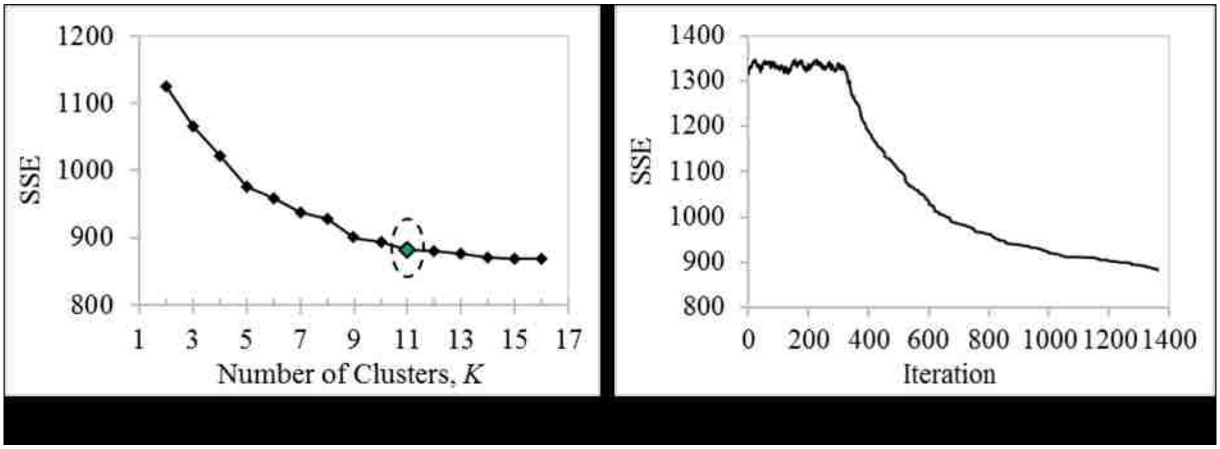

Figure 2.6 SSE trend over the number of clusters (a), and trajectory of the SSE during optimization for 11-cluster models (b).

The *BIC* for these models are provided in Tables 2.4 and 2.5. To compare the goodness of fit, overall *BIC* values were calculated. The overall *BIC* for the 7-cluster models obtained from the proposed approach was 3,008, whereas the *BIC* for the 11-cluster models, obtained from the existing state-of-the-art approach, was 3,171. This difference was the result of similar or better explanatory powers provided by the proposed approach with seven versus 11. That is, the more clusters, the more coefficients for explanatory variables needed to be estimated for a similar goodness of fit; thus, the *BIC* is increased.

Table 2.5 Coefficients Obtained Using the Existing State-of-the-Art CLR Approach

Parameters	C_1 (1,229)	C_2 (1,365)	C_3 (1,413)	C_4 (1,091)	C_5 (1,423)	C_6 (1,412)	C_7 (1,430)	C_8 (1,321)	C_9 (1,272)	C_{10} (1,360)	C_{11} (1,321)
intercept	3.9179	4.1242	4.9547	14.4267	7.4120	3.7039	4.6147	3.7970	7.9487	6.4985	7.5429
age	-0.0430	-0.0391	-0.0455	-0.0419	-0.0504	-0.0365	-0.0358	-0.0459	-0.0502	-0.0369	-0.0320
adt†	-0.0373	-0.0167	-0.0159	-0.0037	-0.0403	-0.0031	-0.0011‡	-0.0370	-0.0059	-0.0030‡	-0.0033
trucks†	0.1137	0.0342	0.0382	0.1256	0.0478	0.0011‡	-0.0266‡	-0.0063‡	0.0515	0.0107‡	0.0310
elevation†	-0.0244‡	0.1784	-0.0490‡	-0.0531‡	-0.2150	0.0134‡	-0.1074	-0.1681	0.1774	-0.0451‡	0.0017‡
precip	-0.0039‡	-0.0173‡	-0.0075‡	-0.1292	-0.0096‡	-0.0037‡	0.0186‡	0.0371	-0.0507	-0.0268	-0.0479
min_temp	0.0105‡	-0.0182‡	0.0851	-0.0933	0.0049‡	-0.0217	-0.0259	-0.0187‡	0.1173	-0.0115‡	0.0465
max_temp	-0.0031‡	0.0135‡	-0.0581	-0.0546	-0.0348	0.0209	0.0182‡	0.0266	-0.1153	-0.0141‡	-0.0647
wet_days	0.0044‡	-0.0080	-0.0031‡	0.0234	0.0023‡	0.0051	-0.0001‡	0.0021‡	-0.0182	0.0044‡	0.0076
freeze_thaw†	0.9951‡	-0.8567‡	6.5932	-20.8922	0.2354‡	-1.3799	0.7758‡	1.4509	5.2168	-2.1911	-1.9284
rut_depth	-0.6360	-1.0021	-1.2161	-1.0259	-0.2717	-0.4595	-0.7600	-1.0184	-1.8027	-0.8827	-1.1997
lane=2	0.0229‡	-0.1578	-0.3482	-0.1019	0.0305‡	-0.0560‡	-0.2602	-0.2159	-0.5342	-0.2792	-0.2630
lane≥3	0.2207	-0.0342‡	-0.4349	-0.3108	0.3095	-0.4315	-0.4773	0.0088‡	-0.3303	-0.2730	-0.3863
sys_id=2	1.3916	0.2777‡	0.5097	-0.1217‡	1.1989	-0.2483	-0.4121	0.8043	-0.6980	-0.5032	0.4380
sys_id=3	1.3306	0.1837‡	0.4183	-0.2667‡	0.6715	-0.3782	-0.4986	0.8235	-0.6427	-0.6491	0.6490
f_class=2	-1.4239	-0.3873‡	-0.2363‡	0.4439	-0.8292	0.3205	0.5201	-0.9277	1.4305	-0.2489‡	-0.7147
f_class=3	-1.3580	-0.5468	-0.3844	0.1211‡	-0.8767	0.1509‡	0.3306	-0.9877	0.9337	0.4276	-0.5483
f_class=4	-1.2475	-0.4867	-0.3948	0.2204‡	-0.5639	0.2707	0.4126	-1.1156	0.8923	0.6946	-0.5923
f_class=5	-1.8532	-0.5072	-0.9079	0.0311‡	-1.4272	0.2050‡	0.1754‡	-2.1799	0.6035	0.4307	-0.4690
f_class=6	-2.0674	-0.7052	-0.7661	-0.7152	-1.7136	0.1425‡	-0.6626	-2.2356	0.2427‡	0.5646	-0.7057
f_class=7	NA	-0.6277	-1.5309	0.4157‡	-0.6797	0.0900‡	0.1752‡	-2.2750	NA	0.7240	NA
category=2	0.0160‡	0.0723‡	-0.3571	0.2248	-0.2728	0.0592‡	-0.1951	-0.0647‡	-0.6593	-0.1255	-0.1354‡
category=3	0.0208‡	0.0147‡	-0.4289	0.1372‡	-0.1665	-0.0549‡	-0.2926	-0.0776‡	-0.7634	-0.1954	-0.4642
category=4	-0.2385	-0.4135	-0.6921	0.0199‡	-0.2206	-0.1261	-0.4187	-0.1467	-1.0083	-0.7392	-0.6837
category=5	-0.3813	-0.3251	-0.8041	-0.0703‡	-0.1786	-0.0797‡	-0.3830	-0.1789	-1.1519	-1.2923	-1.1395
BIC	125	97	420	553	314	226	88	103	337	233	276

† *variable value in thousands.*

‡ *coefficient with p-value > 0.05.*

Note: The quantity included in the parenthesis represents the total number of observations in a cluster.

NA = Not applicable

Investigation of Potential Overfitting

Brusco et al. (2008) noted that clusterwise regression models have great potential for overfitting. Often, a variation in the response variable is governed by clustering. Hence, they recommend investigating the potential presence of overfitting in CLR models. This study adopted a procedure proposed by Brusco et al. (2008) to diagnose overfitting. For the optimum 7-cluster models, the total sum of squares (TSS) was 2,419, the between-clusters sum of squares (BCSS) was 30, the within-clusters sum of squares (WCSS) was 2,389, the sum of squares due to regression (SSR) was 1,456, and the SSE was 933. The BCSS was around 1% of TSS, and SSR was 62% of WCSS. These results indicated that there was no overfitting, as most of the variation in PSI was explained by the within-cluster regressions. SSE accounted for 38% of TSS which suggests that the models still have a relatively high rate of error. A nonlinear functional form should be investigated to reduce the existing errors.

Model Accuracy

The accuracy of the models obtained from both approaches was assessed by calculating the overall root-mean-square error (RMSE) , as follows:

$$RMSE = \sqrt{\frac{\sum_{1}^{\eta}\left(y_{it}^{k} - \hat{y}_{it}^{k}\right)^{2}}{\eta}} \qquad (2.12)$$

where, y_{it}^{k} = the observed PSI, $\hat{y}_{it}^{k}$ = the predicted PSI, and η = the number of predictions.

Both models were applied to the test data set. Memberships of the pavement samples were assigned by mapping the sample IDs and memberships determined by the CLR models. Associated regression models and observed data were used to estimate the PSIs. Predicted PSIs then were compared with the observed PSIs, as shown in Figure 2.7. Results indicated that the both CLR models overestimated the PSI. A possible reason might be the existence of multicollinearity among the explanatory variables.

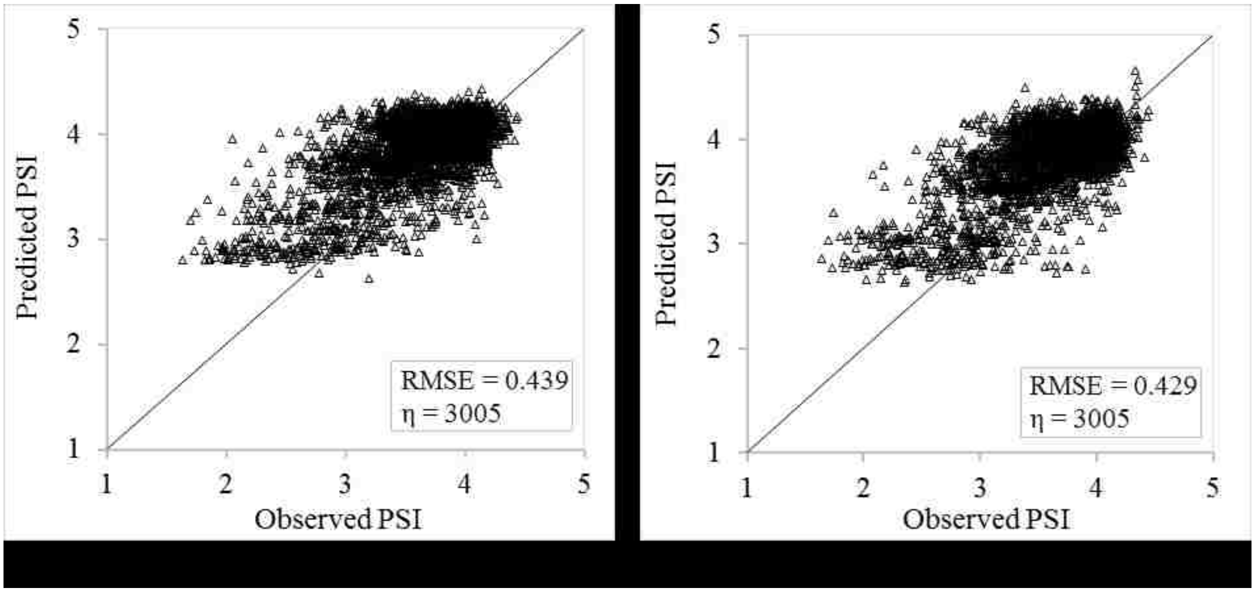

Figure 2.7 Observed versus predicted PSIs: (a) the proposed CLR approach, and (b) the state-of-the-art approach.

The RMSE for 2011 and 2012 predictions were calculated for models obtained from both approaches (Figure 2.7). The RMSE values for models obtained using the proposed CLR and the existing state-of-the-art approach were 0.439 and 0.429, respectively. Even though the prediction accuracy of both models was similar, 7-cluster models obtained using the proposed approach are preferred because they were more parsimonious than 11-cluster models. That is, 7-cluster models were preferred over 11-cluster models that provided the same explanatory power.

2.4 Conclusions

This study proposed and implemented a clusterwise multiple linear regression to develop pavement performance models. A mixed-integer nonlinear mathematical program was formulated to explain the problem. The CLR approach simultaneously divided pavement samples into an optimum number of clusters, and estimated a PPM for each cluster.

In the experiments, various environmental factors were considered as potential explanatory variables including, elevation, annual precipitation, average minimum and maximum temperatures, the number of wet days, the freeze and thaw cycles. The proposed approach enabled consideration of other types of variables, such as economic and social factors. Formulation of mathematical program developed in this study supports a number of explanatory variables, multiple observations per pavement segment, and user-defined constraints on cluster characteristics.

Simulated annealing coupled with OLS was used to solve the mathematical program. For the data used in the experiments, the algorithm found that 7-cluster models provided the optimum solution. Results obtained from the proposed CLR models were compared with results obtained from the state-of-the-art approach. This comparison showed that the proposed CLR approach performed better than the state-of-the-art approach in predicting the PSI of pavement samples.

The analysis showed that overfitting was not an issue for the resulting clusters and regression models. As expected, the use of the BIC as an objective function to determine the best model specification provided a more parsimonious structure compared with that obtained using SSE. This was a consequence of the consistency property of the BIC (Schwarz 1978, Rao and Wu 1989, Yang 2005, Maydeu-Olivares and García-Forero 2010, Vrieze 2012).

CHAPT`ER 3

COMPREHENSIVE CLUSTERWISE LINEAR REGRESSION FOR PAVEMENT MANAGEMENT SYSTEMS

3.1 Introduction

Pavement deteriorates over time due to the combined effects of traffic and environmental factors. To keep pavement in a serviceable condition, highway agencies primarily have two alternatives: 1) permit the pavement to deteriorate until its condition falls below the serviceability limit, and then perform rehabilitation or reconstruction work; or 2) intervene with the deterioration by performing a series of maintenance activities that retard the deterioration process and essentially delay the type of substantial failure that requires major rehabilitation or reconstruction. Considering that a typical cost of the maintenance is 15% to 20% of the cost for rehabilitation or reconstruction (Hajj et al. 2010), agencies are more focused on preserving and maintaining existing facilities (Davies and Sorenson 2000, Labi and Sinha 2003).

However, the challenge is to find the pavement segments that require maintenance as well as appropriate times to execute such activities. Hence, there is a need to develop a proactive approach to identify potential pavement segments for improvement. Pavement performance models (PPMs) – one of several critical components required to achieve this proactive approach – seek to capture historical patterns of pavement deterioration that can be used to estimate an appropriate time for maintenance. Thus, the condition of pavements can be improved before a serviceability limit is reached.

In practice, it is very important to achieve a balance among the number of PPMs; the number of explanatory variables; the resources required to develop, maintain, and use these models; and the associated explanatory power. To seek this balance, PPMs typically are developed using clusters of pavement segments. Instead of estimating the cluster memberships by using statistical methods, a few predefined explanatory variables are used to assign pavement segments into clusters. In terms of performance, clusters thus formed likely include heterogeneous pavement segments.

Existing state-of-the-art methods propose Clusterwise Linear Regression (CLR) to determine pavement clusters and associated PPMs simultaneously, using a single objective function. In CLR, different clusters are formed so that pavement segments assigned within a cluster are homogenous in terms of the effects of the explanatory variables on the dependent variable (Park et al. 2015). In other words, the homogeneity of pavement segments in a cluster is defined by similarities of the observed values of explanatory variables and largely by the proximity of segments with respect to an underlying PPM (Preda and Saporta 2007). Hence, observations of all the pavement segments assigned to a cluster fit the same PPM, with minimum prediction error.

CLR was first implemented by Spath (1979) for data partition and estimation of regression models within each cluster, simultaneously. The approach has been expanded further and implemented in many studies (DeSarbo et al. 1989, Wedel and SteenKamp 1989, Lau et al. 1999, Carbonnea et al. 2011, Schlittgen 2011, Zhen et al. 2012, Tan et al. 2013, Lu et al. 2014). In the field of pavement management, to the best knowledge of the authors, only three studies (Luo and Chou 2006, Luo and Yin 2008, Zhang and Durango-Cohen 2014) have been performed using CLR. In a recent study (Zhang and Durango-Cohen 2014), CLR with multiple explanatory variables was proposed to account for heterogeneity in pavement deterioration. Potential overfitting issues were investigated using the methodology proposed by Brusco et al. (2008). The study used the data collected during the AASHTO Road Test (Highway Research Board 1962), which is no longer the best available data nor representative of existing conditions. However, this data was collected at a single site, and over 50 years ago, when materials and construction techniques were different.

In order to address some of the limitations of previous models, a mathematical programming framework within the CLR approach is proposed to determine simultaneously the optimal number of clusters, the assignment of segments into clusters, and the associated PPMs. The Bayesian Information Criteria (BIC) (Schwarz 1978) was used as the objective function in order to impose penalties for the inclusion of additional parameters. Minimizing BIC reduces unexplained variation, and the search process for the best model specification has the property of consistency (Rao and Wu 1989, Yang 2005, Maydeu-Olivares and García-Forero 2010, Vrieze 2012).

In addition, the proposed framework tests the significance of explanatory variables. To the best of the authors' knowledge, all the existing literature about pavement management and PPMs estimation using CLR suffers from this limitation, which is that variables included in the PPMs are assumed to be significant. However, the effects of insignificant explanatory variables affect clustering and regression analyses. Therefore, the true-cluster members are mixed (Fowlkes et al. 1988), and so it becomes challenging to discover the underlying pavement clusters that exhibit similar performance behaviors (Gupta and Ibrahim 2007).

This problem is illustrated in Figure 3.1, using data from the Pavement Management System (PMS) of the Nevada Department of Transportation (NDOT). Fifty-four randomly selected pavement segments were considered in this example. Each pavement segment was represented by a dependent variable, a Present Serviceability Index (PSI), and two explanatory variables, Age and Average Daily Traffic (ADT). The variables PSI and Age have significant linear relationship, as shown in Figure 3.1a. However, the relationship between PSI and ADT is insignificant, as shown in Figure 3.1b. If ADT was included in a CLR analysis without checking its significance, the resulting clustering and regression models would not represent the underlying relationships among the variables; that is, the underlying pavement-performance behaviors captured by PSI and Age would be blurred.

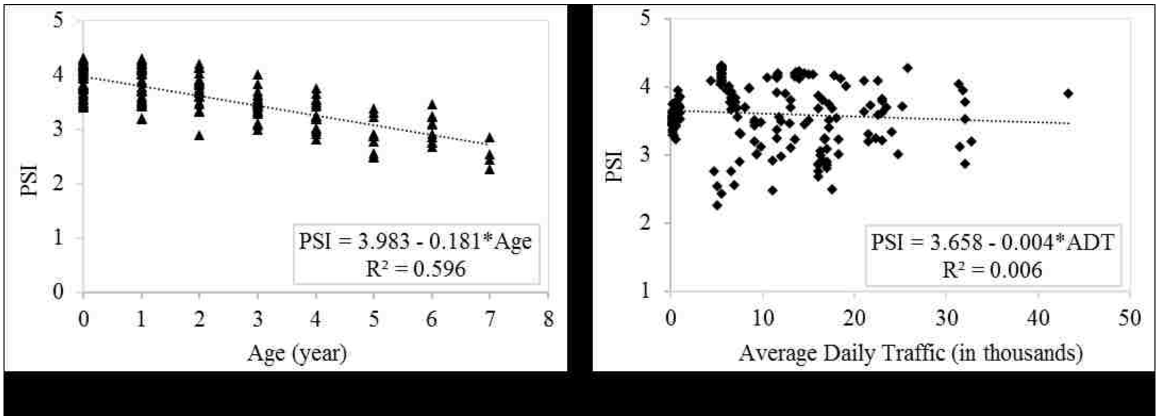

Figure 3.1 Linear relationship between PSI and (a) Age and (b) ADT.

Assignment of pavement segments into clusters using predefined and fixed explanatory variables, instead of estimation, introduces bias into the statistical analysis (Gupta and Ibrahim 2007). The available data are not fully utilized for clustering, as the performance behavior represented by historical PSI is ignored. In addition, clustering using explanatory variables that do not provide any information about the underlying clustering structure does not reveal the true cluster assignments.

A legitimate assignment of pavement segments into clusters that is closer to the true underlying clusters can be obtained using the relevant explanatory variables that exhibit the strongest effects on the performance measure (Fowlkes et al. 1988, Liu and Ong 2008, and Maugis et al. 2009). The strength of the effects of explanatory variables on the dependent variables often is assessed by comparing p-values with the desired level of significance (α). If the p-value of an explanatory variable is greater than α, the explanatory variable is considered as insignificant; in other words, changes in the explanatory variables do not reflect changes in the dependent variable. Therefore, such explanatory variables with p-values less than the desired α need to be excluded during the model development process.

A variable selection procedure can be utilized to select the best subset of potential explanatory variables. This procedure must distinguish between relevant and irrelevant variables, thus providing a true regression model using underlying clusters. A number of variable selection methodologies are available in the literature for data analysis and statistics (Thompson 1978, Tibshirani 1996, Baumann 2003, Efron et al. 2004, Mehmood et al. 2012, Brusco 2014). In this study, the All Subsets regression procedure (Garside 1965, Gorman and Toman 1966, Hocking and Leslie 1967, Mallows 1973, Berk 1978, Efron et al. 2004) was used to select variables to use in the CLR analysis. All (2^P-1) possible subsets of P potential explanatory variables were examined. BIC was used as a criterion for comparing models with different subsets of variables.

It is not recommended to use least squares estimation and variable selection techniques for data with multicollinearity (Gunst and Webster 1975). Strongly-correlated clustering variables may overweight one or more underlying constructs (Ketchen and Shook 1996). Multicollinearity among explanatory variables in a regression equation can make it challenging to identify significant variables correctly (Abdul-Wahab et al. 2005). Therefore, this study investigated the effects of highly-correlated explanatory variables, and the Variance Inflation Factor (VIF) was used to examine potential issues due to multicollinearity. As the degree of collinearity increases, both the variance of regression coefficient and the VIF increase (Yoo et al. 2014). Large VIF is an indicator of multicollinearity (Tacq 1997). In general, a VIF greater than

10 is considered unacceptable (Neter et al. 1996, Midi et al. 2010), even though no formal rule exists in the body of literature.

In order to avoid pre-specifying the significance of potential explanatory variables, this chapter proposes a comprehensive CLR framework that determines, simultaneously, the optimal number of pavement clusters, the assignment of segments into clusters, and the corresponding PPMs using only significant explanatory variables. That is, the proposed framework simultaneously seeks for 1) the optimal number of clusters, 2) the combination of significant explanatory variables that provides the best goodness of fit, and 3) assigns segments into clusters. The significance of the explanatory variables is tested for each cluster model. Hence, different clusters may include different significant explanatory variables.

Considering the simultaneous and extensive search for significant explanatory variables and the optimal number of clusters, the PPMs developed under the proposed framework were expected to provide superior explanatory power compared to existing approaches. The proposed framework was tested using pavement data from the entire State of Nevada. The results illustrated the advantage of solving simultaneously for the three types of parameters listed above.

3.2 Methodology

3.2.1 Problem Formulation

This section describes a mathematical program that was formulated to describe the proposed CLR problem. This chapter uses the same notation and definitions of variables in Chapter 2. However, new variables and constraints were added to address the proposed problem. The following variables are used in this chapter:

I = Number of pavement samples in the network;

i = Subscript for a pavement sample in the network, $i \in I$;

T_i = Number of observation periods for a pavement sample i;

t = Subscript for an observation period for a pavement sample, $t \in T_i$;

O = Total number of observations = $\sum_i^I T_i \ \forall \ i \in I$;

J = Number of explanatory variables;

j = Subscript for an explanatory variable including an intercept, $j = 0, \dots, J$;

x_{ijt}^k = Measurement of an explanatory variable j for a sample i at observation period t that is assigned to a cluster $k \ \forall \ i \in I, j \in J, t \in T_i$;

y_{it}^k = Measurement of dependent variable for a sample i at observation period t that is assigned to a cluster $k \ \forall \ i \in I, t \in T_i$;

K = Optimum number of clusters $(1 \leq k \leq K_{max})$;

k = Subscript for a clusters, $k \in K$;

K_{max} = Maximum number of potential clusters that could be formed;

n = Minimum number of observations required in a cluster;

C_k = Set of pavement samples that are assigned to cluster $k \ \forall \ k \in K$;

δ = Total number of significant explanatory variables including intercepts in all clusters;

v_{jk} = Binary indicator that represents significance of an explanatory variable including an intercept in a cluster $k \ \forall \ j = 0, \dots, J, \ k \in K$;

p_{ik} = Cluster membership of a pavement sample i to a cluster k, $\forall \ i \in I, k \in K$;

β_{jk} = Estimated regression coefficient for an explanatory variable j including an intercept in cluster $k \ \forall \ j = 0, \dots, J, k \in K$;

Various pavement performance measures are available in the literature. PSI, which is a widely accepted measure that serves as a unified standard to measure pavement serviceability

(Shoukry et al. 1997, Terzi 2006, Attoh-Okine and Adarkwa 2013), is easily understood by both road users and legislators (Hudson et al. 2015). This study used PSI as the dependent variable, *y*. Multiple linear regression PPMs were estimated with functional form expressed by:

$$y_{it}^{k} = \beta_{0k} + \sum_{j=1}^{J} \beta_{jk} * x_{ijt}^{k} \qquad (3.1)$$

The objective function was to minimize BIC, as illustrated by Equation 3.2. Intercepts, β_{0k}; coefficients for cluster-specific significant explanatory variables, β_{jk}; the optimum number of clusters, *K*; and the cluster memberships, p_{ik}, were the decision variables to be determined:

$$Min.\ BIC = O+O*ln(2\pi)+O*ln\left(\frac{SSE}{O}\right)+(\delta+K\text{-}1)*ln(O) \qquad (3.2)$$

where, *SSE* is total sum of squared errors expressed by:

$$SSE= \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{t=1}^{T_i}\left[\beta_{0k}+\sum_{j=1}^{J}\beta_{jk} * x_{ijt}^{k} \text{ - } y_{it}^{k}\right]^{2} * p_{ik}\ \forall\ i \in I,\ t \in T_i,\ j \in J,\ k \in K \qquad (3.3)$$

and the quantity (*δ*+*K*-1) is the total number of free parameters to be estimated for *K* clusterwise regression models (DeSarbo and Corn 1988). Intercepts (β_{0k}), coefficients for cluster-specific significant explanatory variables (β_{jk}), the optimum number of clusters (*K*), and cluster memberships (p_{ik}) were the decision variables to be determined.

The proposed mathematical programming included the following constraints:

$$\delta = \sum_{k}\sum_{j} v_{jk}\ \forall\ j = 0,\ldots,J,\ k \in K \qquad (3.4)$$

$$v_{jk} = \begin{cases} 1, & if\ \beta_{jk}\ is\ significant; \\ 0, & Otherwise \end{cases} \quad \forall\ j = 0,\ldots,J,\ k \in K \qquad (3.5)$$

$$\sum_{k} p_{ik} = 1 \ \forall i \in I,\ k \in K \tag{3.6}$$

$$p_{ik} = \begin{cases} 1, & \textit{if sample } i \textit{ is assigned to cluster } k; \\ 0, & \textit{Otherwise} \end{cases} \quad \forall\ i \in I,\ k \in K \tag{3.7}$$

$$C_k = \left\{ i \middle| p_{ik} = 1 \forall i \in I,\ k \in K \right\} \tag{3.8}$$

$$C_{k'} \cap C_{k''} = \text{null} \ \ \forall k' \neq k'',\ k' \text{ and } k'' \in K \tag{3.9}$$

$$\bigcup_{k \in K} |C_k| = I \tag{3.10}$$

$$\sum_{i \in C_k} T_i \geq n \ \forall\ C_k \tag{3.11}$$

$$1 \leq k \leq K_{max} \tag{3.12}$$

$$K_{max} = F(I,\ T_i,\ n) \tag{3.13}$$

Constraint 3.4 provided the total number of significant explanatory variables, including intercepts for all the clusters. The sum of elements in each column of the binary matrix, **V**, of size (*J*+1 x *K*) provided the number of significant explanatory variables and an associated intercept for a particular cluster. According to Constraint 3.5, the element v_{jk} was equal to 1 if an estimated coefficient (β_{jk}) was significant in cluster *k*; otherwise v_{jk} was zero (Equation 3.5). The significance of an explanatory variable as well as an intercept was determined by using the p-value of its estimated regression coefficient.

Constraints 3.6 and 3.7 ensured that a pavement sample was assigned exclusively to a single cluster. A binary indicator variable, p_{ik}, was used to define the membership of a sample.

Indicator p_{ik} equaled 1 if and only if a pavement sample i belonged to cluster k. Otherwise, p_{ik} was zero.

The feasibility of the resulting clustering was guaranteed by Constraints 3.8 – 3.11. Constraints 3.8 – 3.10 prevented the overlap of members among clusters; that is, pavement samples were divided exclusively into *K* clusters. Constraint 3.11 warranted that the number of observations for each cluster was no less than the minimum number of observations, *n*, in order to obtain the statistically reliable estimation of coefficients.

Constraints 3.12 and 3.13 were used to prevent a search beyond a feasible number of clusters. If the pavement sample had more than *n* observations, the sample alone could form a cluster. In reality, none of the pavement samples had more than *n* observations. Hence, samples were grouped into clusters to provide enough observations. All observations of a sample needed to be assigned to the same cluster. Constraint 3.13 denoted the maximum number of feasible clusters, which is represented by function F. The procedure to determine this maximum number is described in Chapter 2.

3.2.2 Solution to the Mathematical Program

This study integrated simulated annealing (SA) with ordinary least square (OLS) to solve the proposed mathematical program. SA determined the cluster memberships (p_{ik}) of the pavement samples. For each accepted cluster, the VIF for all explanatory variables were calculated. Highly correlated explanatory variables that had VIFs greater than a pre-defined limiting VIF were excluded. All subset regressions were utilized to find the best model and to estimate the associated coefficients (β_{jk}). BIC and the level of significance, α, were used as the criteria to select the best model. Hence, selected models included only significant explanatory variables at a given α.

The algorithm utilized to solve the proposed mathematical program is described as follows, and illustrated in Figure 3.2.

Step 1. Set $K = 2$, BIC_{min} = infinity, and $N = 1$.

Step 2. Calculate the maximum number of feasible clusters, K_{max}, utilizing function F, described above, as part of Constraint 3.13.

Step 3. For a given K, randomly assign pavement samples into clusters. If all the clusters have at least n observations, then go to Step 4; otherwise, reassign pavement samples into clusters until all the clusters have at least n observations. Let $C_K^N \; \forall \; 1 \leq k \leq K$ be the valid initial clusters.

Step 4. All subsets regression: Repeat the following steps for all K clusters.

Step 4.1. Calculate VIF for all explanatory variables. Exclude variables that have $VIF > VIF_{max}$. Let $\hat{J}$ be the set of explanatory variables with $VIF < VIF_{max}$.

Step 4.2. Generate all possible $2^{|\hat{J}|} - 1$ subsets of $\hat{J}$.

Step 4.3. Estimate β_{jk} for all subsets, using OLS, and calculate BIC for all the models.

Step 4.4. Rank models in ascending order, using BIC.

Step 4.5. Select the model that has the minimum BIC and all significant explanatory variables with $p\text{-}value < \alpha$.

Step 5. Calculate the total number of free parameters to be estimated, (δ+K-1). Calculate BIC using Equation 3.2.

Step 6. Using the following steps, generate valid neighborhood clusters near to the previous ones.

Step 6.1. Select N_{ps} pavement samples randomly. For each of the selected samples, assign a new membership by generating a random number $u \sim U(1, K)$. If the new membership is the same as previously, regenerate a random number $u' \sim U(1, K)$ until a different outcome is obtained. Repeat this process until the memberships of all selected samples are different from those previously assigned.

Step 6.2. If all clusters have at least *n* observations, go to Step 7; otherwise, repeat Step 6.1. until all clusters have at least *n* observations. Let C_K^{N+1} be the new set of valid neighborhood clusters.

Step 7. For C_K^{N+1}, repeat Step 4 to estimate β_{jk} for all *K* clusters.

Step 8. Calculate the total number of free parameters to be estimated, (δ+*K*-1), and evaluate BIC_K^{N+1}, using the Equation 3.2.

Step 9. Search of a solution.

Step 9.1. Calculate $\Delta BIC = BIC_K^{N+1} - BIC_K^N$.

Step 9.2. Check the following two conditions:

a. If $\Delta BIC < 0$, accept current set of clusters, C_K^{N+1}, and the corresponding β_{jk}; go to Step 10, otherwise, go to Step b.

b. Generate a random number $u" \sim U(0,1)$. Calculate acceptance probability, $p_{accept} = exp\left(\frac{-\Delta BIC}{B*T}\right)$, where *B* is the Boltzmann's constant. If $p_{accept} > u"$, accept current set of clusters, C_K^{N+1}, and the associated β_{jk}; go to Step 10, otherwise, return to Step 6.

Step 10. Counter and temperature update:

Step 10.1. Repeat Steps 6 to 9 for N_{max} times.

Step 10.2. If $\theta < \theta_{min}$, stop the algorithm. Otherwise, reduce the temperature by multiplying the current temperature by λ, set *N* =1, and go to Step 6.

Step 11. Stopping criteria:

Step 11.1. Update BIC_{min} with the smallest between the values obtained in Step 10 and the current BIC_{min}. Set $K_{optimal} = K$.

Step 11.2. Repeat Steps 3 to 10 for $K_{max} - 1$ times.

To seek for a global solution, this algorithm uses a probabilistic approach during the search process. The initial solution is improved repetitively by making small changes until a better solution is obtained (Sridhar and Rajendran 1993, Johnson et al. 1989). The algorithm accepts better solutions and also non-improving (worse) solutions at a certain probability (Dolan et al. 1989, Rutenbar 1989, Aarts et al. 2005). This probability decreases continuously over iterations and depends on 1) the difference between the BICs of the current solution and a newly selected solution, and 2) the current temperature (Nikolaev and Jacobson 2010).

Initially, at a high temperature, the algorithm accepts worse solutions that cause larger increments in BIC. As the temperature goes down, the algorithm accepts worse solutions with relatively smaller increments in BIC. Finally, when the temperature drops to zero, the algorithm no longer accepts worse solutions. This enables occasional 'uphill' moves, which help the algorithm to escape from the local minima. Thus, the algorithm tries to explore the entire solution space to seek for a global solution (Dolan et al. 1989). Previous studies have shown that the algorithm converges to a global minimum when an infinitely slow cooling schedule is utilized (Román-Román et al. 2012).

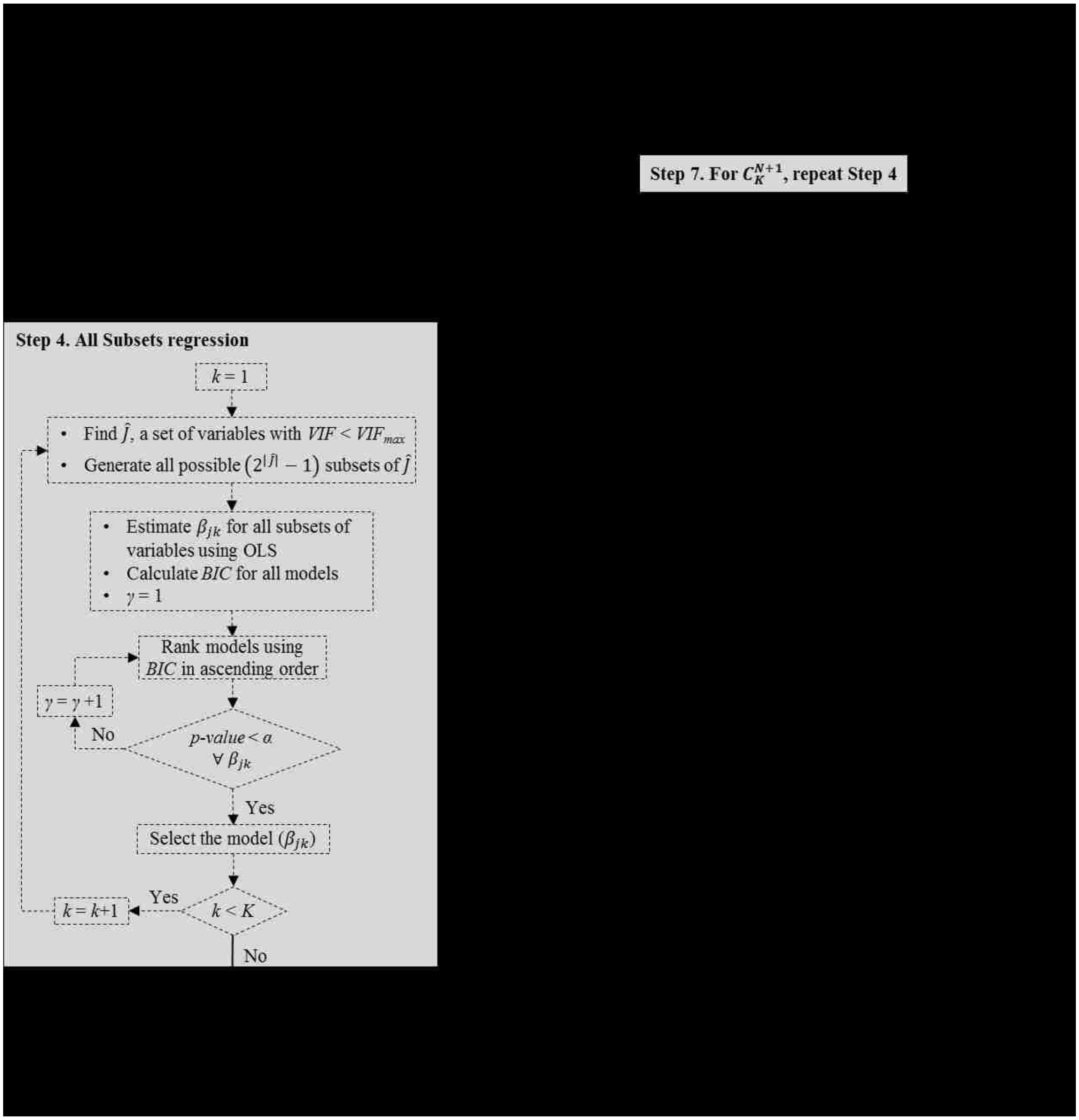
Step 7. For C_K^{N+1}, repeat Step 4
Step 4. All Subsets regression
$k = 1$
Find $\hat{J}$, a set of variables with $VIF < VIF_{max}$
Generate all possible $(2^{|\hat{J}|} - 1)$ subsets of $\hat{J}$
Estimate β_{jk} for all subsets of variables using OLS
Calculate BIC for all models
$\gamma = 1$
Rank models using BIC in ascending order
$\gamma = \gamma + 1$
No
$p\text{-}value < \alpha$ $\forall\, \beta_{jk}$
Yes
Select the model (β_{jk})
Yes
$k = k+1$
$k < K$
No

Figure 3.2 Algorithm used to solve the comprehensive clusterwise linear regression.

3.3 Numerical Experiment and Results

3.3.1 Experimental Research Data

Experiments were performed using the PMS of NDOT. The data included condition monitoring and roadway inventory data collected throughout the State of Nevada. Potential explanatory variables used in this study are illustrated in Table 3.1. A total of 4,138 samples having 14,638 observations (from 2001 to 2010) and 3,005 observations (2011 and 2012) were available for model estimation and validation, respectively. The detailed descriptions of the data were provided in Chapter 2.

Table 3.1 Variables Used in the Pavement Performance Models

Variable	Description
age	Age of the last M&R treatment performed on a segment
adt	One direction average daily traffic
trucks	One direction average daily trucks
elevation	Elevation at midpoint of a segment (m)
precip	Average annual precipitation (cm/year)
min_temp	Minimum average yearly air temperature (^{0}C)
max_temp	Maximum average yearly air temperature (^{0}C)
wet_days	Total number of wet days (days that moisture was recorded) over the course of one year
freeze_thaw	Total number of freeze-thaw cycles that a pavement experienced over the course of one year
rut_depth	Average ride rut depth (cm)
lane=2	Dummy variable for a segment that has 2 lanes (1 = yes, 0 = no)
lane≥3	Dummy variable for a segment that has 3 or more lanes (1 = yes, 0 = no)
sys_id=2	Dummy variable for a segment that is part of NHS (1 = yes, 0 = no)
sys_id=3	Dummy variable for a segment that is part of STP (1 = yes, 0 = no)
f_class=2	Dummy variable for a segment classified as functional class 2 (1=yes, 0 = no)
f_class=3	Dummy variable for a segment classified as functional class 3 (1 = yes, 0 = no)
f_class=4	Dummy variable for a segment classified as functional class 4 (1 = yes, 0 = no)
f_class=5	Dummy variable for a segment classified as functional class 5 (1 = yes, 0 = no)
f_class=6	Dummy variable for a segment classified as functional class 6 (1 = yes, 0 = no)
f_class=7	Dummy variable for a segment classified as functional class 7 (1 = yes, 0 = no)
category=2	Dummy variable for a segment grouped in prioritization category 2 (1 = yes, 0 = no)
category=3	Dummy variable for a segment grouped in prioritization category 3 (1 = yes, 0 = no)
category=4	Dummy variable for a segment grouped in prioritization category 4 (1 = yes, 0 = no)
category=5	Dummy variable for a segment grouped in prioritization category 5 (1 = yes, 0 = no)

3.3.2 Estimation Parameters

The existing literature does not provide hard-and-fast rules to define the limiting VIF beyond the one that indicates a serious multicollinearity problem (Petraitis et al. 1996). Many studies (Myers 1990, Neter et al. 1996, Chatterjee and Hadi 2000) suggested that a multicollinearity problem was serious for greater than 10 VIFs. In this study, all explanatory variables with VIF > 10 were excluded from the final models. Other estimation parameters that were required were set using experience of the research team (Paz et al. 2015a and b, Khadka and Paz 2017, Paz and Khadka 2017) and sensitivity analyses. Table 3.2 provides the parameter values used in this study.

Table 3.2 Estimation Parameters Used in the Experiments

Parameter	Value	Remarks
θ_0	10	Initial temperature
θ_{min}	10e-17	Final minimum temperature
B	30	Boltzmann constant
λ	0.97	Cooling rate
N_{max}	5	Number of neighborhood solutions generated at each temperature level
n	800	Minimum number of observations required in a cluster
N_{ps}	100	Number of pavement samples, which memberships were changed to generate a neighborhood cluster
VIF_{max}	10	Limiting VIF
α	5%	Level of Significance

3.3.3 Results and Discussion

Function F in Constraint 3.13 was used to determine the maximum number of feasible clusters for the data set used in this study. The algorithm found that 16 was the maximum number of feasible clusters that fulfilled the requirements imposed by the constraints for feasible partitions.

The solution algorithm proposed in the section, Solution to the Mathematical Program, sought for the optimum number of clusters by exploring each of all feasible clusters (i.e., K = 2 to 16). Thus, the algorithm determined that 6-cluster CLR models provided the optimum solution with the lowest BIC.

Figure 3.3a shows the BIC trend over the number of clusters that were considered in this experiment. Figure 3b shows the convergence of the objective function, BIC, over iterations when the six-cluster CLR models were used. After 983 iterations, the BIC decreased from the initial value of 9,283 to the final value of 6,443, with an improvement of 31%.

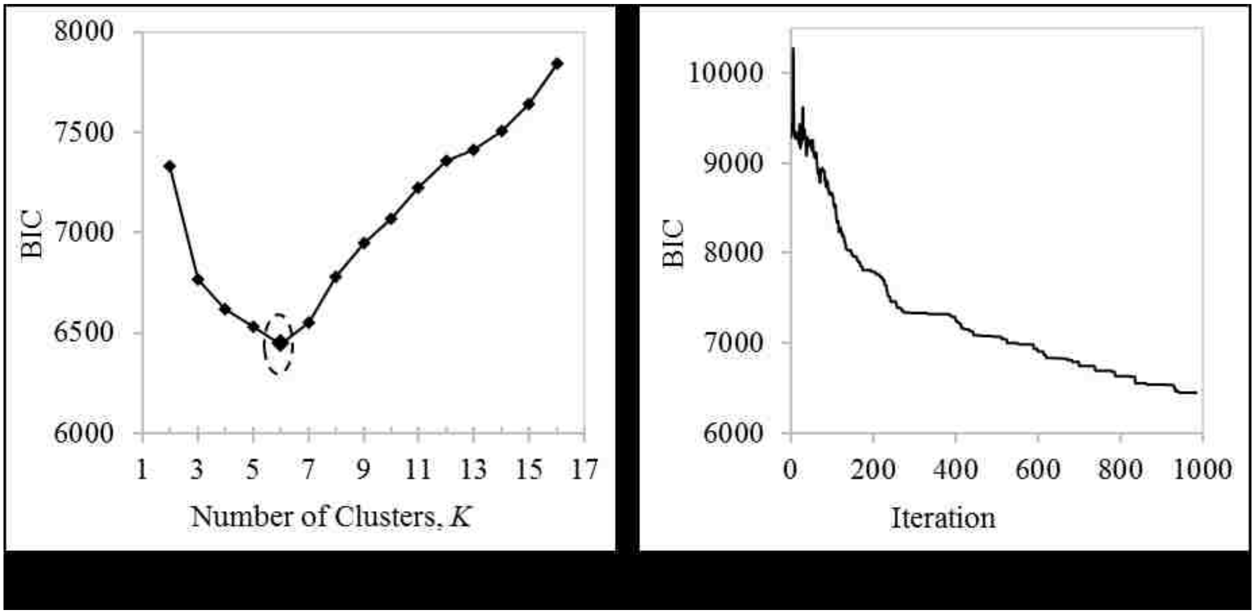

Figure 3.3 BIC versus (a) the number of clusters and (b) iterations for 6-cluster models.

Coefficients for the variables, *trucks* and *freeze_thaw*, were positive. This is counter-intuitive because a pavement deteriorates faster when it is subject to a large number of trucks and frequent freeze-and-thaw cycles. Hence, additional data analysis was performed to investigate the data quality. The analysis showed average positive trends of PSI for these variables. This indicates significant data collection or management errors for *trucks* and *freeze-thaw*. Hence, these two variables were excluded from the models, and new model parameters were estimated. Table 3.3 provides the estimated parameters for 6-cluster models.

Table 3.3 Estimated Model Parameters Using the Proposed CLR Approach

Parameters	Cluster #1			Cluster #2			Cluster #3		
	β_{j1}	VIF	p-value	β_{j2}	VIF	p-value	β_{j3}	VIF	p-value
intercept	4.392	-	< 0.0001	4.552	-	< 0.0001	4.674	-	< 0.0001
age	-0.039	1.0	< 0.0001	-0.022	1.0	< 0.0001	-0.028	1.0	< 0.0001
adt†	-0.013	1.2	< 0.0001	-0.012	1.8	< 0.0001	-0.008	2.2	< 0.0001
rut_depth	-0.509	1.1	< 0.0001	-1.108	1.1	< 0.0001	-1.314	1.1	< 0.0001
lane=2	-	-	-	-0.191	4.4	< 0.0001	-0.358	4.4	< 0.0001
lane≥3	-	-	-	-0.202	1.8	< 0.0001	-0.289	2.5	< 0.0001
f_class=2	-0.185	1.0	0.002	-	-	-	-	-	-
f_class=3	-0.110	1.6	< 0.0001	-	-	-	-	-	-
f_class=4	-0.259	1.5	< 0.0001	-	-	-	-	-	-
f_class=5	-1.052	1.4	< 0.0001	-	-	-	-	-	-
f_class=6	-1.181	1.1	< 0.0001	-	-	-	-	-	-
f_class=7	-0.284	1.0	0.006	-	-	-	-	-	-
category=2	-	-	-	-0.202	2.6	< 0.0001	-0.325	2.8	< 0.0001
category=3	-	-	-	-0.323	4.2	< 0.0001	-0.465	4.4	< 0.0001
category=4	-	-	-	-0.664	2.6	< 0.0001	-0.684	2.9	< 0.0001
category=5	-	-	-	-1.149	2.8	< 0.0001	-0.808	2.8	< 0.0001
No. of Obs.		2,376			2,483			2,442	
BIC		658			1,069			1,470	

Parameters	Cluster #4			Cluster #5			Cluster #6		
	β_{j4}	VIF	p-value	β_{j5}	VIF	p-value	β_{j6}	VIF	p-value
intercept	4.605	-	< 0.0001	4.557	-	< 0.0001	4.401	-	< 0.0001
age	-0.033	1.0	< 0.0001	-0.028	1.0	< 0.0001	-0.037	1.0	< 0.0001
adt†	-0.006	1.8	< 0.0001	-0.005	2.2	< 0.0001	-0.013	1.4	< 0.0001
rut_depth	-1.459	1.1	< 0.0001	-1.295	1.1	< 0.0001	-0.902	1.0	< 0.0001
lane=2	-0.213	4.8	< 0.0001	-0.260	4.9	< 0.0001	-	-	-
lane≥3	-0.405	1.9	< 0.0001	-0.294	2.4	< 0.0001	-	-	-
f_class=2	-	-	-	-	-	-	0.468	1.2	< 0.0001
f_class=3	-	-	-	-	-	-	-0.086	1.5	< 0.0001
f_class=4	-	-	-	-	-	-	-0.258	1.4	< 0.0001
f_class=5	-	-	-	-	-	-	-0.864	1.3	< 0.0001
f_class=6	-	-	-	-	-	-	-1.288	1.1	< 0.0001
f_class=7	-	-	-	-	-	-	-0.634	1.0	< 0.0001
category=2	-0.263	3.0	< 0.0001	-0.194	2.7	< 0.0001	-	-	-
category=3	-0.325	4.0	< 0.0001	-0.287	4.2	< 0.0001	-	-	-
category=4	-0.650	3.2	< 0.0001	-0.639	3.1	< 0.0001	-	-	-
category=5	-0.808	3.4	< 0.0001	-1.130	2.9	< 0.0001	-	-	-
No. of Obs.		2,414			2,340			2,583	
BIC		1,009			1,273			870	

† variable value in thousands.

- = *Not applicable.*

This study used a 5% significance level. Results showed that seven explanatory variables – *elevation*, *precip*, *min_temp*, *max_temp*, *wet_days*, *sys_id=2*, and *sys_id=3* – were not included in any of the resultant six models. As the constraints for significant variables were imposed, the algorithm excluded these seven variables because they were either associated with high VIF, causing multicollinearity, or were statistically insignificant. Hence, the resultant models only had statistically significant explanatory variables.

Table 3.3 also includes the VIFs of the significant explanatory variables. All the VIF values were less than five, which indicated that the associated explanatory variables in each model did not have strong correlations among each other. Hence, the resultant models were free from serious multicollinearity problems.

The six models included different significant explanatory variables. In addition, the common variables had different estimated coefficients. These differences indicated that pavement samples across the clusters were heterogeneous by the effect of explanatory variables, and exhibited different types of performance behavior. For example, the samples exhibited different deterioration rates as they got older. The estimated coefficients for *age* were -0.039 and -0.022 for Clusters #1 and #2, respectively. However, pavement samples in Clusters #1 and #2 performed similarly with respect to traffic-loading conditions. That is, the estimated coefficients for *adt* in Clusters #1 and #2 were -0.013 and -0.012, respectively.

Only four variables – *intercept*, *age*, *adt*, and *rut_depth* – were common for all six models; and all of them had a negative sign, except for the intercept. All the estimated intercept values were realistic. The PSI of a newly constructed pavement was about 4.5 (Christopher et al. 2006). However, the intercepts differed across the models. The negative signs of *age* and *adt* indicated that the conditions deteriorated when a pavement became older and was subjected to greater traffic loadings, respectively. Similarly, the PSI of a pavement segment decreased as rutting along the pavement became deeper.

It was observed that Clusters #2 to #5, which had as significant variables *category=2*, *category=3*, *category=4*, and *category=5*, also had variables *lane=2* and *lane=3* as significant. In contrast, the variable *f_class* was not significant in these clusters. The estimated coefficients of the variables *category=2*, *category=3*, *category=4*, and *category=5* were negative, and the coefficient increased as the category level went up. This indicated that the average PSIs in these four category levels (i.e., from 2 to 5) were smaller than for that of Category 1, and decreased as the level went up. This was expected, because NDOT assigned the highest priority – in terms of maintaining good conditions – to the roadway segments identified as Category 1 and the lowest priority to the roadway segments identified as Category 5 (NDOT 2011). The variable *f_class* was significant only in Clusters 1 and 6. The coefficients for all six classes were negative, except for the *f_class=2* in Cluster 6. The positive sign indicated that the pavement segments classified as Class 2 had a higher average PSI than for the segments classified as Class 1. It also was observed that for both clusters, the coefficient increased as the class number went up, except for the *f_class=7*. A possible reason was that the estimation was based on only 44 observations (Functional Class 7), which might not represent the reality.

3.3.4 Model Performance

Brusco et al. (2008) proposed a procedure to diagnose the presence of overfitting in the resultant CLR models. Five different metrics were calculated for the optimum 6-cluster models, which are included in Table 3.4. The results showed that the between-clusters sum of squares (BCSS) was equal to 4, which was less than 1% of the total sum of squares (TSS). The sum of squares due to regression (SSR) was equal to 1,130, which was 47% of the within-clusters sum of squares (WCSS). This indicates that there was no overfitting, as most of the variation in PSI was explained by the within-cluster regressions. However, SSE accounted for 53% of the TSS. This indicated that the resultant models had relatively high errors, which could be due to the nature of the data. In addition, the estimated linear function might not have been the best to use to explain the pavement deterioration.

Table 3.4 Metrics Calculated to Investigate the Presence of Overfitting in the Models

Metric	Value	Remarks
TSS	2,419	-
BCSS	4	0.17% of TSS
WCSS	2,415	-
SSR	1,130	47% of WCSS
SSE	1,284	53% of TSS

The prediction accuracy of the models was evaluated by calculating the root-mean-square error (RMSE), the normalized root-mean-square error (NRMSE), and the mean absolute error (MAE), using Eqs. 3.14, 3.15, and 3.16, respectively.

$$RMSE = \sqrt{\frac{\sum_{1}^{\eta}\left(y_{it}^{k} - \hat{y}_{it}^{k}\right)^{2}}{\eta}} \tag{3.14}$$

$$NRMSE = \frac{RMSE}{y_{\max} - y_{\min}} \tag{3.15}$$

$$MAE = \frac{1}{\eta}\sum_{1}^{\eta}\left|y_{it}^{k} - \hat{y}_{it}^{k}\right| \tag{3.16}$$

where, y_{it}^{k} = the observed PSI, $\hat{y}_{it}^{k}$ = the predicted PSI, y_{max} = the maximum observed PSI, y_{min} = the minimum observed PSI, and η = the number of predictions.

The estimated model coefficients were applied to the test data set – which is described in the section, Data Resources – to estimate PSIs for 2011 and 2012. The overall RMSE, NRMSE, and MAE values for all the models were 0.47, 0.17, and 0.36, respectively. This indicated that the resultant models were robust. In addition, to diagnose the variation in the prediction errors, the RMSE, NRMSE, and MAE were calculated separately for all six models.

Table 3.5 provides the RMSE, NRMSE, and MAE values for all the models as well as the individual models. It was observed that the differences between RMSE and MAE values were approximately equal for all the models, which indicated that the prediction errors were well distributed among the clusters.

Table 3.5 Model Performance Error Metrics: RMSE, NRMSE, and MAE for Each of the Clusters

Cluster	RMSE	NRMSE	MAE
1	0.47	0.18	0.37
2	0.46	0.18	0.37
3	0.49	0.18	0.37
4	0.47	0.17	0.35
5	0.48	0.18	0.36
6	0.49	0.19	0.38
Overall	0.47	0.17	0.36

Figure 3.4a shows the scattered plot of predicted versus observed PSIs for 2011 and 2012. The degree of prediction error of the models is reflected by the relative positions of data points from the 45^0 line. Data points above the 45^0 line are over-predicted while those under the 45^0 line are under-predicted. Figure 3.4b provides the percentages of observations that were within different ranges of error. For example, about 74% of the total observations were contained within a ±15% range of error. Figure 3.5 shows individual scattered plot of predicted versus observed PSIs for all six models.

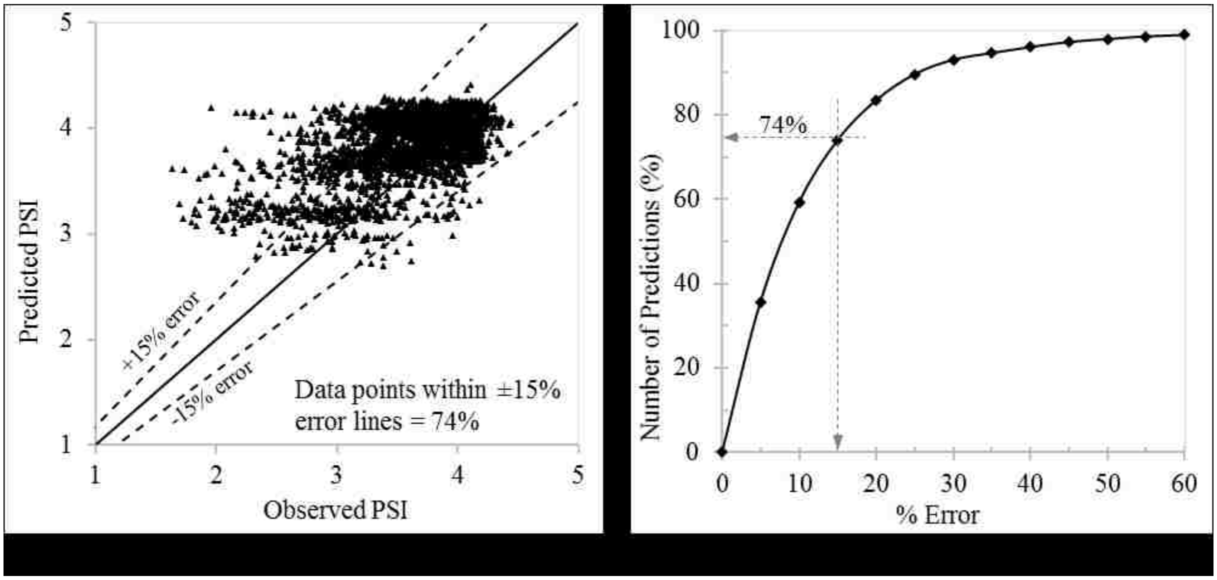

Figure 3.4 Observed versus predicted PSIs (a), and percentage of predictions within different ranges of error (b).

3.4 Conclusions

In this chapter, a comprehensive mathematical program is proposed to estimate PPMs that minimize the estimation error by simultaneously finding 1) the optimum number of pavement clusters, 2) cluster memberships of the samples, 3) cluster-specific significant explanatory variables, and 4) regression coefficients. To solve the mathematical program, Simulated Annealing integrated with All Subsets Regression was implemented. The algorithm has the capability to identify potential explanatory variables that cause serious multicollinearity in a model. In addition, the algorithm addresses multicollinearity by removing the potential explanatory variables one at a time until the effect of multicollinearity is minimal.

In this study, VIF was used to measure the effect of multicollinearity in a model. After addressing the multicollinearity issue, the proposed algorithm identified the relevant explanatory variables to include in the models. All possible combinations of the explanatory variables were evaluated to select the best model for each cluster. Hence, the resultant CLR models included cluster-specific significant explanatory variables that were free from multicollinearity.

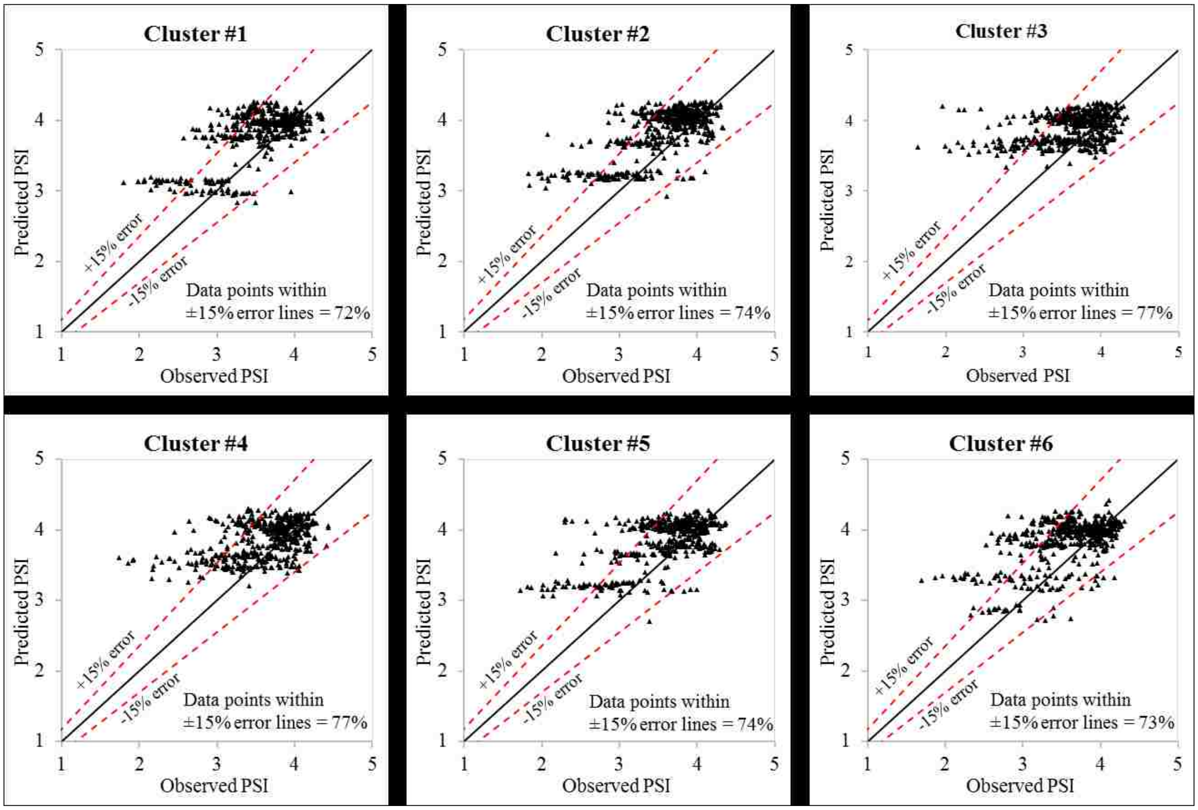

Figure 3.5 Observed versus predicted PSIs for each of 6-cluster models.

The algorithm explored all the feasible clusters that could be formed for the data used in the experiments, and found that 6-cluster models were the optimum solution. The algorithm determined traffic-loading conditions of both ADT and the number of trucks, age, rut-depth, function class, prioritization category, freeze-and-thaw cycles, and the number of lanes as significant explanatory variables. In the literature, all these variables were considered to be the most critical factors for pavement deterioration (Saraf and Majidzadeh 1992, Prozzi and Madanat 2004, Kim and Kim 2006, Salama et al. 2006). Both the magnitude and sign of the estimated regression coefficients were as expected, and were realistic. This indicates that the proposed algorithm is very effective when selecting the underlying explanatory variables that are truly relevant.

The resulting CLR models first were analyzed to investigate the presence of overfitting. The results show that the models did not possess any overfitting issues. In order to investigate the predictive capability of models, RMSE, NRMSE, and MAE were calculated for all the models as well as for individual models. The overall RMSE, NRMSE, and MAE values of 0.47, 0.17, and 0.36, respectively, indicated that the resultant models were robust.

In addition, the results showed that both the differences between the RMSE and MAE values for all six models were approximately equal. This indicated that the prediction error was well distributed among the models. Even so, the models still were associated with prediction errors. Moreover, the linear functional form used in this study did not exactly fit the data used in the experiments. Hence, it is worth investigating the proposed methodology by allowing nonlinear relationships between the pavement performance measures and multiple explanatory variables. Various forms of power and sigmoidal models (Sadek et al. 1996, Luo and Chou 2006, Zhang and Durango-Cohen 2014, and Chen and Mastin 2015) could be investigated.

Finally, the results indicated that each cluster had almost an equal number of members (i.e., pavement samples. However, it is unlikely that the underlying clusters had equally distributed pavement samples. An interesting aspect worthy of investigation would be to explore the likelihood of distribution of the pavement samples and the associated physical characteristics.

CHAPTER 4

ESTIMATION OF PAVEMENT PERFORMANCE MODELS USING NONLINEAR CLUSTERWISE REGRESSION

4.1 Introduction

A variety of modelling approaches, including empirical, mechanistic, and mechanistic-empirical approaches, have been investigated and implemented for pavement performance modelling (George et al. 1989, Li et al. 1997, Zheng 2005, Hong and Prozzi 2006, Bardaka et al. 2014, Chen and Mastin 2015). Mechanistic models are developed using mechanical and engineering properties of pavement materials, such as stress, strain, and deflection (Lytton 1987, Schmitt et al. 2008). However, quantification of the exact mechanistic behaviour is very challenging because pavement deterioration is a very complex process governed by many factors (Prozzi and Madanat 2003). Hence, modelling approaches that use mechanical properties are less preferred (Schram 2008). In contrast, empirical modelling approaches commonly are used to develop pavement performance models (PPMs). Empirical models are estimated using historical pavement data and statistical techniques (Prozzi and Madanat 2003). Model specifications, such as functional form and potential explanatory variables, are chosen based on physical considerations, estimation error, and experience (George et al. 1989, Madanat et al. 2008).

Empirical models can be categorized further into deterministic and probabilistic types (Li et al. 1997, Sundin and Braban-Lexdoux 2001, Ortiz-Garcia et al. 2006, Schram 2008, AASHTO 2012, Chen and Mastin 2015). A deterministic model provides an estimate of a performance measure that represents pavement conditions. In contrast, a probabilistic model estimates the probability that a pavement transitions from one condition level to another.

Numerous deterministic models have been investigated to develop PPMs for individual pavement segments as well as for pavement clusters.

A wide variety of deterministic models are available in the literature (Anastasopoulos and Mannering 2014). Examples include:

- Multiple regression models (Sadek et al. 1996, Hand et al. 1999, Agarwal et al. 2006, Kim and Kim 2006);
- Panel data models, i.e., random-effect and mixed-effect models (Prozzi and Madanat 2003, Archilla 2006, Lee 2007, Yu et al. 2007, Ker and Lee 2011, Khraibani et al. 2012);
- Logistic regression models (Wang 2013); and
- Clusterwise regression (CR) models (Luo and Chou 2006, Luo and Yin 2008, Zhang and Durango-Cohen 2014).

Taking into consideration the need to minimize the overall estimation errors, this study focused on the estimation of CR models.

The existing state of the art for clusterwise regression determines pavement clusters and associated PPMs simultaneously, using a single objective function. CR provides a valid statistical approach to incorporate cluster analysis into regression analysis such that the uncertainty in clustering is considered while simultaneously estimating the regression models (Luo and Chou 2006, Kang and Ghosal 2008, Hsu, 2015). CR estimates the underlying regression models and associated subpopulations (clusters) by searching a mixture of unknown number of regression models that could be formed with the available data (Kang and Ghosal 2009).

In the field of pavement management, Luo and Chou (2006) introduced CR to model pavement deterioration. A sigmoidal (S-shaped) functional form was used to relate a pavement condition rating based on pavement age. Later, Luo and Yin (2008) extended their study to model development of pavement distresses in flexible pavements. In both studies, only pavement age was used as an explanatory variable. In a recent study (Zhang and Durango-Cohen 2014), a CR model with multiple explanatory variables was proposed to estimate pavement serviceability.

A nonlinear model specification, presented by Prozzi and Madanat (2003), was used to estimate the trends in the Present Serviceability Index (PSI). A logarithmic transformation was used to linearize the adopted model. Parameters were estimated using ordinary least squares. The study used data collected during the AASHO Road Test (Highway Research Board 1962). However, the test was performed in a relatively controlled environment, and the data was collected in a single site. Clearly, the experiment characteristics were not representative of all other locations.

Previous studies using CR for pavement performance modelling suffer from other limitations. First, it was found that the explanatory power of variables used in clustering as well as regression analyses was not tested. All user-defined variables were assumed to be significant, and were included in the final models. Second, the proposed mathematical programs could not find the optimal number of clusters for the given data. Hence, time-consuming ‘trial and error’ methods were required to find the optimal number of clusters. Third, the objective function was to minimize the sum of squared errors of prediction (SSE). Given that SSE decreases monotonically as a function of the number of clusters, the optimum number of clusters with minimum SSE always is the total number of data points available in the data (Kodinariya and Makwana 2013). Therefore, minimization of SSE is not the best objective function to use for seeking an optimal number of clusters. Fourth, the PPMs were restricted to be either linear or nonlinear, irrespective of which functional form provided the best results.

To address these limitations from previous studies, this chapter proposes a mathematical programming framework within the CR approach in order to determine simultaneously 1) an optimal number of clusters, 2) the assignment of segments into clusters, and 3) the associated significant PPM parameters. The explanatory power of the variables was tested to include only significant explanatory variables in the final models.

A comprehensive solution algorithm was utilized to solve the proposed problem. The proposed approach could identify variables that cause multicollinearity issues in the models, and could address the problems, if required. The mathematical program and solution algorithm were designed to explore all possible combinations of potentially significant explanatory variables in order to select the best model specification. The Bayesian Information Criteria (BIC) (Schwarz 1978) was used as the objective function to obtain models that balance the goodness of fit and complexity in terms of number of clusters and explanatory variables. The relevance of the nonlinear functional form within the proposed framework was investigated using pavement data from the entire State of Nevada. The results were expected to illustrate the advantage of using nonlinear functional form while solving simultaneously for the three types of parameters listed above.

4.2 Methodology

4.2.1 Problem Formulation

This section provides 1) notation and definitions, 2) details about the performance measure used to evaluate pavement condition, 3) the functional form chosen to estimate PPMs, and 4) the proposed mathematical program and solution algorithm.

Notation and Definitions

The following variables were used to formulate the mathematical program:

I = Number of pavement samples in the network;

i = Subscript for a pavement sample in the network, $i \in I$;

T_i = Number of observation periods for a pavement sample i;

t = Subscript for an observation period for a pavement sample, $t \in T_i$;

O = Total number of observations = $\sum_i^I T_i \; \forall \; i \in I$;

J = Number of continuous explanatory variables;

j = Subscript for a continuous explanatory variable including an intercept, $j = 0,\ldots,J$;

H = Number of categorical explanatory variables;

h = Subscript for a categorical explanatory variable, $h \in H$;

x_{ijt}^{k} = Measurement of a continuous explanatory variable j for a sample i at observation period t that is assigned to a cluster $k \; \forall \; i \in I, j \in J, t \in T_i$;

x_{iht}^{k} = Measurement of a categorical explanatory variable h for a sample i at observation period t that is assigned to a cluster $k \; \forall \; i \in I, h \in H, t \in T_i$;

y_{it}^{k} = Measurement of dependent variable for a sample i at observation period t that is assigned to a cluster $k \; \forall \; i \in I, t \in T_i$;

K = Optimum number of clusters $(1 \leq k \leq K_{max})$;

k = Subscript for a clusters, $\forall \; k \in K$;

K_{max} = Maximum number of potential clusters that could be formed;

n = Minimum number of observations required in a cluster;

C_k = Set of pavement samples that are assigned to cluster $k \; \forall \; k \in K$;

δ = Total number of significant explanatory variables including intercepts in all clusters;

v_{jk} = Binary indicator to represent significance of a continuous explanatory variable including an intercept in a cluster $k \; \forall \; j = 0, \ldots, J, \; k \in K$;

ω_{hk} = Binary indicator to represent significance of a categorical explanatory variable in a cluster $k \; \forall \; h \in H, \; k \in K$;

p_{ik} = Cluster membership of a pavement sample i to a cluster k, $\forall \; i \in I, \; k \in K$;

β_{jk} = Estimated regression coefficient for a continuous explanatory variable j including an intercept in cluster $k \; \forall \; j = 0, \ldots, J, \; k \in K$;

ϑ_{hk}= Estimated regression coefficient for a categorical explanatory variable h in cluster $k \; \forall \; h \in H, \; k \in K$;

Pavement Performance Measure

Pavement performance is defined as an overall assessment of the serviceability pattern of a pavement (Highway Research Board 1962). Performance may be described by serviceability measurements of a pavement over the evaluation period (Li 2005, Hudson et al. 2007). Serviceability represents the degree of service that a pavement is intended to provide under existing conditions (Namakura and Michael 1963, Garcia-Diaz and Riggins 1984).

A variety of pavement performance indices are available in the literature. They were developed to evaluate various aspects of pavements, such as structural, safety, functional, skid resistance, and surface distress (Garcia-Diaz and Riggins 1984, Hand et al. 1999, Li 2005). Some indices were proposed to evaluate an individual aspect of pavement, whereas others were developed to characterize a combination of aspects (Zhang et al. 1993). For example, the International Roughness Index provides the riding quality of a pavement surface. Similarly, structural capacity may be evaluated using a structural number.

The Present Serviceability Index is a commonly used functional performance measure in pavement performance modelling (Hand et al. 1999). PSI serves as a unified standard to measure the riding comfort from the driver's point of view (Shoukry et al. 1997, Garcia-Diaz and Riggins 1984, Terzi 2006, Attoh-Okine and Adarkwa 2013). In addition, PSI is easily understood by road users and legislators (Hudson et al. 2015). In this study, the PSI was used as the dependent variable.

Model Functional Form

Identification of the potential functional form is the most important step when formulating a modelling approach (Darter 1980, Sadek et al. 1996). The selected functional form must represent the actual physical phenomenon – in this study, the deterioration trend – and provide the best fit with the given data. In addition, the selected function form must satisfy all boundary conditions (Wolters and Zimmerman 2010). Both simple and complex functional forms have been utilized to develop PPMs (Sadek et al. 1996, Li et al. 1997, de Melo Silva et al. 2000, Luo and Yin 2008). The simple linear model forms have a few inherent characteristics that typically restrict them in achieving a high level of accuracy for a variety of conditions. Hence, flexible model forms that typically are nonlinear are preferred (Shekharan 2000).

The literature revealed that nonlinear models typically are more appropriate when representing pavement deterioration over time (Hong and Prozzi 2006). Various functional forms have been proposed and implemented for pavement performance modelling. Nonlinear functional forms – including exponential, sigmoidal, and polynomial types – were used by many state departments of transportations (DOTs). For example, sigmoidal models were developed by Texas (TXDOT) and North Carolina (NCDOT) to predict pavement distresses, and by Minnesota (MnDOT) to predict the ride qualify index (Stampley et al. 1995, Gharaibeh et al. 2012, Wolters and Zimmerman 2010, Chen et al. 2014). The City of Cincinnati adopted power and exponential performance models to predict pavement conditions (Rajagopal 2006). Similarly, power functional forms were used by the departments of transportation for Washington State (WSDOT), Oklahoma (ODOT), and Louisiana (LDOTD) to develop PPMs as a function of pavement age (Kay et al. 1993, Khatta et al. 2008, Wolters and Zimmerman 2010).

The property of the power functional form is suitable for describing historical trends of pavement deterioration (Chan et al. 1997). This study investigates the appropriateness of using a power functional form in the context of CR to estimate PPMs. The functional form proposed for the PPMs can be expressed as:

$$y_{it}^{k}=\beta_{0k} * \prod_{j=1}^{J}\left(x_{ijt}^{k}\right)^{\beta_{jk}} * \prod_{h=1}^{H} exp\left(\vartheta_{hk} * x_{iht}^{k}\right) \qquad (4.1)$$

The model parameters can be estimated in two ways. In order to utilize ordinary least squares (OLS), the model can be linearized by performing logarithmic transformation. Alternatively, model parameters are estimated by using nonlinear regression techniques. Considering that direct estimation of nonlinear regression models requires a high amount of computational time, this study pursued the estimation by using logarithmic transformation. The linearized functional form used in the proposed CR analysis was expressed as:

$$ln\left(y_{it}^{k}\right)=ln\left(\beta_{0k}\right)+ \sum_{j=1}^{J}\beta_{jk}*ln\left(x_{ijt}^{k}\right) +\sum_{h=1}^{H}\left(\vartheta_{hk}*x_{iht}^{k}\right) \qquad (4.2)$$

Mathematical Program

The objective function involves minimization of BIC, expressed as:

$$Min.\ BIC = O+O*ln(2\pi)+O*ln\left(\frac{SSE}{O}\right)+(\delta+K\text{-}1)*ln(O) \qquad (4.3)$$

where, SSE is total sum of squared errors, expressed as:

$$SSE = \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{t=1}^{T_i} \left[ln\left(\beta_{0k}\right) + \sum_{j=1}^{J} \beta_{jk} * ln\left(x_{ijt}^{k}\right) + \sum_{h=1}^{H} \left(\vartheta_{hk} * x_{iht}^{k}\right) - ln\left(y_{it}^{k}\right)\right]^2 * p_{ik} \qquad (4.4)$$
$$\forall\ i \in I,\ t \in T_i,\ j \in J,\ h \in H,\ k \in K$$

and the quantity (δ + K -1) is the total number of free parameters to be estimated for K clusterwise regression models (DeSarbo and Corn 1988). Decision variables to be determined included the optimum number of clusters, K; coefficients for cluster-specific significant explanatory variables, β_{0k}, β_{jk}, and ϑ_{hk}; and cluster memberships, p_{ik}.

The proposed mathematical programming included the following constraints:

$$\delta = \sum_k \left(\sum_j v_{jk} + \sum_h \omega_{hk}\right) \forall\ j = 0, \ldots, J,\ h \in H,\ k \in K \qquad (4.5)$$

$$v_{jk} = \begin{cases} 1, & if\ \beta_{jk}\ is\ significant; \\ 0, & Otherwise \end{cases} \quad \forall\ j = 0, \ldots, J,\ k \in K \qquad (4.6)$$

$$\omega_{hk} = \begin{cases} 1, & if\ \vartheta_{hk}\ is\ significant; \\ 0, & Otherwise \end{cases} \quad \forall\ h \in H,\ k \in K \qquad (4.7)$$

$$\sum_k p_{ik} = 1\ \forall i \in I,\ k \in K \qquad (4.8)$$

$$p_{ik} = \begin{cases} 1, & if\ \ sample\ i\ is\ assigned\ to\ cluster\ k; \\ 0, & Otherwise \end{cases} \quad \forall\ i \in I,\ k \in K \qquad (4.9)$$

$$C_k = \left\{ i \middle| p_{ik} = 1 \forall i \in I,\ k \in K \right\} \qquad (4.10)$$

$$C_{k'} \cap C_{k''} = \text{null}\ \ \forall k' \neq k'',\ k'\ \text{and}\ k'' \in K \qquad (4.11)$$

$$\cup_{k \in K} |C_k| = I \qquad (4.12)$$

$$\sum_{i \in C_k} T_i \geq n \; \forall \, C_k \tag{4.13}$$

$$1 \leq k \leq K_{max} \tag{4.14}$$

$$K_{max} = F(I,\ T_i,\ n) \tag{4.15}$$

Constraint (4.5) provided the total number of significant explanatory variables, including intercepts in all clusters. In Constraint (4.6), v_{jk} equaled 1 if coefficient β_{jk} was significant; otherwise, v_{jk} equaled 0. Similarly, in Constraint (4.7), ω_{hk} equaled 1 if coefficient ϑ_{hk}, was significant; otherwise, ω_{hk} equaled 0.

Significance was determined using p-value and α. Constraints (4.8 and 4.9) were used to assign cluster memberships to the samples. The indicator p_{ik} equaled 1 if and only if sample i was in cluster k; otherwise p_{ik} equaled 0. The sets of samples assigned to K clusters were provided by Constraint (4.10). Assignment of a sample to multiple clusters was restricted by Constraint (4.11). Constraint (4.12) ensured all available samples were assigned to clusters. The minimum number of observations required for statistical significance was defined by Constraint (4.13). Constraint (4.14) stated a range of feasible clusters for the available data.

The maximum number of clusters was a function (F) of I, T_i, and n. Table 4.1 provides the step-by-step procedure to calculate this number.

Table 4.1 Function F to Calculate K_{max}

```
**Inputs:** $I$, $T_i$, and $n$
**Output:** $K_{max}$
If (Total observations < $n$) then
    $K_{max} = 0$
else
    Create a matrix, M of size ($\tau_{max}$ x 2):
        • $\tau_{max}$ is maximum number of observations of a pavement sample in the data set
        • $m_{\tau,1}$ includes with all integers from 1 to $\tau_{max}$ in an ascending order
        • $m_{\tau,2}$ includes number of samples that have $\tau$ observations
    If ($m_{\tau,1} \geq n$) then
        $K_{max} = \sum_{\tau \geq n} m_{\tau,2}$
        Update $m_{\tau,2}$ with 0 for all $\tau \geq n$
    else
        Repeat
        If $\sum_{\tau} m_{\tau,2} = 0$ then
            Update $K_{max}$
        else
            Remove rows with $m_{\tau,2} = 0$ from M
            Initialize counters: $\psi = \gamma$ = number of rows in M
            M’ = M
            If $\sum_{\tau}(m_{\tau,1} * m_{\tau,2}) < n$ then
                 Update $K_{max}$
            else
                $S = m_{\psi,1}$; $m_{\psi,2} = m_{\psi,2} - 1$
                Repeat
                If ($m_{\gamma,2} = 0$) then
                    $\gamma = \gamma$ -1
                        If ($\gamma$ = 0) then M = M’, $\gamma$ = number of rows in M; $n = n + 1$; $S = 0$ end
                else
                    If ($S > n$) then $S = S - m_{\gamma,1}$; $m_{\gamma,2} = m_{\gamma,2} + 1$; $\gamma = \gamma$ -1 end
                        If ($\beta$ = 0) then M = M’, $\gamma$ = number of rows in M; $\psi = \psi - 1$; $S = 0$ end
                            If ($\psi$ = 0) then
                                $\psi = \gamma$ = number of rows in M, $n = n$+1ψ
                            else
                                $S = m_{\psi,1}$; $m_{\psi,2} = m_{\psi,2} - 1$
                            end
                    $S = S + m_{\gamma,1}$; $m_{\gamma,2} = m_{\gamma,2} - 1$
                end
                Until $S = n$
                $K_{max} = K_{max} + 1$
            end
        end
        Until no sample is available for clustering
    end
Update $K_{max}$
end
```

4.2.2 Solution to the Mathematical Program

This section provides a description of the solution algorithm utilized to find the optimal solution for the proposed mathematical programming problem. The exiting literature does not provide an exact algorithm to solve such a combinatorial problem efficiently (Meneses and Ferreira 2012). In addition, no single approach is known to be superior to other methods (Marler and Arora 2004). Hence, the selection of an appropriate solution approach is problem-specific and depends on user preferences, such as the availability of software and trade-offs between computational time and the quality of the results.

In this study, Simulated Annealing (SA) integrated with Function All Subsets Regression (ASR) was utilized to solve the proposed mathematical problem. The ASR sought for the best model parameters that could provide a balance between goodness of fit and model complexity. The criteria used to select the best model were BIC and α. All potential model specifications were tested. In addition, the ASR took care of potential multicollinearity that might be present in the models. Table 4.2 provides an algorithmic description of the function ASR.

Table 4.2 Function All Subsets Regression

Inputs: K, cluster memberships, observations of all explanatory variables and a dependent variable
Outputs: Models parameters and set of significant explanatory variables

```
1. Set $k$ =1
2. Repeat
      2.1. Calculate GVIFs for all explanatory variables used in the model
      2.2. Remove the explanatory variable with the largest GVIF and recalculate GVIFs with the
           remaining variables
   Until all explanatory variables have GVIF less than $VIF_{max}$. Let $\hat{J}$ is number of such variables
3. Generate all possible subsets of $\hat{J}$
4. For all subsets, estimate model parameters and BIC using ordinary least squares method
5. Select the model that has minimum BIC and all variables with p-value < $\alpha$
6. If $k < K$, go to Step 7; otherwise, Step 8
7. Set $k = k + 1$ and go to Step 2
8. Return the model parameters and associated significant explanatory variables of the best models
   selected for all $K$ clusters in Step 5
9. End
```

The SA portion of the solution algorithm determined the cluster memberships of pavement samples and called the ASR function to estimate the optimal solution. SA was chosen because it:

1) Can escape from local optima (occasionally) by accepting moves that degrades solutions,
2) Is a simple and efficient search algorithm to solve combinatorial optimization problems, and
3) Is easy to implement.

SA has been used successfully to solve similar problems (DeSarbo *et al.* 1989, Selim and Alsultan 1991, Sun *et al.* 1994). Table 4.3 provides the master algorithm utilized to solve the proposed mathematical problem.

4.3 Numerical Experiment and Results

4.3.1 Experimental Research Data

Experimental research data were extracted from the Pavement Management System database of the Nevada Department of Transportation. The data included pavement conditions and roadway inventory data that was collected over a 12-year period (from 2001 to 2012, inclusive) for the entire State of Nevada. A total of 17,642 observations of flexible pavements were available for the experiments. Out of this, 14,637 observations (2001 to 2010) were used to develop PPMs and 3,005 observations (2011 and 2012) were used for validation.

Table 4.4 includes variables used in this study. PSI is used as the dependent variable. The descriptions of the data were provided in Chapter 2.

Table 4.3 Master Algorithm: Simulated Annealing Integrated with All Subsets Regression

```
Inputs: Observations of all explanatory variables and a dependent variable for pavement samples
Outputs: $K_{optimal}$, models parameters, cluster memberships, and clusters-specific significant explanatory variables

1. Set $K$ =2, initial temperature = $\theta_0$, and final temperature = $\theta_{min}$, $BICmin = \infty$
2. Call Function F to calculate $K_{max}$
3. Repeat
    3.1. Randomly generate a valid $K$ initial clusters of pavement samples, $C'_k$
    3.2. Call Function ASR to estimate the best model parameters
    3.3. Evaluate the objective function, BIC ($C'_k$) using Equation 4.3
    3.4. Set current temperature, $\theta = \theta_0$
    3.5. Repeat the following steps for $N_{max}$ times
        a. Randomly generate valid $K$ neighborhood clusters, $C''_k$
        b. Call Function ASR to estimate the best model parameters
        c. Evaluate the objective function, BIC ($C''_k$) using Equation 4.3
        d. Calculate ΔBIC = BIC ($C''_k$) - BIC ($C'_k$)
        e. If ΔBIC < 0, let $BIC_K$ = BIC ($C''_k$) and $C'_k = C''_k$, and go to Step 3.6. Otherwise, do the following:
            • Generate a random number $u \sim U(0,1)$. Calculate acceptance probability, $p_{accept} = \exp\left(\frac{-\Delta BIC}{B * \theta}\right)$, where $B$ is a Boltzmann's constant
            • If $p_{accept} > u$, let $BIC_K$ = BIC ($C''_k$) and $C'_k = C''_k$, and go to Step 3.6. Otherwise, go back to Step 3.5
    3.6. If $\theta < \theta_{min}$ then
            • Update $\theta = \lambda * \theta$, where $\lambda$ is the cooling rate
            • Go back to Step 3.5
        else
            • Go to Step 3.7
        end
    3.7. If $BIC_K < BIC_{min}$ then
            • Update $BIC_{min} = BIC_K$ and $K_{optimal} = K$
        else
            • If $K < K_{max}$ then set $K = K+1$; otherwise go to Step 4
        end
4. Return $K_{optimal}$, cluster, $C'_k$, models parameters, and set of clusters-specific significant explanatory variables
5. End
```

Table 4.4 Variables Used in the Pavement Performance Models

Variable	Description
age	Age of the last M&R treatment performed on a segment
adt	One direction average daily traffic
trucks	One direction average daily trucks
elevation	Elevation at midpoint of a segment (m)
precip	Average annual precipitation (cm/year)
min_temp	Minimum average yearly air temperature (^{0}C)
max_temp	Maximum average yearly air temperature (^{0}C)
wet_days	Total number of wet days (days that moisture was recorded) over the course of one year
freeze_thaw	Total number of freeze-thaw cycles that a pavement experienced over the course of one year
rut_depth	Average ride rut depth (cm)
lane=2	Dummy variable for a segment that has 2 lanes (1 = yes, 0 = no)
lane≥3	Dummy variable for a segment that has 3 or more lanes (1 = yes, 0 = no)
sys_id=2	Dummy variable for a segment that is part of NHS (1 = yes, 0 = no)
sys_id=3	Dummy variable for a segment that is part of STP (1 = yes, 0 = no)
f_class=2	Dummy variable for a segment classified as functional class 2 (1=yes, 0 = no)
f_class=3	Dummy variable for a segment classified as functional class 3 (1 = yes, 0 = no)
f_class=4	Dummy variable for a segment classified as functional class 4 (1 = yes, 0 = no)
f_class=5	Dummy variable for a segment classified as functional class 5 (1 = yes, 0 = no)
f_class=6	Dummy variable for a segment classified as functional class 6 (1 = yes, 0 = no)
f_class=7	Dummy variable for a segment classified as functional class 7 (1 = yes, 0 = no)
category=2	Dummy variable for a segment grouped in prioritization category 2 (1 = yes, 0 = no)
category=3	Dummy variable for a segment grouped in prioritization category 3 (1 = yes, 0 = no)
category=4	Dummy variable for a segment grouped in prioritization category 4 (1 = yes, 0 = no)
category=5	Dummy variable for a segment grouped in prioritization category 5 (1 = yes, 0 = no)

4.3.2 Estimation Parameters

Several estimation parameters were required to initiate and utilize the proposed solution algorithm. It was very important to select appropriate starting values of the parameters, because the convergence rate of the algorithm largely depended on them. Experience from previous research (Paz et al. 2015a and b, Khadka and Paz 2017, Paz and Khadka 2017) and results from the sensitivity analysis were used to choose the estimation parameter values to be used in the experiments. Table 4.5 provides the estimation parameters specified in the experiments.

Table 4.5 Estimation Parameters Used in the Experiments

Parameter	Value	Remarks
θ_0	10	Initial temperature
θ_{min}	10e-17	Final minimum temperature
B	80	Boltzmann constant
λ	0.97	Cooling rate
N_{max}	5	Number of neighborhood solutions generated at each temperature level
n	800	Minimum number of observations required in a cluster
N_{ps}	80	Number of pavement samples, which memberships were changed to generate a neighborhood cluster
VIF_{max}	5	Limiting VIF
α	5%	Level of Significance

4.3.3 Results and Discussion

Function F determined the maximum number of feasible clusters for the data used in this study to be 16. The solution algorithm explored all the feasible number of clusters (i.e., K = 2 to 16) to seek for the optimum number of clusters. Figure 4.1a shows the optimum values of the objective function for each of the feasible number of clusters. The algorithm returned five-cluster models with the lowest BIC as a part of the optimum solution. Figure 4.1b visualizes the convergence curve of BIC for the five-cluster models. Initially, BIC was -26,955, and after 1,437 iterations, the algorithm found the optimal solution to have a BIC of -30,010.

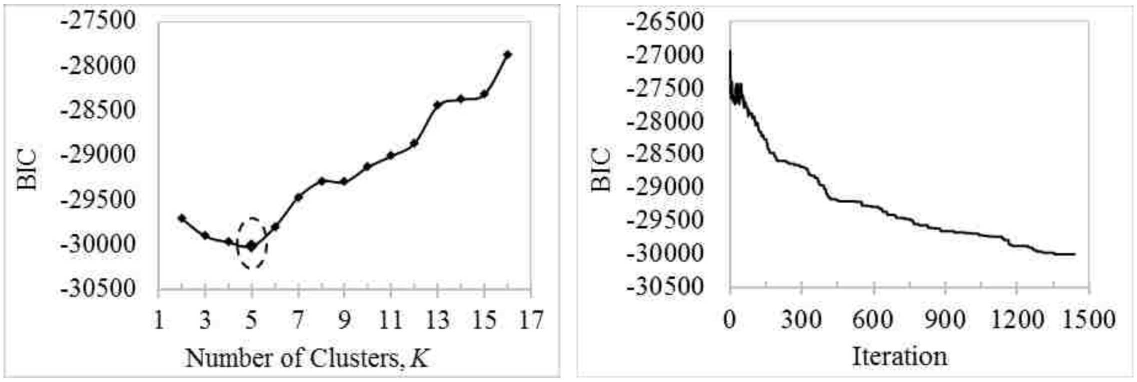

Figure 4.1 BIC versus the number of clusters (a), and convergence curve of BIC for five-cluster models (b).

The estimated parameters for the five-cluster models are presented in Table 4.6. Explanatory variables identified and included in the models were *ln*(*age*), *ln*(*adt*), *ln*(*rut_depth*), *ln*(*precip*), *ln*(*min_temp*) as well as all the dummy variables for the number of lanes, prioritization category, and functional class. In the pavement modelling literature, these variables were considered to be critical factors that affect pavement performance (Saraf and Majidzadeh 1992, Prozzi and Madanat 2004, Salama et al. 2006). Only two variables, *ln*(*age*) and *ln*(*rut_depth*), were included in all five models. The variable *ln*(*precip*) was significant only in the model for Cluster #3. Other variables were common in a few models. For example, *ln*(*adt*) was included in models for Clusters #1, #2, and #4.

Other explanatory variables were excluded from the resultant models because they were either causing multicollinearity in the models or were statistically insignificant ($\alpha = 0.05$). The variance-inflation factor (VIF) was used as a criterion to investigate potential multicollinearity in the models. Because the models included a few categorical explanatory variables that had more than one level, generalized variance-inflation factors (GVIF) were calculated, as suggested by Fox and Monette (1992). For each model, the explanatory variables with GVIFs greater than the limiting VIF were dropped, one at a time, starting with the one that had the largest GVIF. Table 4.6 shows the GVIFs of significant explanatory variables that were included in the final models. The GVIF values were less than the limiting VIF, which indicated that the models were free from serious multicollinearity.

Table 4.6 Estimated Parameters of Five-Cluster Nonlinear Models (α = 0.05)

Clusters, *k*	Variables, *j*	Coefficients, β_{jk}	p-value	Bootstrap 95% CI	GVIF
1	*Intercept*	1.464	<0.000	(1.438, 1.490)	-
(2,818)	*ln*(*age*)	-0.021	<0.000	(-0.026, -0.017)	1.01
	ln(*adt*)	-0.007	<0.000	(-0.011, -0.003)	1.89
	ln(*rut_depth*)	-0.104	<0.000	(-0.134, -0.074)	1.07
	lane=2	-0.067	<0.000	(-0.080, -0.053)	1.54
	lane≥3	-0.155	<0.000	(-0.178, -0.132)	
	f_class=2	-0.034	0.041	(-0.067, -0.001)	1.13
	f_class=3	-0.049	<0.000	(-0.061, -0.038)	
	f_class=4	-0.084	<0.000	(-0.097, -0.071)	
	f_class=5	-0.149	<0.000	(-0.164, -0.134)	
	f_class=6	-0.305	<0.000	(-0.337, -0.274)	
	f_class=7	-0.359	<0.000	(-0.467, -0.251)	
2	*Intercept*	1.568	<0.000	(1.540, 1.596)	-
(2,968)	*ln*(*age*)	-0.027	<0.000	(-0.032, -0.022)	1.02
	ln(*adt*)	-0.007	0.001	(-0.010, -0.003)	2.06
	ln(*rut_depth*)	-0.060	<0.000	(-0.090, -0.03)	1.08
	lane=2	-0.052	<0.000	(-0.067, -0.038)	1.58
	lane≥3	-0.114	<0.000	(-0.137, -0.091)	
	f_class=2	-0.040	0.015	(-0.072, -0.009)	1.16
	f_class=3	-0.056	<0.000	(-0.068, -0.044)	
	f_class=4	-0.128	<0.000	(-0.142, -0.114)	
	f_class=5	-0.393	<0.000	(-0.411, -0.375)	
	f_class=6	-0.465	<0.000	(-0.498, -0.432)	
	f_class=7	-0.490	<0.000	(-0.579, -0.401)	
3	*Intercept*	1.922	<0.000	(1.776, 2.067)	-
(2,962)	*ln*(*age*)	-0.021	<0.000	(-0.026, -0.016)	1.01
	ln(*precip*)	-0.027	<0.000	(-0.042, -0.013)	1.95
	ln(*min_temp*)	-0.116	<0.000	(-0.15, -0.082)	1.91
	ln(*rut_depth*)	-0.149	<0.000	(-0.177, -0.121)	1.02
	f_class=2	-0.046	0.004	(-0.076, -0.015)	1.02
	f_class=3	0.010	0.020	(0.002, 0.018)	
	f_class=4	-0.051	<0.000	(-0.060, -0.042)	
	f_class=5	-0.166	<0.000	(-0.176, -0.155)	
	f_class=6	-0.257	<0.000	(-0.282, -0.232)	
	f_class=7	-0.051	0.010	(-0.090, -0.012)	
4	*Intercept*	1.536	<0.000	(1.466, 1.607)	-
(2,942)	*ln*(*age*)	-0.010	<0.000	(-0.015, -0.005)	1.01
	ln(*adt*)	-0.046	<0.000	(-0.050, -0.043)	1.60
	ln(*min_temp*)	-0.121	<0.000	(-0.142, -0.101)	1.17
	ln(*rut_depth*)	-0.210	<0.000	(-0.24, -0.179)	1.02
	lane=2	-0.037	<0.000	(-0.047, -0.027)	1.25
	lane≥3	-0.152	<0.000	(-0.173, -0.131)	
5	*Intercept*	1.876	<0.000	(1.803, 1.949)	-
(2,948)	*ln*(*age*)	-0.021	<0.000	(-0.026, -0.016)	1.01
	ln(*min_temp*)	-0.091	<0.000	(-0.111, -0.071)	1.14
	ln(*rut_depth*)	-0.170	<0.000	(-0.202, -0.138)	1.05
	lane=2	-0.079	<0.000	(-0.092, -0.065)	1.49
	lane≥3	-0.137	<0.000	(-0.156, -0.118)	
	category=2	-0.072	<0.000	(-0.085, -0.059)	1.21
	category=3	-0.102	<0.000	(-0.117, -0.088)	
	category=4	-0.181	<0.000	(-0.198, -0.164)	
	category=5	-0.254	<0.000	(-0.271, -0.238)	

Note: number in parenthesis represents the number of observations in a clusters

As the model parameters were estimated after logarithmic transformation was performed, the normality of the residuals was investigated. The Anderson-Darling test was performed using the R package, ‘nortest’ (R Core Team 2015). Table 4.7 includes the results of the test, which indicated that the residuals were non-normal and the estimated p-values of the regression coefficients were inaccurate. Hence, bootstrap (Efron 1979, Efron and Tibshirani 1993) was used to calculate 95% confidence intervals (CI) of all the estimated coefficients. Table 4.6 includes the lower and upper bounds of the 95% CIs for all coefficients. Table 4.6 shows that the bootstrap results confirmed the significance of all explanatory variables in the model. For example, the coefficient for *f_class = 3* in the case of Cluster #3 is 0.01 with a p-value of 0.02, and the bootstrap CI (0.002, 0.018) fell to the right of 0; that is, *f_class = 3* was positive and significant.

Table 4.7 Anderson-Darling Normality Test (Nonlinear Models)

Clusters	A-value	p-value
1	92.876	< 2.2e-16
2	113.85	< 2.2e-16
3	102.96	< 2.2e-16
4	75.497	< 2.2e-16
5	58.248	< 2.2e-16

It was observed that models for Cluster #1 and #2 included the same significant explanatory variables. In addition, the coefficients of the corresponding variables were nearly equal and had the same sign. These similarities suggested that these two clusters could be merged, but the objective function of the resultant model would increase slightly. This can be observed in Figure 4.1b, where the BIC for number of clusters equal to 4 was slightly higher than for 5.

The models had different cluster-specific explanatory variables, that is, each model had a unique set of explanatory variables. In addition, the associated coefficients varied among the resultant models. These differences indicated that average performance behavior of pavement samples across clusters were different.

4.3.4 Model Performance

The proposed algorithms discussed in the Solution Algorithm section were also utilized to estimate the optimal number of clusters and associated model parameters, using a linear functional form expressed by:

$$y_{it}^{k} = \beta_{0k} + \sum_{j=1}^{J} \beta_{jk} * x_{ijt}^{k} + \sum_{h=1}^{H} \vartheta_{hk} * x_{iht}^{k} \qquad (4.16)$$

Six-cluster linear models were found to be part of the optimum solution. The estimated model parameters are provided in Table 4.8. The primary observation was that the variables *min_temp* and *precip*, which were significant in the nonlinear models, were not included in any of the linear models. As expected, the intercepts and coefficients of the corresponding explanatory variables were different across the models.

The performances of the resultant nonlinear and liner models were compared. Both models were applied to the test data set to estimate PSIs for 2011 and 2012. For nonlinear models, the predicted PSIs were transformed back to the original scale by taking exponential.

The scatter plots of observed versus predicted PSIs for both models are shown in Figure 4.2. The figure shows that the nonlinear models had less scattered data points beyond the ±15% error lines. About 81% of the total data points are within the ±15% range of error. In case of linear models, approximately 74% of the total data points are within the ±15% error lines.

Table 4.8 Estimated Parameters of Six-Cluster Linear Models (α = 0.05)

Variables, *j*	Clusters, *k*	Coefficients, β_{jk}	p-value	Bootstrap 95% CI	GVIF	Clusters, *k*	Coefficients, β_{jk}	p-value	Bootstrap 95% CI	GVIF
Intercept	1	4.393	<0.000	(4.362, 4.423)	-	6	4.401	<0.000	(4.374, 4.428)	-
age	(2,376)	-0.040	<0.000	(-0.045, -0.034)	1.0	(2,583)	-0.037	<0.000	(-0.042, -0.031)	1.0
$adt^{\dagger}$		-0.013	<0.000	(-0.015, -0.012)	1.1		-0.014	<0.000	(-0.015, -0.012)	1.2
rut_depth		-0.200	<0.000	(-0.285, -0.115)	1.0		-0.355	<0.000	(-0.432, -0.278)	1.0
f_class=2		-0.186	0.002	(-0.303, -0.069)	1.0		0.468	<0.000	(0.350, 0.586)	1.0
f_class=3		-0.111	<0.000	(-0.140, -0.081)			-0.086	<0.000	(-0.113, -0.060)	
f_class=4		-0.259	<0.000	(-0.293, -0.226)			-0.258	<0.000	(-0.289, -0.228)	
f_class=5		-1.052	<0.000	(-1.092, -1.012)			-0.864	<0.000	(-0.905, -0.823)	
f_class=6		-1.182	<0.000	(-1.265, -1.098)			-1.288	<0.000	(-1.365, -1.211)	
f_class=7		-0.284	0.006	(-0.492, -0.076)			-0.634	<0.000	(-0.811, -0.457)	
Intercept	2	4.552	<0.000	(4.496, 4.608)	-	4	4.605	<0.000	(4.548, 4.661)	-
age	(2,483)	-0.022	<0.000	(-0.028, -0.016)	1.0	(2,414)	-0.033	<0.000	(-0.038, -0.027)	1.0
$adt^{\dagger}$		-0.012	<0.000	(-0.014, -0.01)	1.3		-0.006	<0.000	(-0.008, -0.005)	1.3
rut_depth		-0.436	<0.000	(-0.527, -0.346)	1.0		-0.574	<0.000	(-0.667, -0.482)	1.1
lane=2		-0.191	<0.000	(-0.24, -0.141)	1.6		-0.213	<0.000	(-0.266, -0.160)	1.6
lane≥3		-0.202	<0.000	(-0.294, -0.111)			-0.405	<0.000	(-0.484, -0.326)	
category=2		-0.202	<0.000	(-0.248, -0.156)	1.2		-0.263	<0.000	(-0.311, -0.215)	1.2
category=3		-0.323	<0.000	(-0.38, -0.266)			-0.326	<0.000	(-0.383, -0.268)	
category=4		-0.664	<0.000	(-0.73, -0.599)			-0.650	<0.000	(-0.716, -0.584)	
category=5		-1.149	<0.000	(-1.216, -1.083)			-0.808	<0.000	(-0.874, -0.742)	
Intercept	3	4.674	<0.000	(4.612, 4.736)	-	5	4.557	<0.000	(4.496, 4.618)	-
age	(2,442)	-0.028	<0.000	(-0.034, -0.022)	1.0	(2,340)	-0.028	<0.000	(-0.034, -0.021)	1.0
$adt^{\dagger}$		-0.008	<0.000	(-0.010, -0.006)	1.5		-0.005	<0.000	(-0.006, -0.003)	1.5
rut_depth		-0.517	<0.000	(-0.618, -0.417)	1.0		-0.510	<0.000	(-0.607, -0.413)	1.1
lane=2		-0.358	<0.000	(-0.414, -0.302)	1.7		-0.260	<0.000	(-0.317, -0.203)	1.7
lane≥3		-0.289	<0.000	(-0.391, -0.187)			-0.294	<0.000	(-0.394, -0.195)	
category=2		-0.325	<0.000	(-0.376, -0.273)	1.2		-0.194	<0.000	(-0.247, -0.141)	1.2
category=3		-0.465	<0.000	(-0.529, -0.401)			-0.287	<0.000	(-0.348, -0.226)	
category=4		-0.684	<0.000	(-0.756, -0.613)			-0.639	<0.000	(-0.709, -0.569)	
category=5		-0.808	<0.000	(-0.881, -0.735)			-1.130	<0.000	(-1.203, -1.057)	

Note: † = variable value in thousands, - = Not applicable, and quantity in parenthesis represents the number of observations in a clusters

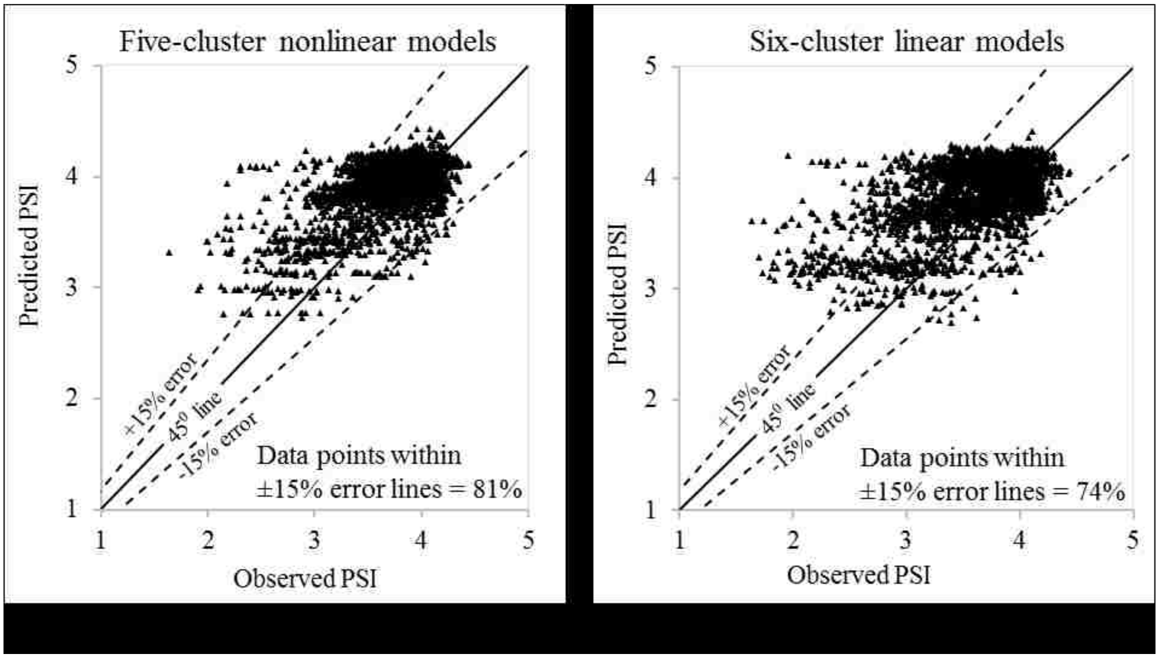

Figure 4.2 Observed versus prediction PSI: a) linear models, and b) nonlinear models

The prediction accuracy of the models was measured using root-mean-square error (RMSE), normalized root-mean-square error (NRMSE), and mean absolute errors (MAE). The overall values for RMSE, NRMSE, and MAE for all the nonlinear models were 0.41, 0.15, and 0.33; whereas for the linear models were 0.47, 0.17, and 0.36, respectively. All three metrics for the nonlinear models were less than those for the linear models. This indicated that the nonlinear models were more accurate than the linear models in estimating PSIs of pavement samples. Table 4.9 provides the RMSE, NRMSE, and MAE values for all the models as well as for individual nonlinear and linear models.

Table 4.9 Overall and Individual RMSE, NRMSE, and MAE

Clusters	Five-cluster nonlinear models				Six-cluster linear models			
	No. of Obs.	RMSE	NRMSE	MAE	No. of Obs.	RMSE	NRMSE	MAE
1	567	0.41	0.19	0.32	474	0.47	0.18	0.37
2	560	0.39	0.16	0.31	525	0.46	0.18	0.37
3	610	0.42	0.17	0.34	507	0.49	0.18	0.37
4	621	0.41	0.18	0.33	495	0.47	0.17	0.35
5	647	0.42	0.16	0.33	494	0.48	0.18	0.36
6	-	-	-	-	510	0.49	0.19	0.38
Overall	3005	0.41	0.15	0.33	3005	0.47	0.17	0.36

4.4 Conclusions

This chapter discussed some of the modelling approaches that were used in the pavement performance literature. Clusterwise regression (CR) was introduced as the existing state-of-the-art approach to estimate PPMs. In addition, the chapter outlined some of the limitations of existing state-of-the-art approach. To address these limitations, a comprehensive mathematical programming approach and solution algorithm were proposed.

The proposed mathematical programming approach used a single objective function to simultaneously divide the pavement samples into an optimum number of clusters and estimate the corresponding model parameters. Bayesian Information Criteria was used as the objective function, and Simulated Annealing integrated with All Subsets Regression was utilized to solve the mathematical problem. During the optimization process, the algorithm sought for the potential explanatory variables that posed serious multicollinearity issues in a model. The variables that posed the largest multicollinearity effect in the models were dropped individually until all models were free from serious multicollinearity issues. All possible linear and nonlinear model specifications were examined to determine the best model for each cluster. A power functional form was used to estimate nonlinear PPMs.

Five-cluster models were found as the optimum solution. Variables *ln*(*age*), *ln*(*adt*), *ln*(*rut_depth*), *ln*(*precip*), *ln*(*min_temp*) as well as all the dummy variables for the number of lanes, prioritization category, and functional class were identified as the significant and as having realistic coefficients. All these variables were considered to be critical factors affecting pavement performance. Linear PPMs were estimated using the same algorithmic framework discussed in this chapter. The variables *min_temp* and *precip*, which were significant in the nonlinear models, were not included in any of the linear models. Similar to nonlinear models, the estimated coefficients and associated sign were as expected, and were realistic.

The performances of the nonlinear and linear models were compared by means of validation of the prediction capabilities. In addition, RMSE, NRMSE, and MAE were used to compare the explanatory power of the models even further. Results showed that the nonlinear models were more accurate than the linear models in estimating PSIs. However, the nonlinear models overestimated the PSIs for a few pavement samples.

CHAPTER 5

CONCLUSIONS AND FUTURE RESEARCH

This chapter summarizes this dissertation and highlights its significance. Recommendations for possible extensions and directions for future research are also discussed in this chapter.

5.1 Summary and Conclusions

This study proposes a generalized Clusterwise Regression (CR) approach to estimate pavement clusters and associated PPMs, simultaneously. The proposed approach integrated clustering, variable selection, and regression techniques to seek for the true underlying pavement clusters and the best model specification for PPMs. The resultant PPMs included cluster-specific significant explanatory variables. That is, significant explanatory variables could be different across models.

A mixed-integer nonlinear mathematical program with BIC as the objective function was formulated to describe the problem. The program was flexible enough to handle multiple explanatory variables, multiple observations per pavement segments, and user-defined constraints on cluster characteristics. In addition, the program assigned all observations of a pavement sample to the same cluster exclusively. An iterative search based optimization procedure was implemented to explore all feasible clusters that could be formed for a given data set. All possible combinations of explanatory variables were explored and potential multicollinearity issues were addressed. A comprehensive algorithm was implemented in the software R to solve the proposed mathematical problem. The algorithm included Simulated Annealing coupled with: (i) Ordinary Least Squares (OLS) for estimation of the linear models, and (ii) All Subset Regression for estimation of the nonlinear models. Parameters of the nonlinear models were estimated after linearizing the adopted model using a logarithmic transformation. The estimated model parameters were then transformed back to the original scale by taking exponential. The optimization parameters required by the solution algorithm were determined based on previous experience and extensive sensitivity analysis.

The study results highlighted the ability of variable selection procedure to distinguish between significant and insignificant explanatory variables. They also illustrated the importance of testing the significance of explanatory variables while seeking for the best model specification. The results showed average daily traffic, pavement age, rut-depth along the pavement, average annual precipitation and minimum temperature, function class, prioritization category, and the number of lanes, as significant for explaining pavement performance. Further, the estimated model parameters were as expected in magnitude and signs. The models were accurate in estimating pavement performance or condition with minimal errors. No overfitting issues were observed. These results together implied that the proposed approach was effective in developing adequate PPMs. A detailed analysis of the experiment results suggested that nonlinear CR models were superior than the linear CR models in terms of prediction accuracy for the data used in this study.

5.2 Research Contributions

The primary contribution of this dissertation is a comprehensive framework including a mathematical programming formulation and solution algorithm to determine simultaneously the optimal number of clusters, sample memberships to clusters, cluster-specific significant explanatory variables and associated coefficients, and the best functional form between linear and power models for pavement performance. The existing literature does not provide this kind of generalized modeling approach which seeks balance between goodness of fit and model complexity and facilitates superior model development.

5.3 Future Research

This study investigated the appropriateness of the power functional form within the clusterwise regression framework to estimate PPMs. Various other functional forms have been used to explain pavement performance behavior. It would be worthwhile expanding the proposed framework to consider all potential functional forms in order to determine the best models that have minimal estimation error. In addition, as pavement condition data also is panel data, varieties of panel data models can be explored, such as fixed-effect, mixed-effect, and random-effect models.

This study focused on partitioning the pavement data such that the resultant PPMs had minimum estimation errors. However, it did not investigate the resulting distribution of the pavement samples across clusters in order to identify the most critical or dominant variable in each cluster. Therefore, this study did not provide any justification of whether the assignment of pavement samples to the clusters truly represented the underlying clustering structure. Future work could include validating the assignment of pavement samples to the clusters.

Due to unavailability of data, a few potential explanatory variables that were proven to be significant in previous studies were not used. For example, the effect of pavement structure was not considered in the analysis. The structural numbers of pavements could be used to study the effect of the pavement structure on performance.

This study accounted for the effect of historical maintenance activities by setting pavement age to zero when a new maintenance activity was performed. However, routine maintenance works were ignored. In addition, the effects of various types of the maintenance activities were assumed to be the same. To address this limitation, maintenance types could be used as an extra explanatory variable in the analysis.

Pavement performance is governed by environmental, subsurface, and load-related factors, among others. The effects of these factors are not independent from one another, and potential interactions of various factors largely affect pavement performance. For example, extensive rutting and shoving on a flexible pavement can occur due to heavy traffic and high temperatures (Huang, 2003; Aguiar-Moya and Prozzi, 2011). Hence, adequate PPMs should capture the effects of the potential interactions of governing factors. In such circumstances, the PPMs could have a significantly large number of predictors; as a result, estimation of the model parameters could become very challenging. A non-parametric modelling approach could be utilized to estimate model parameters accurately. This study did not consider the effects of potential interactions of explanatory variables. Future research is recommended as well on the use of a non-parametric modelling approach, which is flexible enough to handle a large numbers of predictors. Details about non-parametric modelling can be found in Kang and Ghosal (2008), Attoh-Okine et al. (2009) and Ghahramani (2013).

Optimization parameters were obtained using past experience of the research team as well as sensitivity analysis. However, the existing literature (Park and Kim 1988, Johnson et al. 1987) included a few standard methods to determine such parameters. Future research could use optimization parameters that were determined from these proven techniques.

Simulated Annealing primarily was used for clustering. As clustering methods are highly sensitive to the choice of algorithms, different types of algorithms could be investigated to select the best. For example, genetic algorithms (Shekharan, 2000), artificial neural networks (Attoh-

Okine 1999), particle swarm optimization (ver der Merew and Engelbrecht 2003), and combination of these could be explored.

REFERENCES

Aarts, E., Korst, J., and Michiels, W. (2005). Simulated Annealing. *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*, Springer US, 187–210, DOI: 10.1007/0-387-28356-0_7.

AASHTO. (2012). *Pavement management guide, 2nd Edition*. Washington, DC.

Abdul-Wahaba, S.A., Bakheitb, C.S., and Al-Alawia, S.M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, 20(10), 1263–1271.

Agarwal, P., Das, A., and Chakroborty, P. (2006). Simple Model for Structural Evaluation of Asphalt Concrete Pavements at the Network Level. *J. Infrastruct. Syst.*, 12(1), 41-49. DOI: 10.1061/(ASCE)1076-0342(2006)12:1(41).

Aguiar-Moya, J.P., and Prozzi, J. (2011). Development Of Reliable Pavement Models. *Report No. SWUTC/11/161025-1*, Texas Transportation Institute Texas, A&M University System, Texas. ⟨http://static.tti.tamu.edu/swutc.tamu.edu/publications/technicalreports/161025-1.pdf⟩ (December 2016).

Anastasopoulos, P., and Mannering, F. (2014). Analysis of Pavement Overlay and Replacement Performance Using Random Parameters Hazard-Based Duration Models. *J. Infrastruct. Syst.*, 21(1), 04014024. DOI: 10.1061/(ASCE)IS.1943-555X.0000208.

Anily, S., and Federgruen, A. (1987). Simulated annealing methods with general acceptance probabilities. *Journal of Applied Probability* 24 (3): 657–667.

Archilla, A. (2006). Repeated Measurement Data Analysis in Pavement Deterioration Modeling. *J. Infrastruct. Syst.*, 12(3), 163-173. DOI: 10.1061/(ASCE)1076-0342(2006)12:3(163).

Attoh-Okine, N. (1999). Analysis of learning rate and momentum term in backpropagation neural network algorithm trained to predict pavement performance, *Advances in Engineering Software*, 30(4), 291–302.

Attoh-Okine, N., and Adarkwa, O. (2013). *Pavement Condition Surveys –Overview of Current Practices.* Delaware Center for Transportation 245. ⟨https://sites.udel.edu/dct/files/2013/10/Rpt-245-Pavement-Condition-Okine-DCTR422232-1pzk0uz.pdf⟩ (January 2016).

Attoh-Okine, N., Coogerb, K., and Mensaha, S. (2009). Multivariate adaptive regression

(MARS) and hinged hyperplanes (HHP) for doweled pavement performance modelling, *Construction and Building Materials*, 23(9), 3020–3023.

Bardaka, E., Labi, S., and Haddock, J.E. (2014). Using Enhanced Econometric Techniques to Verify the Service Life of Asset Intervention A Case Study for Indiana. *Transportation Research Record 2431*, Transportation Research Board, Washington, DC, 16–23.

Baumann, K. (2003). Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry*, 22(6), 395–406, DOI: 10.1016/S0165-9936(03)00607-1.

Berk, K.N. (1978). Comparing Subset Regression Procedures. *Technometrics*, 20(1), 1–6.

Brusco, M.J. (2014). A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. *Computational Statistics & Data Analysis*, 77, 38–53.

Brusco, M.J., Cradit, J.D., Steinley, D., and Fox, G.L. (2008). Cautionary Remarks on the Use of Clusterwise Regression. *Multivariate Behavioral Research*, 43(1), 29–49.

Buchheit, R.B., Garrett Jr., J.H., McNeil, S., and Chen, P. (2005). Automated Procedure to Assess Civil Infrastructure Data Quality: Method and Validation. *Journal of Infrastructure Systems*, 11(3), 180–189.

Carbonnea, R.A., Caporossi, G., and Hansen, P. (2011). Globally optimal clusterwise regression by mixed logical-quadratic programming. *European Journal of Operational Research*, 212(1), 213–222.

Chan, P., Oppermann, M., and Wu, S.S. (1997). North Carolina's Experience in Development of Pavement Performance Prediction and Modeling. *Transportation Research Record 1592*, Transportation Research Board, Washington, DC, 80–88, DOI: 10.3141/1592-10.

Chatterjee, S., and Hadi, A.S. (2000). *Regression analysis by example*, John Wiley and Sons, New York, USA.

Chen, D., and Mastin, N. (2015). Sigmoidal models for predicting pavement performance conditions. *J. Perform. Constr. Facil.*, 30(4). DOI:10.1061/(ASCE)CF.1943-5509.0000833.

Chen, D., Cavalline, T.L., Ogunro, V.O., and Thompson, D.S. (2014). Development and Validation of Pavement Deterioration Models and Analysis Weight Factors for the

NCDOT Pavement Management System. *FHWA/NC/2011-01_Phse II*, NCDOT, Raleigh, NC.

Christopher, B.R., Schwartz, C., and Boudreau, R. (2006). Geotechnical Aspects of Pavements. *Report No. FHWA NHI-05-037,* U.S. Department of Transportation, Washington, D.C.

Collins, N.E., Eglese, R.W., and Golden, B.L. (1988). Simulated Annealing – An Annotated Bibliography. *American Journal of Mathematical and Management Sciences* 8 (3–4): 209–307.

Darter, M.I. (1980). Requirements for Reliable Predictive Pavement Models. *Transportation Research Record 766*, Transportation Research Board, Washington, DC, 25-31.

Davies, R.M., and Sorenson, J. (2000). Pavement preservation: Preserving our investment in highways. *Public Roads*, 63(2), 63–69.

de Melo Silva, F., Van Dam, T., Bulleit, W., Ross Ylitalo, R. (2000). Proposed Pavement Performance Models for Local Government Agencies in Michigan. *Transportation Research Record 1699*, Transportation Research Board, Washington, DC, 81–86. DOI: 10.3141/1699-11.

DeSarbo, W.S., and Corn, W.L. (1988). A Maximum Likelihood Methodology for Clusterwise Linear Regression. *Journal of Classification*, 5(2), 249–282.

DeSarbo, W.S., Oliver, R.L., and Rangaswamy, A. (1989). A Simulated Annealing Methodology for Clusterwise Linear Regression. *Psychometrika*, 54(4), 707–736.

Dolan, W.B., Cummings, P.T., and LeVan, M.D. (1989). Process optimization via simulated annealing: Application to network design. *AIChE Journal*, 35(5), 725–736, DOI: 10.1002/aic.690350504.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics*, 7(1), 1–26.

Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2), 407–499.

Farhan, J., and Fwa, T.F. (2015). Improved Imputation of Missing Pavement Performance Data Using Auxiliary Variables. *Journal of Transportation Engineering*, 141(1), 1–8.

Fowlkes, E.B., Gnanadesikan, R., and Kettenring, J.R. (1988). Variable Selection in Clustering. *Journal of Classification*, 5(2), 205–228.

Fox, J., and Monette, G. (1992). Generalized Collinearity Diagnostics, *Journal of the American Statistical Association,* 87(417), 178-183.

Garcia-Diaz, A., and Riggins, M. (1984). Serviceability and Distress Methodology for Predicting Pavement Performance. *Transportation Research Record 997*, Transportation Research Board, Washington, DC, 56-61.

Garside, M.J. (1965). The Best Subset in Multiple Regression Analysis. *Applied Stat. Journ. of the Royal Statist. Society, Series C.*, 14(2), 196–200.

George, K.P., Rajagopal, A.S., and Lim, L.K. (1989). Models for predicting pavement deterioration. *Transportation Research Record 1215*, Transportation Research Board, Washington, DC, 1–7.

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling, *Phil Trans R Soc A*, 371: 20110553. DOI: 10.1098/rsta.2011.0553.

Gharaibeh, N., Wimsatt, A., Saliminejad, S., Menendez, J.R., Weissmann, A.J., Weissmann, J., and Chang-Albitres, C. (2012). Implementation of New Pavement Performance Prediction Models in PMIS. *Report. FHWA/TX-12/5-6386-01-1*.

Gorman, J.W., and Toman, R.J. (1966). Selection of Variables for Fitting Equations to Data. *Technometrics*, 8(1), 27–51, DOI: 0.1080/00401706.1966.10490322.

Gunst, R.F., and Webster, J.T. (1975). Regression analysis and problems of multicollinearity. *Communications in Statistics*, 4(3), 277–292, DOI: 10.1080/03610927308827246.

Guo, J.Q., and Zheng, L. (2005). A modified simulated annealing algorithm for estimating solute transport parameters in streams from tracer experiment data. *Environmental Modelling & Software* 20: 811–815.

Gupta, M., and Ibrahim, J.G. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association*, 102(479), 867–880.

Hajj, E.Y., Loria, L., and Sebaaly, P.E. (2010). Performance Evaluation of Asphalt Pavement Preservation Activities. *Transportation Research Record: Journal of the Transportation Research Board*, 2150, 36–46.

Hand, A.J., Seebaly, P.E., and Epps, J.A. (1999). Development of Performance Models Based on Department of Transportation Pavement Management System Data. *Transportation Research Record 1684*, Transportation Research Board, Washington, DC, 215-222.

Highway Research Board. (1962). The AASHO road test. *Special Rep. No. 61A-E*, National Academy of Science, National Research Council, Washington, DC.

Hocking, R.R., and Leslie, R.N. (1967). Selection of the Best Subset in Regression Analysis. *Technometrics*, 9(4), 531–540.

Hong, F., and Prozzi, J.A. (2006). Estimation of Pavement Performance Deterioration Using Bayesian Approach. *J. Infrastruct. Syst.*, 12(2), 77-86. DOI: 10.1061/(ASCE)1076-0342(2006)12:2(77).

Hong, F., and Prozzi, J.A. (2010). Roughness Model Accounting for Heterogeneity Based on In-Service Pavement Performance Data. *J. Transp. Eng.* 136 (3): 205–213.

Hong, F., and Prozzi, J.A. (2015). Pavement Deterioration Model Incorporating Unobserved Heterogeneity for Optimal Life-Cycle Rehabilitation Policy. *J. Infrastruct. Syst.* 21(1): 1–11.

Hsiao, C. (2003). *Analysis of panel data*, *2nd Ed.* Cambridge University Press, Cambridge, U.K.

Hsu, D. (2015). Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. *Applied Energy*, 160, 153–163.

Huang, Y.H. (2003). *Pavement Analysis and Design*, 2nd Edition. Prentice Hall, Inc. New Jersey.

Hudson, W.R., Haas, R., and Perrone, E. (2015). Measures of Pavement Performance must consider the Road User. *Proc.,9th International Conference on Managing Pavement Assets*, Alexandria, VA.

Hudson, W.R., Monismith, C.L., Shook, J.F., Finn, F.N., and Skok Jr, E.L. (2007). AASHO road test effect on pavement design and evaluation after 50 years. *Transportation Research Circular E-C118*, 17–30.

Johnson, D.S., Aragon, C.R., McGeoch, L.A., and Schevon, C. (1989). Optimization by simulated annealing: an experimental evaluation; part I, graph partitioning. *Operations research*, 37(6), 865–892.

Kang, C., and Ghosal, S. (2008). Clusterwise regression using Dirichlet process mixtures. *Advances in Multivariate Statistical Methods*, 305–325.

Kay, R.K., Mahoney, J.P., and Jackson, N.C., 1993. The WSDOT Pavement Management System – A 1993 Update. *Technical Rep. WA-RD 274.1*, Washington State Dept. of Transportation, Olympia, WA.

Ker, H.W., and Lee, Y.H. (2011). Preliminary Analysis of Flexible Pavement Performance Data Using Linear Mixed Effects Models. *Electrical Engineering and Applied Computing*, 90, 351-363.

Ketchen, D.J., and Shook, C.L. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, 17(6), 441–458.

Khadka, M., and Paz, A. (2017). Estimation of optimal pavement performance models for highways. *Tenth International Conference on the Bearing Capacity of Roads, Railways and Airfields*. Athens, Greece.

Khatta, M.J., Baladi, G., Zhang, Z., and Ismail, S. (2008). Review of Louisiana's pavement management system—Phase I. *Transportation Research Record 2084*, Transportation Research Board, Washington, DC, 18–27.

Khraibani, H., Lorino, T., Lepert, P., and Marion, J.M. (2012). Nonlinear Mixed-Effects Model for the Evaluation and Prediction of Pavement Deterioration. *J. Transp. Eng.*, 138(2), 149-156. DOI: 10.1061/(ASCE)TE.1943-5436.0000257.

Kim, S., and Kim, N. (2006). Development of performance prediction models in flexible pavement using regression analysis method. *KSCE J. of Civil Engineering*, 10(2), 91–96.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by Simulated Annealing. *Science* 220 (4598): 671–680.

Kodinariya, T.M., and Makwana, P.R. (2013). Review on determining number of cluster in k-means clustering. *International Journal of Advanced Research in Computer Science and Management Studies*, 1(6), 90–95.

Labi, S., and Sinha, K.C. (2003). Life-cycle evaluation of highway pavement preventive maintenance. [CD-ROM], *82nd annual meeting of the Transportation Research Board*, National Research Council, Washington DC.

Lau, K., Leung, P., and Tse, K. (1999). A mathematical programming approach to clusterwise regression model and its extensions. *European Journal of Operational Research*, 116(3), 640–652.

Lee, D. (2007). Dynamic Prediction Model of As-Built Roughness in Asphaltic Concrete Pavement Construction. *J. Transp. Eng.*, 133(2), 90-95. DOI: 10.1061/(ASCE)0733-947X(2007)133:2(90).

Li, N., Hass, R., and Xie, W.C. (1997). Investigation of Relationship Between Deterministic and Probabilistic Prediction Models in Pavement Management. *Transportation Research Record 1592,* Transportation Research Board, Washington, DC, 70–79, DOI: 10.3141/1592-09.

Li, Z. (2005). *A Probabilistic and Adaptive Approach to Modeling Performance of Pavement Infrastructure*. Thesis (PhD). University of Texas at Austin.

Liu, H.H., and Ong, C.S. (2008). Variable selection in clustering for marketing segmentation using genetic algorithms. *Expert Systems with Applications*, 34(1), 502–510.

Lu, H., Huang, S., Li, Y., and Yang, Y. (2014). Panel Data Analysis Via Variable Selection and Subject Clustering. *Data Mining for Services*, 3, Springer-Verlag, 31–76.

Luo, Z., and Chou, E.Y. (2006). Pavement Condition Prediction Using Clusterwise Regression. *Transportation Research Record 1974*, Transportation Research Board, Washington, DC, 70–77, DOI: 10.3141/1974-11.

Luo, Z., and Yin, H. (2008). Probabilistic Analysis of Pavement Distress Ratings with the Clusterwise Regression Method. *Transportation Research Record 2084*, Transportation Research Board, Washington, DC, 38–46, DOI: 10.3141/2084-05.

Lytton, R.L. (1987). Concepts of Pavement Performance Prediction and Modeling. Vol. 2, Proceedings, Second North American Conference on Managing Pavements, Ontario Ministry of Transportation, Toronto, Canada.

Madanat, S.M., Nakat, Z., and Sathaye, N. (2008). Development of Empirical-Mechanistic Pavement Performance Models using Data from the Washington State PMS Database. *Research Report UCPRC-RR-2005-05*, Univ. of Cal., Davis and Berkeley, ⟨http://www.dot.ca.gov/newtech/researchreports/reports/2008/ucprc-rr-2005-05.pdf⟩ (December 2016)

Mallows, C.L. (1973). Some Comments on Cp. *Technometrics*, 15(4), 661–675.

Marler, R., and Arora, J. (2004). Survey of multi-objective optimization methods for engineering.

Structural and Multidisciplinary Optimization, 26, 369– 395.

Maugis, M., Celeux, G., and Martin-Magniette, M.L. (2009). Variable Selection for Clustering with Gaussian Mixture Models. *Journal of the International biometric society*, 65(2), 701-709.

Maydeu-Olivares, A., and García-Forero, C. (2010). Goodness-of-fit testing. *International Encyclopedia of Education*, 7, 190–196.

Mehmood, T., Liland, K.H., Snipen, L., and Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 118, 62–69.

Meneses, S., and Ferreira, A. (2013). Pavement maintenance programming considering two objectives: maintenance costs and user costs, *International Journal of Pavement Engineering*, 14(2), 206-221, DOI: 10.1080/10298436.2012.727994.

Midi, H., Sarkar, S.K., and Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3), 253–267.

Myers, R.H. (1990). *Classical and Modern Regression with Applications, 2nd edition*, PWS Kent, Boston, MA.

Nakamura, V.F., and Michael, H.L. (1963). Serviceability Ratings of Highway pavements. *Highway Research Record 40*, Highway Research Board, 21-36.

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). *Applied linear statistical models*, Irwin, Chicago, Illinois, USA.

Nikolaev A., and Jacobson, S. (2010). Simulated annealing. *Handbook of Metaheuristics, International Series in Operations Research and Management Science*, 146, Springer, Berlin, 1–39, DOI: 10.1007/978-1-4419-1665-5_1.

Ortiz-García, J., Costello, S., and Snaith, M. (2006). Derivation of Transition Probability Matrices for Pavement Deterioration Modeling. *J. Transp. Eng.*, 132(2), 141-161. DOI: 10.1061/(ASCE)0733-947X(2006)132:2(141).

Park, M.W., and Kim, Y.D. (1998). A systematic procedure for setting parameters in simulated annealing algorithms. *Comput. Oper Res.* 25 (3): 207–217.

Park, Y.W., Jiang, Y., Klabjan, D., and Williams, L. (2015). Algorithms for Generalized Clusterwise Linear Regression. 〈http://dynresmanagement.com/uploads/3/3/2/9/3329212/gclr.pdf〉 (January 2016).

Paz, A., and Khadka, M, (2017). Limitations of Existing Pavement Performance Models and a Potential Solution. *World Conference on Pavement and Asset Management (WCPAM2017)*, Baveno, Italy.

Paz, A., Molano, V., and Sanchez, M. (2015a). Holistic Calibration of Microscopic Traffic Flow Models: Methodology and Real World Application Studies. *Engineering and Applied Sciences Optimization: Dedicated to the memory of Professor M.G. Karlaftis*, 38, Ed. 1. Springer International Publishing.

Paz, A., Molano, V., Martinez, E., Gaviria, C., and Arteaga, C. (2015b). Calibration of Traffic Flow Models Using a Memetic Algorithm. *Transportation Research Part-C: Emerging Technologies*, 55, 432–443.

Petraitis, P.S., Dunham, A.E., and Niewiarowski, P.H. (1996). Inferring multiple causality: the limitations of path analysis. *Functional Ecology*, 10(4), 421–431.

Pierce, L.M., McGovern, G., and Zimmerman, K.A. (2013). *Practical Guide for Quality Management of Pavement Condition Data Collection*, FHWA, U.S. Department of Transportation. ⟨http://www.fhwa.dot.gov/pavement/management/qm/data_qm_guide.pdf⟩ (December 2016)

Preda, C., and Saporta, G. (2007). PCR and PLS for clusterwise regression on functional data. *Selected Contributions in Data Analysis and Classification*, Springer-Berlin, 589–598. DOI: 10.1007/978-3-540-73560-1.

Prozzi, J.A., and Madanat, S.M. (2003). Incremental Nonlinear Model for Predicting Pavement Serviceability. *J. Transp. Eng.*, 129(6), 635-641. DOI: 10.1061/(ASCE)0733-947X(2003)129:6(635).

Prozzi, J.A., and Madanat, S.M. (2004). Development of pavement performance models by combining experimental and field data. *J. of Infrastructure Systems*, 10(1), 9–22.

Pulugurta, H. (2007). Development of Pavement Condition Forecasting Models. PhD diss., University of Toledo.

R Core Team. (2015). *R: A language and environment for statistical computing*, Vienna: R Foundation for Statistical Computing. ⟨http://www.R-project.org⟩ (November 2016).

Rajagopal, A. (2006). Developing Pavement Performance Prediction Models and Decision Trees for the City of Cincinnati. *FHWA/OH-2006/14*, Ohio DOT, Columbus, OH.

Rao, C.R., and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2), 369–374.

Román-Román, P., Romero, D., Rubio, M.A., and Torres-Ruiz, F. (2012). Estimating the parameters of a Gompertz-type diffusion process by means of Simulated Annealing. *Applied Mathematics and Computation*, 218(9), 5121–5131.

Rose, J., Klebsch, W., and Wolf, J. (1990). Temperature measurement and equilibrium dynamics of simulated annealing placement. *IEEE Transactions on Computer Aided Design* 9 (3): 253–259.

Roshan, S.B., Jooibari, M.B., Teimouri, R., Asgharzade-Ahmadi, G., Falahati-Naghibi, M., and Sohrabpoor, H. (2013). Optimization of friction stir welding process of AA7075 aluminum alloy to achieve desirable mechanical properties using ANFIS models and simulated annealing algorithm. *International Journal of Advanced Manufacturing Technology* 69 (5-8): 1803–1813.

Rutenbar, R.A. (1989). Simulated Annealing Algorithms: An Overview. *IEEE Circuits and Devices Magazine*, 5(1), 19–26.

Sadek, A.W., Freeman, T.E., and Demetsky, M.J. (1996). Deterioration Prediction Modeling of Virginia’s Interstate Highway System. *Transportation Research Record 1524*, Transportation Research Board, Washington, DC, 118–129, DOI: 10.3141/1524-14.

Salama, H., Chatti, K., and Lyles, R. (2006). Effect of Heavy Multiple Axle Trucks on Flexible Pavement Damage Using In-Service Pavement Performance Data. *J. Transp. Eng.*, 132(10), 763-770.

Saraf C.L., and Majizzadeh, K. (1992). Distress prediction models for a network level PMS. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1344, 38–48.

Schlittgen, R. (2011). A weighted least-squares approach to clusterwise regression. *Advances Statistical Analysis*, 95(2), 205–217.

Schmitt, R.L., Owusu-Ababio, S., and Denn, K.D. (2008). Database Development for an HMA Pavement Performance Analysis System. *Report No. 07-11*, University of Wisconsin-Madison, ⟨http://wisconsindot.gov/documents2/research/06-13hmadatabase-f.pdf⟩ (December 2016).

Schram, S.A. (2008). *Mechanistic-Empirical Modelling and Reliability in Network-Level Pavement Management*. Thesis (PhD). North Dakora State University.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.

Selim, S.Z., and Alsultan, K. (1991). A simulated annealing algorithm for the clustering problem. *Pattern Recognition* 24 (10): 1003–1008.

Shahin, M.Y. (1994). *Pavement Management for Airports, Roads, and Parking Lots*. Boston, MA: Springer US.

Shekharan, A.R. (2000). Solution of Pavement Deterioration Equations by Genetic Algorithms. *Transportation Research Record 1699*, Transportation Research Board, Washington, DC, 101–106, DOI: 10.3141/1699-14.

Shoukry, S.N., Martinelli, D.R., and Reigle, J.A. (1997). Universal Pavement Distress Evaluator Based on Fuzzy Sets, *Transportation Research Record 1592*, Transportation Research Board, Washington, D.C., 180–186.

Spath, H. (1979). Algorithm 39: Clusterwise linear regression. *Computing*, 22(4), 367–373.

Sridhar, J., and Rajendran, C. (1993). Scheduling in a cellular manufacturing system: a simulated annealing approach. *International Journal of Production Research*, 31(12), 2927-2945, DOI: 10.1080/00207549308956908

Stampley, B.E., Miller, B., Smith, R.E., Scullion, T. (1995). Pavement Management Information System Concepts, Equations and Analysis Models. Texas Transportation Institute, Research Report TX 96/1989-1, August.

Steinbach, M., Ertoz, L., and Kumar, V. (2003). Challenges of clustering high dimensional data. *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition*. Springer-Verlag. 273-309. doi: 10.1007/978-3-662-08968-2_16.

Sun, L.X., Xie, Y.L., Song, X.H., Wang, J.H., Yu, R.Q. (1994). Cluster analysis by simulated annealing, *Computers & Chemistry*, 18(2), 103-108.

Sundin, S., and Braban-Lexdoux, C. (2001). Artificial intelligence-based decision support technologies in pavement management. *Comput. Aided Civ. Infrastruct. Eng.*, 16(2), 143–157.

Tacq. J. (1997). *Multivariate analysis techniques in social science research: From problem to analysis*, Sage Publications, London.

Tan, S.G., and Cheng, D. (2014). Quality Assurance of Performance Data for Pavement Management Systems. *2014 Geohubei International Conference. ASCE GSP 246: Design, Analysis, and Asphalt Material Characterization for Road and Airfield Pavement*. 163-169.

Tan, T., Suk, H.W., Hwang, H., and Lim, J. (2013). Functional fuzzy clusterwise regression analysis. *Adv. Data Anal. Classif.*, 7(1), 57–82.

Terzi, S. (2006). Modeling the Pavement Present Serviceability Index of Flexible Highway Pavements Using Data Mining. *Journal of Applied Sciences*, 6(1), 193–197.

Thompson, M.L. (1978). Selection of Variables in Multiple Regression: Part I. A Review and Evaluation. *International Statistical Review*, 46(1), 1–19, DOI: 10.2307/1402505.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58, 267–288.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a dataset via the gap statistic. *J. Roy. Stat. Soc. B* 63 (2): 411–423.

van der Merwe, D.W., and Engelbrecht, A.P. (2003). Data clustering using particle swarm optimization. Proc. IEEE Congress on Evolutionary Computation 2003 (CEC 2003), Canbella, Australia, 215-220. DOI: 10.1109/CEC.2003.1299577

Vrieze, S.I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychol Methods*, 17(2), 228–243.

Wang, Y. (2013). Ordinal Logistic Regression Model for Predicting AC Overlay Cracking. *J. Perform. Constr. Facil.*, 27(3), 346-353. DOI: 10.1061/(ASCE)CF.1943-5509.0000327.

Wedel, M., and SteenKamp, J. (1989). Fuzzy clusterwise regression approach to benefit segmentation. *Int. J. Res. Mark.*, 6(4), 241–258.

Wolters, A.S., and Zimmerman, K.A. (2010). Research of Current practices in pavement performance modeling. *FHWA –PA-2010-007-080307*, PennDOT, Harrisburg, PA.

⟨http://ntl.bts.gov/lib/32000/32900/32969/Research_of_Current_Practices_in_Pavement_Performance_Modeling.pdf⟩ (December 2016).

Wooldridge, J.M. (2006). *Introductory econometrics: A modern approach*. Mason, OH: Thomson/South-Western.

Wu, J. (2014). Model-based clustering and model selection for binned data. PhD diss., SUPELEC.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92(4), 937–950.

Yoo, W., Mayberry, R., Bae, S., Singh, K., (Peter) He, Q., and Lillard, J.W. (2014). A Study of Effects of MultiCollinearity in the Multivariable Analysis. *Int J Appl Sci Technol.*, 4(5), 9–19.

Yu, J., Chou, E., and Luo, Z. (2007). Development of Linear Mixed Effects Models for Predicting Individual Pavement Conditions. *J. Transp. Eng.*, 133(6), 347-354. DOI: 10.1061/(ASCE)0733-947X(2007)133:6(347).

Zhang, W., and Durango-Cohen, P. (2014). Explaining Heterogeneity in Pavement Deterioration: Clusterwise Linear Regression Model. *J. Infrastruct. Syst.*, 20(2). DOI: 10.1061/(ASCE)IS.1943-555X.0000182.

Zhang, Z., Singh, N., and Hudson, W.R. (1993). Comprehensive Ranking Index for Flexible Pavement Using Fuzzy Sets Model. *Transportation Research Records 1397*, Transportation Research Board, Washington, D.C., 96-102.

Zhen, Z., Yan, L., and Nan, K. (2012). Clusterwise linear regression with the least sum of absolute deviations - An MIP approach. *Int. J. Oper. Res.*, 9(3), 62172.

Zheng, L. (2005). *A Probabilistic and Adaptive Approach to Modeling Performance of Pavement Infrastructure*. Thesis (PhD). The University of Texas at Austin.

CURRICULAM VITAE

Graduate College
University of Nevada, Las Vegas

Mukesh Khadka

1600 E Rochelle Ave.
Las Vegas, NV 89119
E-mail: imukeshkhadka@gmail.com

Degrees:

Bachelor of Engineering in Civil Engineering, 2007
Institute of Engineering, Tribhuvan University, Nepal

Master of Engineering in Construction, Engineering and Infrastructure Management, 2011
Asian Institute of Technology, Thailand

Dissertation Title: Generalized Clusterwise Regression for Simultaneous Estimation of Optimal Pavement Clusters and Performance Models.

Dissertation Examination Committee:

Committee Chair, Alexander Paz, Ph.D.
Committee Member, Mohamed Kaseko, Ph.D.
Committee Member, Moses Karakouzian, Ph.D.
Committee Member, Pramen P. Shrestha, Ph.D.
Graduate College Representative, Ashok K. Singh, Ph.D.