COMPREHENSIVE CLUSTERWISE LINEAR REGRESSION FOR PAVEMENT MANAGEMENT SYSTEMS

Mukesh Khadka
Ph.D. Candidate
Cell: (702) 683-2722
Email: khadkam@unlv.nevada.edu

Alexander Paz, Ph.D., P.E., Corresponding Author
Associate Professor
Director, Transportation Research Center
Office: (702) 895-0571
Cell: (702) 688-3878
Fax: (702) 895-3936
Email: apaz@unlv.edu
http://web.unlv.edu/centers/trc/paz/

Civil and Environmental Engineering and Construction
University of Nevada, Las Vegas
4505 Maryland Parkway, PO Box 454007, Las Vegas, NV 89154-4007

ABSTRACT

A comprehensive mathematical program was formulated to determine simultaneously 1) an optimum number of pavement clusters, 2) cluster memberships of pavement samples, 3) cluster-specific significant explanatory variables, and 4) estimated regression coefficients for Pavement Performance Models (PPMs). Simulated Annealing coupled with All-Subset Regression was proposed to solve the mathematical programming. The proposed algorithm was capable to identify and address potential multicollinearity issues. All possible combinations of the explanatory variables were examined to select the best model that provided a balance among 1) the number of PPMs; 2) the number of explanatory variables; 3) the resources required to develop, maintain, and use these models; and 4) the explanatory power. For the dataset used in this research, 6-cluster models were determined as part of the optimum solution. The predictive capabilities of the resultant models were investigated, and results showed that the models provided few prediction errors without any overfitting issues.

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 requiring 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 so that the condition of a pavement 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 determine this balance, PPMs typically are developed by using clusters of pavement samples. Instead of estimating the cluster memberships by using statistical methods, a few predefined explanatory variables are used to assign pavement samples into clusters. In terms of performance, clusters formed in this way likely include heterogeneous pavement samples.

The 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, various clusters are formed so that homogenous pavement samples, in terms of the effects of the explanatory variables on the dependent variable of a present regression model, are assigned within a cluster (Park et al. 2015). The homogeneity of pavement samples in a cluster is defined by the effects of the observed values of explanatory variables on the estimated dependent variable, the Present Serviceability Index (PSI), by the regression model. Observations of all the pavement samples assigned to a cluster fit the same PPM such that the overall sum of squared errors (SSE) within clusters is minimal.

CLR first was 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). However, in the field of pavement management, to the best knowledge of the authors, only four 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. The study used the data collected during the AASHO Road Test (Highway Research Board 1962), which is no longer the best available data nor representative of existing conditions. This data was collected at a single site, and over 50 years ago, when materials and construction techniques were different.

The study estimated models with the objective of minimization of the residual sum of squares (RSS). The number of models were determined subjectively using the trends of RSS and Akaike Information Criteria (AIC) over the number of clusters. In addition, the study investigated the presence of overfitting in the CLR models, using a procedure proposed by Brusco et al. (2008). In this current study, overfitting means that most of the variations in the dependent variable appears to be explained by the estimated model; however, the actual relationship between the dependent variable and some of the explanatory variables and/or the functional form of the model is not really captured. Overfitting typically is evidenced during validation when the model is used to estimate values for the dependent variable, using data that was not used for model development. Later in this paper, the section on Model Performance provides a rigorous explanation of a procedure to determine potential overfitting in a model.

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 (Khadka and Paz, 2017b). In this study, the Bayesian Information Criteria (BIC) (Schwarz 1978) was used as the objective function. BIC penalizes more for the inclusion of additional parameters than does AIC (Kadane and Lazar 2004). On the other hand, several studies showed that the number of parameters in a model selected using AIC was overestimated (Geweke and Meese, 1981; Katz, 1981; Koehler and Murphree, 1988; Kadane and Lazar 2004).

BIC is one of the most popular log-likelihood-based information criteria used for model selection. As BIC is an increasing function of the error variance and free parameters to be estimated, minimizing BIC reduces unexplained variations in the dependent variable, the number of explanatory variables, or both (Uzoma and Jeremiah, 2016). In case of a large sample size, BIC is consistent in the sense that the probability of the selected model being the true model approaches ‘1’ (Rao and Wu 1989; Yang 2005; Maydeu-Olivares and García-Forero 2010; Vrieze 2012, Kim et al. 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 the limitation that variables included in the PPMs are assumed to be significant. However, the effects of variables without any evidence of significance can affect clustering and regression analyses. Therefore, heterogenous samples can be assigned together erroneously (Fowlkes et al. 1988); therefore, it becomes challenging to discover the underlying pavement clusters that exhibit similar performance behavior (Gupta and Ibrahim 2007).

This problem is illustrated in Figure 1, using data from the Pavement Management System (PMS) of the Nevada Department of Transportation (NDOT). In this example, 54 randomly selected pavement samples were considered. Each pavement sample was represented by a dependent variable, PSI, and two explanatory variables, Age and Average Daily Traffic (ADT).

The variables PSI and Age had a significant linear relationship (p-value = 0.001), as shown in Figure 1a. The estimated BIC and root mean square error (RMSE) for the model were 85 and 0.2916, respectively. However, the relationship between PSI and ADT was not clear, as shown in Figure 1b. The estimated BIC and RMSE for the model were 251 and 0.4572, respectively. When both Age and ADT were included in the model as explanatory variables, the estimated BIC was increased to 90, with a slight decrease in RMSE by 0.0003. Hence, if an irrelevant variable, ADT in this example, is included in a CLR analysis without checking its significance, it increases the BIC. In addition, it causes a loss of efficiency in the model. The estimated clustering and regression models may not capture the correct underlying relationships among the variables when a variable is included in the model without sufficient evidence of its significance.

Assignment of pavement samples 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 because 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 underlying cluster assignments.

A legitimate assignment of pavement samples into homogeneous clusters to minimize the estimation error can be obtained using the relevant explanatory variables that exhibit the strongest effects on the dependent variable (Fowlkes et al. 1988; Liu and Ong 2008; and Maugis et al. 2009). The strength of the effects of explanatory variables on the dependent variable often is assessed by comparing p-values with the desired level of significance (α). A p-value represents the significance of the estimated coefficient for an explanatory variable. If the p-value for an explanatory variable is greater than α, there is not enough evidence to claim that the estimated coefficient is likely to be different from zero. In other words, changes in the explanatory variable do not reflect changes in the dependent variable. Hence, such explanatory variables having p-values greater than the desired α usually are excluded from the model during model estimation 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 in order to provide the best regression models. Typically, the fewest number of explanatory variables that sufficiently explain most of the variances in the dependent variable are selected as the best model specification. In terms of data analysis and statistics, numerous methodologies for variable selection are available in the literature (Thompson 1978; Tibshirani 1996; Baumann 2003; Efron et al. 2004; Mehmood et al. 2012; Brusco 2014). In this study, the All-Subset 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 for CLR analysis. All (2^P-1) possible subsets of potential explanatory variables, P, 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 under the presence of multicollinearity (Gunst and Webster 1975). Strongly-correlated clustering variables may overweight one or more underlying constructs and produce loss in efficiency (Ketchen and Shook 1996). Typically, multicollinearity inflates the variance of regression parameters and makes correct identification of significant variables challenging (Abdul-Wahab et al. 2005; Dorman et al. 2013; Ohlemüller et al. 2008). However, strongly correlated variables may not be a problem in all cases (Harrell 2001). In addition, if the collinearity between two variables remains constant, their estimated parameters are likely to have low standard errors; the problem would be serious if the standard errors of the correlated variables are high (Washington et al. 2011). The best way to address multicollinearity is to conduct a carefully designed experiment that considers the trade-off between removing and keeping potential explanatory variables that are expected to cause multicollinearity. Judgement and iterations are required to determine the best model specification that minimizes the effects of multicollinearity (Washington et al. 2011).

This study investigated the effects of highly-correlated explanatory variables. The Variance Inflation Factor (VIF), used to examine potential issues due to multicollinearity (Marquardt 1970; Mansfield and Helms 1982), is defined as $1/(1-R_i^2)$, where R_i^2 is the R^2 for an explanatory variable, X_i regressed on the remaining explanatory variables. When no explanatory variables are correlated, the VIF is equal to '1'. As the degree of collinearity increases, both the variance of regression coefficient and the VIF increase (Yoo et al. 2014). Tacq (1997) showed that large VIF is an indicator of multicollinearity. 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 literature.

To avoid prespecifying the significance of potential explanatory variables, this paper 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 likely significant explanatory variables. 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. In the study, the likely significance of the explanatory variables was 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 illustrate the advantage of solving simultaneously for the three types of parameters listed above.

METHODOLOGY

Problem formulation

This section describes a mathematical program that was formulated to describe the proposed CLR problem. Among various pavement performance measures available in the literature, PSI 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). PSI is understood easily 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:

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

The objective function was to minimize BIC, expressed as:

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

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

$$SSE=\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{t=1}^{T_i}\left(\beta_{0k}+\sum_{j=1}^{J}\beta_{0k}*x_{ijt}-y_{it}\right)^2*p_{ik}\ \forall\ i\in I, j\in J, t\in T_i, k\in K \qquad (3)$$

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. In addition, the proposed mathematical programming included the following constraints:

Constraints for significant variables:

$$\delta = \sum_k\sum_j v_{jk}\ \forall\ j = 0,\ldots,J,\ \ k\in K \qquad (4)$$

$$v_{jk}=\begin{cases}1, & if\ \beta_{jk}\ is\ significant;\\ 0, & Otherwise\end{cases}\quad \forall\ j = 0,\ldots,J,\ k\in K \qquad (5)$$

Membership constraints:

$$\sum_k p_{ik} = 1 \; \forall i \in I,\, k \in K \tag{6}$$

$$p_{ik} = \begin{cases} 1, & if\;\; sample\ i\ is\ assigned\ to\ cluster\ k; \\ 0, & Otherwise \end{cases} \quad \forall\, i \in I,\, k \in K \tag{7}$$

Constraints for feasible partitions:

$$C_k = \left\{ i \middle| p_{ik} = 1 \forall i \in I,\, k \in K \right\} \tag{8}$$

$$C_{k'} \cap C_{k''} = \text{null} \;\; \forall k' \neq k'',\, k' \text{ and } k'' \in K \tag{9}$$

$$\bigcup_{k \in K} |C_k| = I \tag{10}$$

$$\sum_{i \in C_k} T_i \geq n \;\forall\, C_k \tag{11}$$

Constraints for range of clusters:

$$1 \leq k \leq K_{max} \tag{12}$$

$$K_{max} = \mathrm{F}(I, T_i, n) \tag{13}$$

The constraint expressed by (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 the constraint expressed by (5), the element v_{jk} was equal to '1' if an estimated coefficient (β_{jk}) was significant in cluster k; otherwise, v_{jk} was '0' (Eq. 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 expressed by (6) and (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 '0'.

The feasibility of the resulting clustering was guaranteed by constraints expressed by (8) - (11). Constraints expressed by (8) – (10) prevented the overlap of members among clusters; that is, pavement samples were divided exclusively into K clusters. Constraint (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 expressed by (12) and (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.

The constraint expressed by (13) denoted the maximum number of feasible clusters. A procedure to calculate this maximum number was denoted by function F (Khadka et al., 2017). The procedure iteratively searched for the best combinations of the pavement samples to form a cluster such that each cluster had the required minimum number of observations. In the first step, it searched pavement samples with *n* or more observations. In this case, each pavement sample could form a cluster and was assigned to an individual cluster. Once all such cases were searched, the procedure searched two or more pavement samples, where a total number of observations equaled to *n*. In this step, all possible combinations of pavement samples with a total number of observations equal to *n* were searched to find the maximum number of combinations. No sample could be assigned to more than one cluster. After assigning all possible combinations, the algorithm seeks for the combination(s) of samples having the minimal number of extra observations in each cluster.

Solution to the mathematical program

This study integrated Simulated Annealing (SA) (Krickpatrick et al. 1983; Cěrny, 1985) with Ordinary Least Square (OLS) to solve the proposed mathematical program, which is described as follows by means of algorithmic steps and a discussion regarding the details. SA was chosen because it provides a probabilistic mechanism to seek a global optimum in a large search space that involves discrete variables, such as cluster membership. Thus, SA was used to determine the cluster memberships (p_{ik}) of the pavement samples. For each accepted cluster, the VIF for all explanatory variables were calculated as discussed in the introduction. Highly correlated explanatory variables that had VIFs greater than a predefined limiting VIF were excluded. All-subset regressions were utilized to find the best model and to estimate the associated regression 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 is illustrated in Figure 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 the constraint expressed by (13).

Step 3. For a given K, randomly assign pavement samples into clusters using the following steps:

Step 3.1. Generate a random number $u \sim U(1, K)$ and assign it to each of the pavement sample used for the estimation of CLR models. When a sample is assigned to a cluster, all observations associated with that sample are assigned to this cluster.

Step 3.2. Find the total number of observations assigned to each of the clusters, (i.e., 1 to *K*).

Step 3.3. If all the clusters have at least *n* observations, then go to Step 4; otherwise, repeat Steps 3.1 and 3.2 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-subset 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-value* < α.

Step 5. Calculate the total number of free parameters to be estimated, (δ+*K*-1). Calculate *BIC* using Eq. 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 $u_1 \sim U(1, K)$. If the new membership is the same as previously, regenerate a random number $u_2 \sim U(1, K)$

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 Eq. 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 the acceptance probability, $p_{accept} = exp\left(\frac{-\Delta BIC}{B*T}\right)$, where B is the Boltzmann's constant. If $p_{accept} > u"$, accept the current set of clusters, C_K^{N+1}, and the corresponding 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 a global solution, this algorithm used a probabilistic approach during the search process. The initial solution was improved repetitively by making small changes until a better solution was obtained (Sridhar and Rajendran 1993; Johnson et al. 1989). The algorithm accepted better solutions as well as non-improving (worse) solutions at a certain probability (Dolan et al. 1989; Rutenbar 1989; Aarts et al. 2005). This probability decreased continuously over iterations, and depended 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 accepted worse solutions, which caused larger increments in BIC. As the temperature went down, the algorithm accepted worse solutions with relatively smaller increments in BIC. Finally, when the temperature dropped to zero, the algorithm no longer accepted worse solutions. This enabled occasional ‘uphill’ moves that helped the algorithm to escape from the local minima. Thus, the algorithm tried to explore the entire solution space to seek for a global solution (Dolan et al. 1989). Previous studies have shown that the algorithm converged to a global minimum when an infinitely slow cooling schedule was utilized (Román-Román et al. 2012).

Application of CLR Models

Luo and Chao (2008) proposed a procedure to apply CLR models to estimate pavement conditions. However, the proposed procedure applies only for cases when pavement age is the only independent variable. In addition, the procedure cannot be used to estimate the condition of a pavement sample that was not used to develop the CLR model. In other words, the procedure cannot be used to determine the cluster memberships of the pavement samples that are not included in the estimation process.

To address this issues, this study proposed a heuristic to closely assign the cluster membership to a pavement sample. It was assumed that the new sample had observations for all the explanatory variables included in the estimated CLR models as well as the dependent variable (i.e. PSI), for at least one year. The following procedure could be used to estimate PSI using CLR models and the observations for a pavement sample:

1. Estimate $\widehat{PSI}_t^k$ separately for all T observations of a sample, using each of the K estimated CLR models.
2. Calculate the overall sum of squared error (SSE) for each of the models, $\sum_k \Delta PSI_t^k = \sum_k \left(\widehat{PSI}_t^k - PSI_t^k\right)^2$
3. The sample is assigned to the model associated with the least overall SSE.

EXPERIMENT AND RESULTS

Data

Experiments were performed using the Pavement Management System (PMS) of the Nevada Department of Transportation (NDOT). The data included condition monitoring and roadway inventory data collected throughout the entire State of Nevada. 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 for the number of lanes were encoded as:
     - $lane \leq 2$ was equal to '1' if the pavement sample had two or less lanes and zero otherwise, and
     - $lane \geq 3$ was equal to '1' if the pavement sample had three or more lanes and zero otherwise.
   - NDOT classifies pavement samples under
     - The Interstate Route (IR),
     - The National Highway System (NHS), or
     - The Surface Transportation Program (STP).

     Two dummy variables, *nhs* and *stp*, were encoded as: *nhs* was equal to '1' if a segment belonged to the NHS; otherwise, *nhs* was equal to '0'. Similarly, *stp* was equal to '1' if a segment belonged to STP; otherwise, *stp* was equal to '0'.
   - NDOT grouped its roadway network into five prioritization categories, 1- 5, using such factors as facility type and traffic volumes (NDOT, 2011). The type and

frequency of maintenance and rehabilitation (M&R) activities vary among these prioritization categories. Four dummy variables – *category=2*, *category=3*, *category=4*, and *category=5* – were encoded as:

  - *category=2* was equal to ‘1’ if the pavement sample is under Prioritization Category 2, ‘0’ otherwise; and
  - *category=3* was equal to ‘1’ if the pavement sample was under Prioritization Category 3, ‘0’ otherwise.

  The same approach was used for the other three dummy variables.

- Code of Federal Regulations (CFR) Title 23 part 470 mandates state agencies to identify the functional class of roads and streets. NDOT divided its roadway network into seven functional classes: (i) Interstate and Highway, (ii) Other Freeways and Expressway, (iii) Principal Arterial-Other, (iv) Minor Arterial, (v) Major Collector, (vi) Minor Collector, and (vii) Local. Six dummy variables – *f_class=2*, *f_class=3*, *f_class=4*, *f_class=5*, *f_class=6* and *f_class=7* – were encoded as follows:

  - *f_class=2* was equal to ‘1’ if the pavement sample was an Interstate and Highway, ‘0’ otherwise;
  - *f_class=3* was equal to ‘1’ if the pavement sample was classified as Other Freeways and Expressway, ‘0’ otherwise.

  The same approach was used for the other four dummy variables.

A total of 4,138 samples – including 14,638 observations from 2001 to 2010 and 3,005 observations from 2011 and 2012 – were available for model estimation and validation, respectively. Table 1 illustrates a subset of data used in the experiments.

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 if the VIF was greater than 10. In this study, all explanatory variables with VIF > 10 were excluded from the final models. Other estimation parameters that were required were set by using previous experience (Paz et al. 2015a; Paz et al. 2015b) and sensitivity analyses. Table 2 provides the parameter values used in this study.

Experiment results and discussion

Function F in the constraint expressed by (13) was used to determine the maximum number of feasible clusters for the dataset 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 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%.

Coefficients for the variables, *trucks* and *freeze_thaw*, were positive. This is counter-intuitive because a pavement deteriorates faster when it is subjected to heavy trucks and frequent freeze-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 could be because pavements having a larger number of trucks and freeze-thaw cycles often are designed to have stronger pavement structures, and are continuously maintained. This research did not use any explanatory variable that relates PSI to pavement structure. Hence, *trucks* and *freeze_thaw*, which were likely to be positively correlated with missing information, such as pavement structure, may have captured this hidden effect. There could be other reasons for the positive coefficients for *trucks* and *freeze-thaw*; however, this investigation did not find enough evidence to justify these positive trends. Hence, these two variables were excluded from the models, and new model parameters were estimated. The effect of these two variables was discussed in Khadka and Paz (2017a). Future research is recommended to investigate this issue. Table 3 provides the estimated parameters for 6-cluster models.

This study used a 5% significance level. Results showed that seven explanatory variables – *elevation*, *precip*, *min_temp*, *max_temp*, *wet_days*, *nhs*, and *stp* – were not included in any of the final estimated 6-cluster 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 4 shows the binary matrix, *V*, associated with the 6-cluster models estimated in this study. Each '1' indicates a 'significant' variable for a particular cluster; '0' indicates otherwise.

Table 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 sample 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. A positive sign indicated that the pavement samples 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 actual conditions.

Model performance

In CLR, minimizing overall SSE translates the maximization of variations in the dependent variable explained by clustering process and regression models (Brusco et al. 2008). CLR does not differentiate between the variations explained by the clustering process and variations explained by regression models. Hence, in some cases, variation in the dependent variable could be minimized by the clustering process even if variations explained by the regression models are small. This creates a potential for overfitting the data in cases when regression relationships are not strong.

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, and are included in Table 5. The results showed that the between-clusters sum of squares (BCSS), which represent the variations explained by the clustering process, was equal to ‘4’, which is 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). The

WCSS represents the sum of the variations across clusters, which is the sum of SSE and SSR. This indicated that there was no overfitting, as most of the variations in PSI was explained by within-cluster regressions. However, SSE accounted for 53% of the TSS, which indicated that the resultant models had relatively high errors, possibly due to the nature of the data. In addition, the estimated linear function might not have been the best to use in order to explain the pavement deterioration.

The prediction accuracy of the models was evaluated by calculating the RMSE, the normalized root-mean-square error (NRMSE), and the mean absolute error (MAE), using (14), (15), and (16), respectively:

$$RMSE = \sqrt{\sum_{i}^{\eta}(y_{it} - \hat{y}_{it})^2/\eta} \qquad (14)$$

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \qquad (15)$$

$$MAE = \frac{1}{\eta}\sum_{1}^{\eta}|y_{it} - \hat{y}_{it}| \qquad (16)$$

where y_{it} = the observed PSI, $\hat{y}_{it}$ = 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 dataset described in the Data section 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 6 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.

Figure 4a shows a scattered plot of predicted versus observed PSIs for 2011 and 2012. The degree of prediction error of the models was reflected by the relative positions of the data points from the 45^0 line. Data points above the 45^0 line were over-predicted, while those under the 45^0 line were under-predicted. Results indicated that the predicted PSIs ranged from 2.70 to 4.42, whereas the observed PSIs ranged from 1.64 to about 4.44. In particular, the CLR models overestimated PSIs that were at the lower end of the data. Possible reasons for overestimation could be that this study did not include any explanatory variables that captured the pavement structure. In addition, improvements by routine maintenance activities were ignored.

Figure 4b provides the percentages of observations that were within different ranges of error. For example, about 74% of the total number of predictions were contained within a ±15% range of error. Figure 5 shows individual scattered plots of predicted versus observed PSIs for all six models.

CONCLUSIONS AND RECOMMENDATIONS

In this paper, 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-subset regression was implemented. The algorithm has the capability to identify potential explanatory variables that cause serious multicollinearity in a model.

VIF was used to measure the effect of multicollinearity in a model. In this study, multicollinearity was addressed using a traditional approach where correlated variables were removed one at a time until the effect of multicollinearity became minimal. However, a better way to address multicollinearity is to consider the trade-off between removing and keeping potential explanatory variables that are expected to cause multicollinearity. Future research is recommended to integrate such an experiment in the CLR framework.

After addressing the multicollinearity issue, the proposed algorithm identified the relevant explanatory variables to be included in the models. All possible combinations of the explanatory variables were evaluated to select the best model for each cluster. Hence, the estimated CLR models included cluster-specific significant explanatory variables that were free from multicollinearity.

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 the significant explanatory variables to be 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. 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 was very effective when selecting the explanatory variables that were relevant.

The estimated CLR models first were analyzed to investigate the presence of overfitting, and the results showed that the models did not possess any overfitting issues. To investigate the predictive capability of the 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 estimated models had small estimation errors. In addition, the results showed that both the differences between the RMSE and MAE values for all six models were approximately equal, which indicated that the prediction error was well distributed among the models. Even so, the models still were associated with prediction errors.

The linear functional form used in this study did not exactly fit the data used in the experiments. Hence, it would be worth investigating the proposed methodology by using nonlinear relationships between 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. Further investigation would be required to identify any dominant explanatory variables of the pavement samples that define a cluster.

This study also proposed a heuristic to assign cluster membership to a pavement sample. It was assumed that a new sample had observations for all explanatory variables included in the estimated CLR models as well as the dependent variable (i.e. PSI). Future research is recommended to develop a procedure to assign cluster membership to pavement samples that were not included during model estimation and lack data about the dependent variable.

The proposed algorithm was designed to search for a global minimum; however, a large amount of computational time is required. Another avenue for future research would be to develop faster and more efficient combinatorial algorithms that would guarantee global optimality.

ACKNOWLEDGMENTS

Many thanks to the Nevada Department of Transportation for providing the data that was used for analysis, and to our Technical Writer at UNLV's Howard R. Hughes College of Engineering, Mrs. Julie Longo, for her help reviewing this manuscript. Special thanks to Mr. Fabien Sortais for his help with the implementation of the solution algorithm. The authors would like to thank all the anonymous reviewers and editor for their comments that contributed to improving the final version of the paper.

NOTATION

The following symbols are used in this paper:

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 i, $t \in T_i$;

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}= 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}= 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 cluster, $k \in K$;

K_{max} = Maximum number of potential clusters that could be formed using the given data;

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 $\forall$ $j = 0, \ldots, J, k \in K$;

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 $\forall$ $j = 0, \ldots, J, k \in K$;

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.

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.

Attoh-Okine, N. and Adarkwa, O. (2013). “Pavement Condition Surveys–Overview of Current Practices.” 〈www.sites.udel.edu/dct/files/2013/10/Rpt-245-Pavement-Condition-Okine-DCTR422232-1pzk0uz.pdf〉 (January 2016).

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.

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.

Cěrny, V. (1985). “A Thermodynamical Approach to the Travelling Salesman Problem; An Efficient Simulation Algorithm.” *J. of Optimization Theory and Applic.* 45, 41-55.

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.

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.

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

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.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J. and Münkemüller, T. (2013). “Collinearity: a review of methods to deal with it and a simulation study evaluating their performance.” *Ecography*, 36(1), 27-46.

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

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

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.

Geweke, J. F., and Meese, R. A. (1981). “Estimating Regression Models of Finite but Unknown Order.” *International Economics Review*, 22, 55–70.

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.

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.

Harrell, F. E. (2001). *Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis (Second Ed.)*, Springer, Cham, Heidelberg, New York Dordrecht, London.

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.

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.

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.

Joseph B. Kadane, J. B. and Lazar, N. A. (2004). “Methods and Criteria for Model Selection.” *Journal of the American Statistical Association*, 99(465) 279-290.

Katz, R. W. (1981). “On Some Criteria for Estimating the Order of a Markov Chain.” *Technometrics*, 23, 243–249.

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. (2017a). “Limitations of Existing Pavement Performance Models and a Potential Solution.” *World Conference on Pavement and Asset Management*, Baveno, Italy.

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

Khadka, M., Paz, A., Cristian, A., and Hale, D. K. (2017). “Simultaneous Generation of Optimum Pavement Clusters and Associated Performance Models.” Manuscript submitted for publication in *Transportmetrica A: Transport Science*. Under review.

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.

Kim, Y., Kwon, S., and Choi, H. (2012). “Consistent Model Selection Criteria on High Dimensions.” *Journal of Machine Learning Research*, 13(1), 1037-1057.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P. (1983). “Optimization by Simulated Annealing.” *Science, New Series*, 220(4598), 671-680.

Koehler, A. B. and Murphree, E. S. (1988). “A Comparison of the Akaike and Schwarz Criteria for Selecting Model Order.” *Applied Statistics*, 37, 187-19.

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.

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. J. (2006). “Pavement Condition Prediction Using Clusterwise Regression.” *Transportation Research Record: Journal of the Transportation Research Board*, No. 1974, 70–77.

Luo, Z. and Yin, H. (2008). Probabilistic Analysis of Pavement Distress Ratings with the Clusterwise Regression Method, *Transportation Research Record: Journal of the Transportation Research Board*, No. 2084, 38–46.

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

Mansfield, E. R. and Helms, B. P. (1982). “Detecting Multicollinearity.” *The American Statistician*, 36(3a), 158-160. DOI: 10.1080/00031305.1982.10482818.

Marquardt, D. W. (1970). “Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation.” *Technometrics*, 12(3), 591–612.

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.

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.

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

Nevada Department of Transportation (2011). *Pavement Management System Overview*, Material Division, Nevada Department of Transportation, 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.

Ohlemuller, R., Anderson, B. J., Araujo, M. B., Butchart, S. H., Kudrna, O., Ridgely, R. S., and Thomas. C. D. (2008). “The coincidence of climactic and species rarity: high risk to small range species from climate change.” *Biology Letters*, 4, 568–572.

Park, Y. W., Jiang, Y., Klabjan, D. and Williams, L. (2015). “Algorithms for Generalized Cluster-wise Linear Regression.” 〈http://www.dynresmanagement.com/publications〉 (January 2016).

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.

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.

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.

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: Journal of the Transportation Research Board*, No. 1524, 118–129.

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.

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

Shoukry, S. N., Martinelli, D. R. and Reigle, J. A. (1997). "Universal Pavement Distress Evaluator Based on Fuzzy Sets." *Transportation Research Record: Journal of the Transportation Research Board*, No. 1592, 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

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

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.

Uzoma, U. and Jeremiah, O. O. (2016). “An Alternative Approach to AIC and Mallow’s Cp Statistic-Based Relative Influence Measures (RIMS) in Regression Variable Selection.” *Open Journal of Statistics*, 6(1), 70-75, DOI: 10.4236/ojs.2016.61009

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.

Washington, S. P., Karlaftis, M. G., and Mannering, F. L. (2011). *Statistical and Econometric Methods for Transportation Data Analysis (Second Ed.)*, Chapman Hall/CRC, Boca Raton, FL.

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

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.

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.

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.

Table 1. A Subset of Data used in the Experiments

Sample ID	1				2			
Year	2001	2002	2003	2004	2001	2002	2003	2004
psi	3.82	3.73	3.71	3.62	3.96	3.88	3.86	3.53
age	0	1	2	3	0	1	2	3
adt	725	825	950	950	725	825	950	950
trucks	20	20	19	20	20	20	19	20
elevation	4750	4750	4750	4750	4750	4750	4750	4750
precip	8.25	8.25	6.65	6.65	6.65	6.65	6.65	6.65
min_temp	33	33	36	36	36	36	36	36
max_temp	65	65	67	67	67	67	67	67
wet_days	45	45	41	41	41	41	41	41
freeze_thaw	176	176	154	154	154	154	154	154
rut_depth	0.09	0.08	0.08	0.09	0.01	0.01	0.01	0.01
lane≤2	1	1	1	1	1	1	1	1
lane≥3	0	0	0	0	0	0	0	0
nhs	0	0	0	0	0	0	0	0
stp	1	1	1	1	1	1	1	1
f_class=2	0	0	0	0	0	0	0	0
f_class=3	0	0	0	0	0	0	0	0
f_class=4	0	0	0	0	0	0	0	0
f_class=5	1	1	1	1	1	1	1	1
f_class=6	0	0	0	0	0	0	0	0
f_class=7	0	0	0	0	0	0	0	0
Category=2	0	0	0	0	0	0	0	0
Category=3	1	0	0	0	0	0	0	0
Category=4	0	1	1	1	0	1	1	1
Category=5	0	0	0	0	0	0	0	0

Table 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

Table 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	-1.293	1.1	< 0.0001	-2.814	1.1	< 0.0001	-3.338	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	-3.706	1.1	< 0.0001	-3.289	1.1	< 0.0001	-2.291	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		

Note: *†* = variable value in thousands, and - = Not applicable

Table 4. Binary Matrix Showing Significance of the Explanatory Variables in the Estimated 6-Cluster Models

	Cluster					
Explanatory Variables	*k* = 1	*k* = 2	*k* = 3	*k* =4	*k* =5	*k* = 6
intercept	1	1	1	1	1	1
age	1	1	1	1	1	1
adt	1	1	1	1	1	1
elevation	0	0	0	0	0	0
precip	0	0	0	0	0	0
min_temp	0	0	0	0	0	0
max_temp	0	0	0	0	0	0
wet_days	0	0	0	0	0	0
rut_depth	1	1	1	1	1	1
lane≤2	0	1	1	1	1	0
lane≥3	0	1	1	1	1	0
nhs	0	0	0	0	0	0
stp	0	0	0	0	0	0
f_class=2	1	0	0	0	0	1
f_class=3	1	0	0	0	0	1
f_class=4	1	0	0	0	0	1
f_class=5	1	0	0	0	0	1
f_class=6	1	0	0	0	0	1
f_class=7	1	0	0	0	0	1
Category=2	0	1	1	1	1	0
Category=3	0	1	1	1	1	0
Category=4	0	1	1	1	1	0
Category=5	0	1	1	1	1	0

Table 5. 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

Table 6. RMSE, NRMSE, and MAE for Each Cluster

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