




Research Article

Simultaneous Generation of Optimum Pavement Clusters and Associated Performance Models

Mukesh Khadka ¹, Alexander Paz ², Cristian Arteaga ¹ and David K. Hale³

¹Civil and Environmental Engineering and Construction, University of Nevada, Las Vegas, 4505 Maryland Parkway, P.O. Box 454007, Las Vegas, NV 89154-4007, USA

²Professor and TMR Chair, Civil Engineering and Built Environment, Queensland University of Technology, 2 George St, Brisbane, QLD 4000, Australia

³Transportation Project Manager, Transportation Solutions Division, USA

Correspondence should be addressed to Alexander Paz; paz.alexander@gmail.com

Received 7 June 2018; Accepted 26 November 2018; Published 12 December 2018

Academic Editor: A. M. Bastos Pereira

Copyright © 2018 Mukesh Khadka et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With regard to developing pavement performance models (PPMs), the existing state-of-the-art proposes Clusterwise Linear Regression (CLR) to determine the pavement clusters and associated PPMs simultaneously. However, the approach does not determine optimal clustering to minimize error; that is, the number of clusters and explanatory variables are prespecified to determine the corresponding coefficients of the PPMs. In addition, existing formulations do not address issues associated with overfitting as there is no limit to include parameters in the model. In order to address this limitation, this paper proposes a mathematical program within the CLR approach to determine simultaneously (1) an optimal number of clusters, (2) assignment of segments into clusters, and (3) regression coefficients for all prespecified explanatory variables required to minimize the estimation error. The Bayesian Information Criteria is proposed to limit the number of optimal clusters. A simulated annealing coupled with ordinary least squares was used to solve the mathematical program.

1. Introduction

Typically, pavement performance models (PPMs) are developed using a two-step approach. First, pavement segments with similar characteristics are grouped into clusters using a few critical factors, such as pavement type, age, and traffic volume. Then, PPMs for each of the clusters are developed using statistical techniques. The objective of clustering is to group the pavement segments that perform similarly over time. However, in practice, the performances of pavement segments within a cluster differ significantly because clusters are formed using only a few critical factors [1, 2]. A major challenge is to select characteristics that define clusters and the corresponding segments associated with them [3]. If inappropriate characteristics are used, clusters may include homogeneous segments with different performance behavior or heterogeneous segments with similar performance behavior [4]. 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 [1].

Figures 1(a) and 1(b) 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 the 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 to 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 1(c) illustrates that segments “SR445 (SB), MP: 40-39” and “SR156 (WB), MP: 3-2” had homogeneous performance behavior. Similarly, segments “SR892 (SB), MP: 25-24” and “SR445 (NB), MP: 36-37” showed a consistent behavior (see Figure 1(d)). This suggests that factors other than the Prioritization Category are critical in causing the differences in performance behavior. Influencing factors

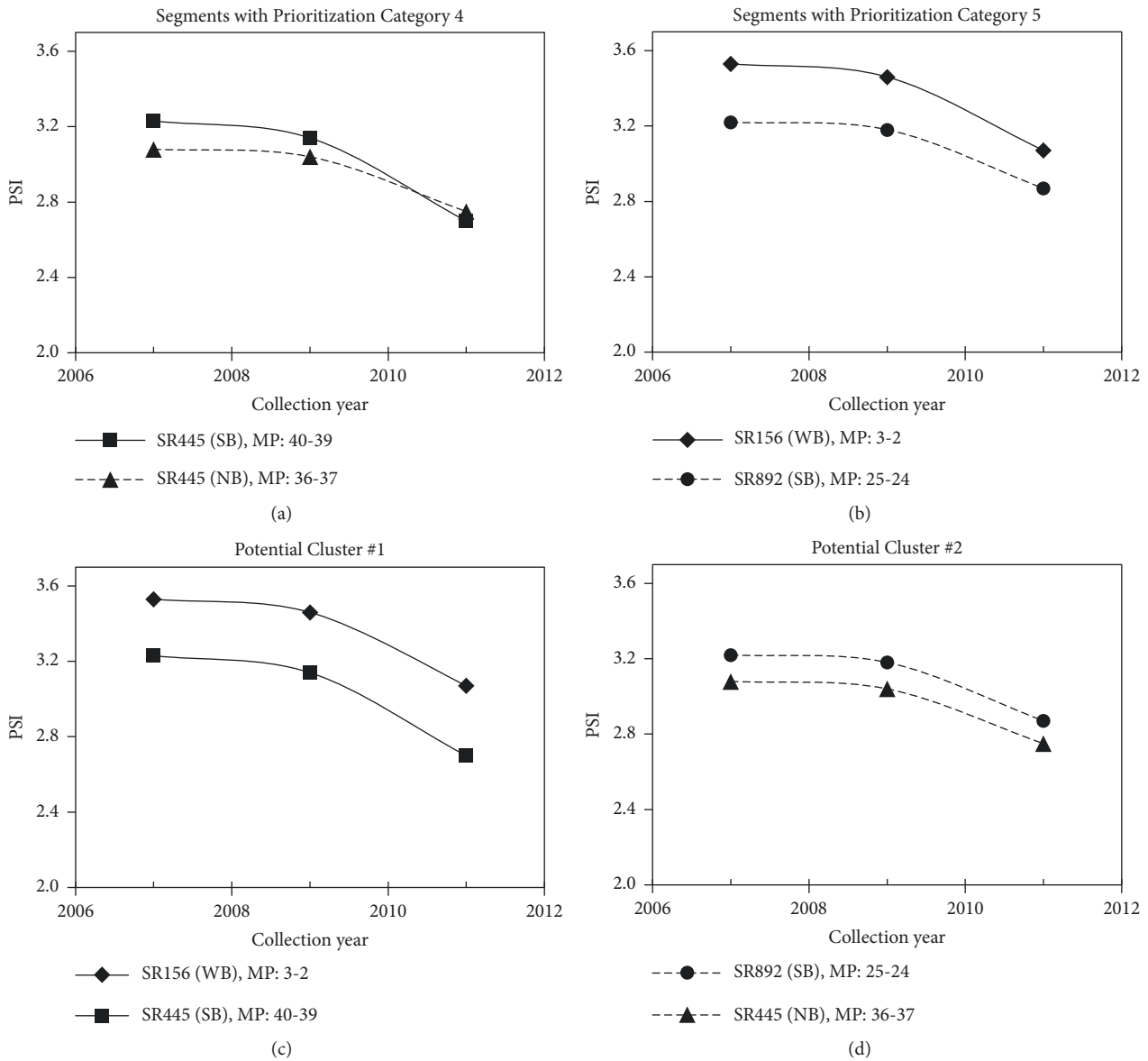


FIGURE 1: (a) and (b) Heterogeneous performance behavior of pavement segments from the same Prioritization Category; (c) and (d) potential clusters with pavement segments having homogeneous performance behavior. NB: north bound; SB: south bound; WB: west bound; MP: milepost.

could include subgrade type, traffic loading characteristics, or any hidden effects.

To address the limitations of the two-step approach, Spath [5] proposed using Clusterwise Linear Regression (CLR) to determine clusters and associated regression models simultaneously. CLR assigns pavement segments that have similar regression effects on the pavement performance such that the overall sum of squared errors is minimal. Hence, CLR minimizes the overall prediction error by simultaneously determining the explanatory variables' coefficients and assigning each pavement segment into an appropriate cluster.

In the context of pavement management, CLR first was used by Luo and Chou [1] to model the deterioration of pavement conditions. First, the pavement network was

clustered by using a few critical pavement characteristics. 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 [2] expanded their research, using CLR to formulate the development of pavement distresses. Both studies [1, 2] used pavement age as the only explanatory variable.

To address some of the limitations present in the studies completed by Luo and Chou [1] and Luo and Yin [2], Zhang and Durango-Cohen [6] developed CLR models using

multiple explanatory variables. Data used in the study were collected during the AASHO Road Test [7], conducted in the late 1950s in Ottawa, Illinois. The data were collected from a single site in a relatively controlled environment; clearly, the site characteristics were not representative of all other locations. For example, the data did not represent the varieties of soil and climatic conditions. In addition, construction techniques and materials have changed substantially since this test was conducted. In this study, the number of clusters was determined manually using an “elbow” criterion. Experiments were run only for a number of clusters equal to 1, 5, 10, 15, 20, and 25. One possible reason for not examining all the possible number of clusters might have been large amount of required computational time given that an exchange algorithm was utilized with 100 instances with various initial assignments.

The previous CLR methods used to estimate PPMs do not test the explanatory power of variables in both cluster 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 also accounted for during assignment of pavement segments into clusters and estimation of the regression coefficients. The presence of any insignificant explanatory variables distorts the underlying regression effects of other significant explanatory variables. This may lead to an incorrect assignment of pavement segments into clusters and estimation of PPMs. Khadka and Paz [8] developed PPMs using CLR while testing the significance of the explanatory variables and obtained superior results relative to the case when significance was assumed.

Another key limitation of the existing CLR is the need to prespecify the number of clusters. In order to avoid prespecifying the number of required clusters, this study proposes a mathematical program to simultaneously determine an optimal number of clusters, the assignment of segments into clusters, and regression coefficients for all 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 [1, 2, 6], the objective function was to minimize the sum of squared errors of prediction (SSE). It is intuitive that SSE decreases monotonically as the number of clusters increases. That is, for a given dataset, the optimum number of clusters always is the total number of data points [9]. An optimum number of clusters are 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 [10]. Usually, this leads to an overfitting problem [11]. Therefore, SSE is not the best objective function to use for searching an optimal number of clusters.

Even though SSE decreases as the number of clusters increases, the rate of improvement diminishes significantly after a relative optimal [12] number of clusters known as the elbow point. Increasing the number of clusters beyond

the elbow point provides a very small reduction in SSE. An SSE versus the number of cluster curve might not exhibit an elbow point distinctly in all cases. Hence, it could be very challenging to choose the right number of clusters.

To address these limitations, this study extended the existing CLR framework by simultaneously (i) incorporating the Bayesian Information Criteria (BIC) [13] as the objective function, (ii) finding the optimal number of clusters, and (iii) finding the maximum number of clusters as required for model estimation. The BIC penalizes more for the inclusion of additional parameters relative to the Akaike Information Criteria (AIC) [14]. The BIC selects simpler models than the AIC when the sample size is greater than eight [15]. Hence, the optimal number of clusters found using BIC provides a balance between model complexity and goodness of fit [16]. On the other hand, several studies showed that the number of parameters in a model selected using AIC was overestimated [14, 17–19]. 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” [20–23]. Galimberti et al. [24] used BIC to propose a unified framework for model-based clustering, linear regression, and multiple cluster structure detection.

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 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. Methodology

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

2.1.1. Definition of a Pavement Sample. The condition of a pavement segment improves when intervention occurs by applying an M&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 historical interventions made on a pavement segment. Figure 2 provides a simplified depiction of how consecutive observations of a pavement segment are divided into two pavement samples.

In this study, a pavement sample, instead of pavement segment, was considered as a single entity during cluster

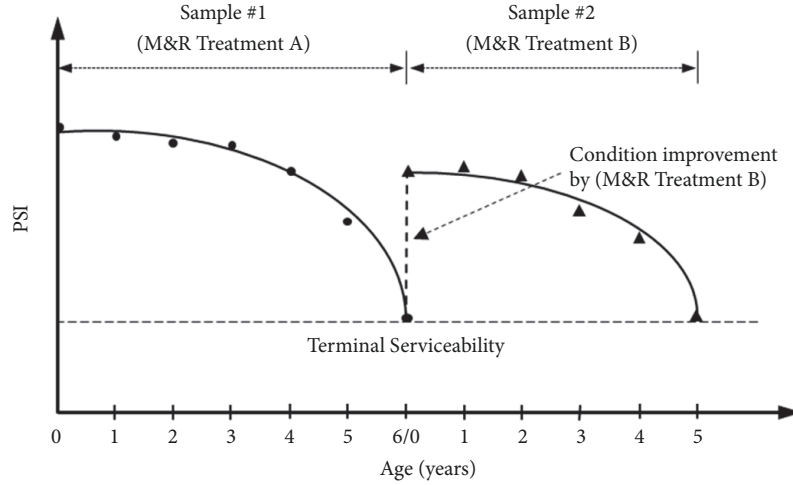


FIGURE 2: A typical pavement performance curve and a simplified depiction of how the observations of a pavement segments are divided into two samples.

analysis. Hence, if a pavement segment consists of two or more pavement samples, these samples could be assigned to different clusters.

2.1.2. Notation and Terms. The following notation and terms are used in describing the proposed model:

- i = Subscript for a pavement sample in the network
- I = Number of pavement samples in the network, indexed $1 \leq i \leq I$
- t = Subscript for an observation period for a pavement sample
- T_i = Number of observation periods for a pavement sample i , indexed $1 \leq t \leq T_i$
- O = Total number of observations = $\sum_i^I T_i \forall i \in I$
- j = Subscript for an explanatory variable
- J = Number of explanatory variables, indexed $1 \leq j \leq J$
- k = Subscript for a cluster
- K_{max} = Maximum number of clusters
- K = Number of optimum clusters
- x_{ijt}^k = Measurement of an explanatory variable j ($1 \leq j \leq J$) for a sample i ($1 \leq i \leq I$) at observation period t ($1 \leq t \leq T_i$) that is assigned to a cluster k ($1 \leq k \leq K_{max}$)
- y_{it}^k = Measurement of dependent variable for a sample i ($1 \leq i \leq I$) at observation period t ($1 \leq t \leq T_i$) that is assigned to a cluster k ($1 \leq k \leq K_{max}$)
- n = Minimum number of observations required in a cluster
- C_k = Set of pavement samples that are assigned to cluster k ($1 \leq k \leq K_{max}$)

p_{ik} = Cluster membership of a pavement sample i to a cluster k , $\forall 1 \leq i \leq I$, and $1 \leq k \leq K_{max}$.

δ_k = Intercept for regression model in cluster $k \forall 1 \leq k \leq K_{max}$

β_{jk} = Slope coefficient for explanatory variable j ($1 \leq j \leq J$) in cluster k ($1 \leq k \leq K_{max}$)

2.1.3. Mathematical Program. A major aspect of model development is to identify variables that explain an actual physical process. Pavement performance models (PPMs) should include explanatory variables that affect pavement deterioration. In this study, explanatory variables from the existing literature were selected. For example, pavement age, traffic and environmental conditions, and structural and material properties of pavement were considered as factors affecting pavement performance [25–28].

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 [29–31]. In addition, PSI reflects the human rider's response and is understood by highway users and legislators [32]. The adopted functional form for the regression model is expressed as

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

This study proposes a mixed-integer, nonlinear mathematical program to determine an optimal number of clusters, assignment of segments into clusters, and regression coefficients for all prespecified 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 partitions 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 as expressed by

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

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 provides the optimal solution. Minimizing BIC reduces unexplained variation in the dependent variable [13]. The total SSE was calculated using

$$\text{SSE} = \sum_{k=1}^K \sum_{i=1}^I \sum_{t=1}^{T_i} \left(\delta_k + \sum_{j=1}^J \beta_{jk} * x_{ijt}^k - y_{it}^k \right)^2 * p_{ik} \quad (3)$$

$\forall i, j, t, k$

Each cluster was associated with a linear regression model with predefined explanatory variables. Deviations of predicted statistics from actual data were calculated separately for each cluster and summed to obtain the total SSE.

Decision variables to be determined were the coefficients for all prespecified explanatory variables, δ_k and β_{jk} ; the optimum number of clusters, K ; and the cluster membership, p_{ik} .

The following constraints were imposed to describe the proposed problem:

$$\sum_k p_{ik} = 1 \quad \forall i \in i, k \quad (4)$$

$$p_{ik} = \begin{cases} 1, & \text{if sample } i \text{ is assigned to cluster } k; \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

$\forall i, k$

$$\sum_{i \in C_k} T_i \geq n \quad \forall C_k \quad (6)$$

$$1 \leq k \leq K_{max} \quad (7)$$

$$K_{max} = F(I, T_i, n) \quad (8)$$

Equations (4) and (5) ensure that each pavement sample was assigned to exactly one cluster. The indicator p_{ik} equals 1 if and only if a pavement sample i belongs to cluster k . Otherwise, it takes a value of zero.

Equation (6) describes a minimum-size constraint, which is imposed to ensure sufficient observations in each cluster for a statistically reliable estimation of coefficients. The total number of observations in a cluster is required to be no less than the minimum number of observations, n .

Equations (7) and (8) are imposed to determine the maximum number of potential clusters. If a pavement sample has more than n observations, regression over these observations could generate statistically reliable estimates of the coefficients. Hence, a cluster could be formed with only

one pavement sample that has more than n observations. If all pavement samples have 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. Equation (8) determines the maximum number of potential clusters in both cases. Function F in this constraint represents the following algorithm to determine K_{max} . A flowchart for this algorithm is provided in Figure 3.

2.1.4. Proposed Algorithm to Determine K_{max}

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, \mathbf{M} , of size $(\tau_{max} \times 2)$ with the following elements, where τ_{max} is maximum number of observations of a pavement segment(s) in the dataset:

- (a) The first column of \mathbf{M} includes all integer values from 1 to τ_{max} .
- (b) The second column includes the number of segments associated with the number of observations.
- (c) If no segments have a particular number of observations in the dataset, 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 \mathbf{M} has all zeros in its second column ($\sum_{\tau} m_{\tau,2} = 0$), then return $K_{max} = K_{max}$ and go to Step 6. Otherwise,

- (a) update \mathbf{M} by removing all rows that have the number of segments equal to zero ($m_{\tau,2} = 0$)
- (b) initialize two indices as $\omega = \vartheta =$ number of rows in \mathbf{M}
- (c) make a copy of \mathbf{M} and let it be represented by \mathbf{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

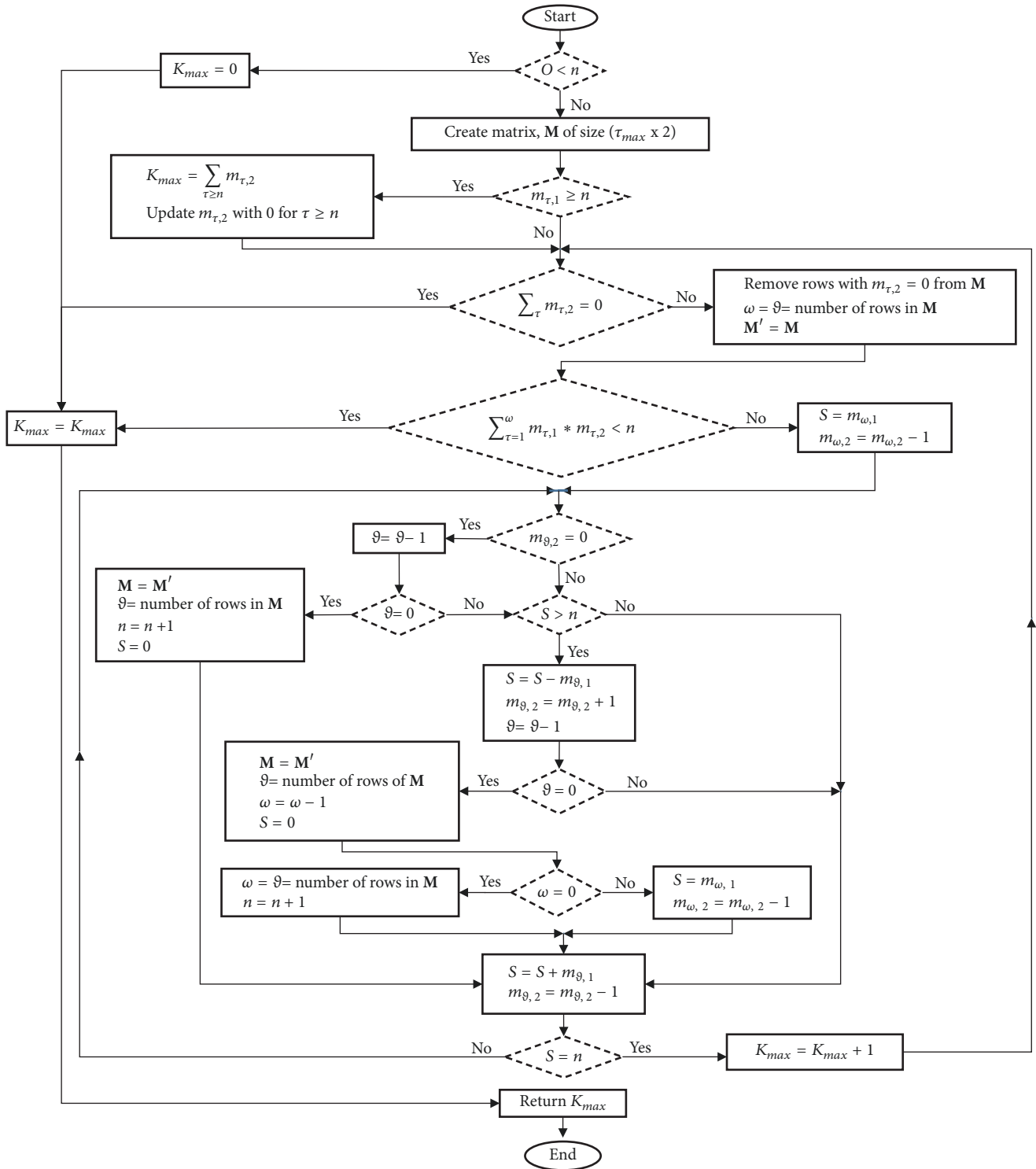


FIGURE 3: Algorithm utilized to calculate the maximum number of potential clusters.

- (i) $\mathbf{M} = \mathbf{M}'$
(ii) set $\vartheta =$ number of rows of \mathbf{M}
(iii) set $n = n + 1$
(iv) set $S = 0$
(v) go to Step 4.6
Otherwise, go to Step 4.3.
Step 4.3. If $(S > n)$, then
(i) $S = S - m_{\vartheta,1}$

(ii) $m_{\vartheta,2} = m_{\vartheta,2} + 1$

(iii) Set $\vartheta = \vartheta - 1$

Otherwise, go to Step 4.6.

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

(i) $\mathbf{M} = \mathbf{M}'$

(ii) set $\vartheta =$ number of rows of \mathbf{M}

(iii) set $\omega = \omega - 1$

(iv) set $S = 0$

Otherwise, go to Step 4.6.

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

(i) update both indices ω and ϑ with the number of rows of \mathbf{M}

(ii) set $n = n + 1$

(iii) go to Step 4.6

Otherwise,

(i) set $S = m_{\omega,1}$

(ii) set $m_{\omega,2} = m_{\omega,2} - 1$

(iii) go to Step 4.6

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

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

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

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 dataset estimate, p_{ik} . For each accepted neighborhood cluster, OLS was utilized to estimate the regression coefficients, δ_k and β_{jk} . The fitting linear models (lm) function, available in the statistical software, R, was used to estimate these coefficients [33]. DeSarbo et al. [34] 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 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. Calculate maximum number of potential clusters, K_{max} , utilizing function F as described above, as part of Constraint 8.

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 δ_k and β_{jk} for all K clusters using OLS.

Step 4. Evaluate objective function, BIC_K^N using (2).

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

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

Step 5.2. For each of the samples 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 a new set of valid neighborhood clusters.

Step 6. Search for a solution:

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

Step 6.2. Evaluate BIC_K^{N+1} using (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_{K+1}^N , and corresponding δ_k 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, $p_{accept} = \exp(-\Delta BIC / (B * T))$, 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:

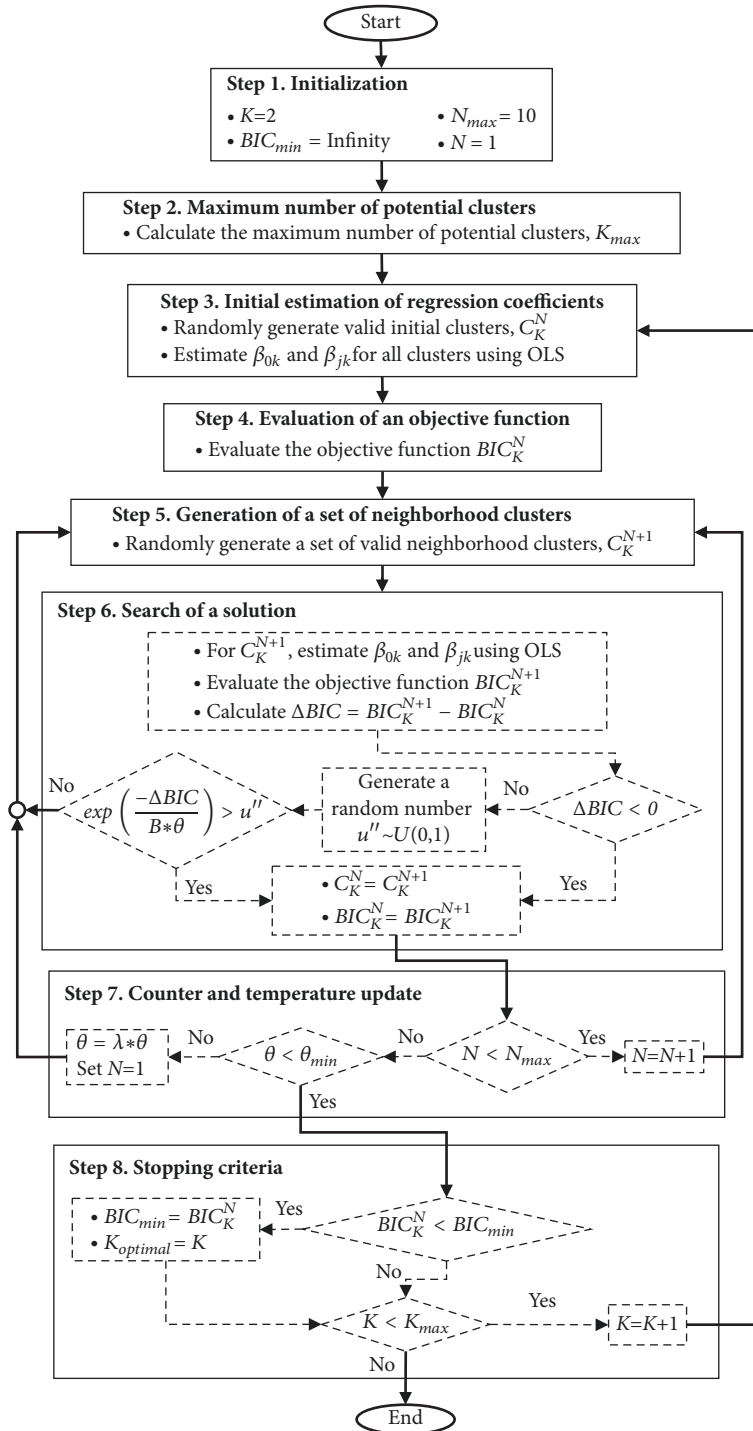


FIGURE 4: Algorithm utilized to solve the clusterwise multiple linear regression.

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

Step 7.2. If $\theta < \theta_{min}$, stop the algorithm. Otherwise, reduce the temperature by multiplying the current temperature by the prespecified 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 [35].

3. Experiment and Results

3.1. Data Resources. Data used in this study were extracted from the PMS database of NDOT [36]. The data consisted of various classes, location data, segment data, contract data, environmental data, traffic data, and pavement condition data, collected for flexible pavement in the entire State of Nevada.

A detailed data analysis was performed to check for inconsistent and missing information in the dataset, and some were found. Some of the missing data were synthesized based on associated information available in the dataset. In preparing the PMS data, the following filters were applied:

- (i) Only one-mile segments were selected for consistency.
- (ii) Only pavement segments with the most recent maintenance contracts awarded in 2001 or later were used in the study.
- (iii) PSI of a pavement should deteriorate over time if no M&R treatment occurred. If PSI of a segment in any year increased by 0.1 or more points from the previous year without any M&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.
- (iv) If the PSI of any year decreased by one or more points from the previous year, all observations for that year were excluded from the analysis.
- (v) 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.
- (vi) Only PSI values used were within the interval of the mean, minus three standard deviations to the mean plus three standard deviations.
- (vii) Pavement samples that did not consist of data regarding conditions for at least two consecutive years were excluded.

- (viii) Data analysis showed that an improvement in 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, collected from 2001 to 2010, were used; the remaining 3,005 observations, collected in 2011 and 2012 (about 17% of the total number), were used as test dataset to check the accuracy of the CLR models. Tables 1 and 2 provide the descriptive statistics of continuous and categorical variables used in this paper, respectively.

3.2. Parameters of the Algorithm. Performance of the SA algorithm generally depends on the values of the 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 [37, 38]. A body of literature exists regarding various methodologies for finding the most appropriate values for annealing parameters in SA [37, 39–43].

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 by Anily and Federgruen [44]. 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 amount of time [39].

Effective values to be assigned to the optimization parameters depend on the type and complexity of the problem. These values may not be obvious to determine, but rather might be determined by trial and error for a given problem [40]. In this study, values assigned to the optimization parameters were determined using experience gained from previous research [45–51] that involved SA and other comparable algorithms. Table 3 lists the parameter values used in this study.

The minimum number of observations required in a cluster, n , is one of the parameters to be defined by the analyst for each dataset and application as it is required for any other statistical analysis regarding the minimum sample size. That is, the proposed methodology and contributions are not restricted or affected by the sample size. The analyst must have available sufficient data for each cluster to be able to obtain reliable estimates. A too small n will result in statistically unreliable models and a potentially time-consuming search process as the maximum number of feasible clusters; K_{max} will be very large. In contrast, a too large n will result in insufficient number of clusters required to provide the optimum goodness of fit. Hence, sensitivity analysis is required to achieve balance between reliable estimates and K_{max} .

3.3. 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

TABLE 1: Descriptive statistics for the continuous variables.

Variable	Minimum	Maximum	Mean	Std. deviation
<i>psi</i>	1.60	4.57	4.01	0.41
<i>age</i>	0.00	8.00	2.24	2.01
<i>adt</i>	20.00	132000.00	4844.45	9812.57
<i>trucks</i>	1.00	7731.00	862.29	1082.20
<i>elevation</i>	228.60	2667.00	1368.25	415.19
<i>precip</i>	3.94	89.28	19.33	10.10
<i>min_temp</i>	-6.67	13.33	3.20	4.00
<i>max_temp</i>	7.78	31.67	20.31	4.22
<i>wet_days</i>	11.00	81.00	42.14	15.67
<i>freeze_thaw</i>	0.00	230.00	136.75	51.51
<i>rut_depth</i>	0.00	1.60	0.14	0.14

TABLE 2: Descriptive statistics for the continuous variables.

Variable	Category	Dummy variable	Number of observations	Percent
System ID	IR	-	4,622	31.6
	NHS	<i>nhs</i>	5,063	34.6
	STP	<i>stp</i>	4,952	33.8
Number of Lanes	1	-	8,450	57.7
	2	<i>lane=2</i>	5,612	38.3
	≥ 3	<i>lane≥ 3</i>	575	3.9
Prioritization Category	1	-	5,004	34.2
	2	<i>category =2</i>	3,181	21.7
	3	<i>category =3</i>	3,118	21.3
	4	<i>category =4</i>	1,558	10.6
	5	<i>category =5</i>	1,776	12.1
Functional Class	1	-	4,695	32.1
	2	<i>f_class = 2</i>	107	0.7
	3	<i>f_class = 3</i>	5,106	34.9
	4	<i>f_class = 4</i>	2,604	17.8
	5	<i>f_class = 5</i>	1,830	12.5
	6	<i>f_class = 6</i>	251	1.7
	7	<i>f_class = 7</i>	44	0.3

Note: IR, interstate route; NHS, national highway system; STP, Surface Transportation Program.

TABLE 3: Setup parameters for implementation of the proposed algorithm.

Parameter	Value	Remarks
θ_0	10	Initial temperature
θ_{min}	10e-17	Final 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

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 4.

Figure 5(a) shows the smallest BIC for each of the clusters ($K = 2$ to 16) considered in this experiment. Figure 5(b) 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 values 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.

Different clusters had different numbers of significant explanatory variables. For example, Cluster 2 had 10 insignificant explanatory variables. 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 not significant in four different clusters.

Trucks play a key role in pavement deterioration because they transfer heavy loads to the pavement [52]. Hence, it is expected that variable “trucks” be significant. There could be various reasons precluding the statistical significance of this variable. Two of them include (i) multicollinearity effects, which are not addressed in the existing CLR literature and motivates the expansion of the framework as recommend in this paper under future research and (ii) different lanes mistakenly used to collect pavement performance and truck traffic data. The data used in this study does not include the lane used to collect the information.

The performance of the proposed CLR approach was compared with that of the existing CLR for pavement management. That is, in this section models estimated using the proposed approach were compared with those estimated using the existing CLR [6] described in the Introduction. Experiments using the existing CLR approach were run for all feasible clusters ($K = 2$ to 16). Figure 5(c) 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 5(c) does not exhibit a clear elbow point. Hence, an optimum number of clusters needed 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 the 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 5(d) shows the trajectory of SSE, when 11-cluster CLR models were used, and Table 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.

The *BIC* for these models are provided in Tables 4 and 5. To compare the goodness of fit, overall *BIC* values were calculated. The overall *BIC* for the seven-cluster models obtained from the proposed approach was 3,008, whereas the *BIC* for 11-cluster models, obtained from the existing approach, was 3,171. This difference was the result of similar or better explanatory power provided by the proposed approach with seven clusters 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* increased.

It is observed that the individual BIC for the 11-cluster are slightly smaller than those for the seven-cluster models. Similar to SSE, the more the clusters are used, the smaller the individual model BIC is. However, this does not necessarily translate into a smaller overall BIC.

3.4. Model Assessment and Validation

3.4.1. Potential Overfitting. Brusco et al. [53] 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. [53] and the test dataset to diagnose overfitting. For the optimum seven-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 within-cluster regressions. SSE accounted for 38% of TSS, which suggests that the models still have a relatively high rate of errors. A nonlinear functional form should be investigated to reduce the existing errors.

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

$$RMSE = \sqrt{\frac{\sum_1^{\eta} (y_{it}^k - \hat{y}_{it}^k)^2}{\eta}} \quad (9)$$

where,

y_{it}^k = the observed PSI

\hat{y}_{it}^k = the predicted PSI

η = the number of predictions

Both models were applied to the test dataset. 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 6. Results indicate

TABLE 4: Estimated model parameters 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)	
	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.	Coeff.	Std. err.
<i>intercept</i>	4.145	0.312	6.242	0.369	2.7910	0.311	6.7280	0.358	7.7810	0.394	12.1400	0.401	3.8730	0.331
<i>age</i>	-0.035	0.002	-0.040	0.003	-0.0350	0.003	-0.0327	0.003	-0.0464	0.003	-0.0392	0.004	-0.0498	0.003
<i>adt</i> [†]	-0.006	<0.001	-0.004	<0.001	-0.0262	<0.001	-0.0028	<0.001	-0.0078	<0.001	-0.0053	<0.001	-0.0334	<0.001
<i>trucks</i> [†]	0.0002 [‡]	<0.001	0.0205 [‡]	<0.001	0.0190 [‡]	<0.001	-0.0151 [‡]	<0.001	0.0306	<0.001	0.0557	<0.001	0.0752	<0.001
<i>elevation</i> [†]	0.0066 [‡]	<0.001	0.0182 [‡]	<0.001	-0.0352 [‡]	<0.001	-0.1079	<0.001	-0.1418	<0.001	-0.0131 [‡]	<0.001	0.1060	<0.001
<i>precip</i>	-0.0037 [‡]	0.008	-0.0118 [‡]	0.009	-0.0037 [‡]	0.008	-0.0248	0.009	0.0094 [‡]	0.011	-0.0518	0.012	-0.0252	0.008
<i>min_temp</i>	-0.030	0.009	0.0129 [‡]	0.010	-0.0092 [‡]	0.009	0.0497	0.010	-0.0321	0.011	-0.0447	0.013	0.0532	0.009
<i>max_temp</i>	0.025	0.007	-0.030	0.008	0.0249	0.007	-0.0568	0.008	-0.0124 [‡]	0.009	-0.0554	0.010	-0.0311	0.007
<i>wet_days</i>	0.005	0.002	-0.010	0.002	0.0115	0.002	0.0031 [‡]	0.002	-0.0028 [‡]	0.002	0.0004 [‡]	0.003	-0.0061	0.002
<i>freeze_thaw</i> [†]	-1.697	<0.001	1.513	<0.001	-0.2029 [‡]	<0.001	1.8550	<0.001	-1.4100	0.001	-13.7900	0.001	4.2370	<0.001
<i>rut_depth</i>	-0.632	0.115	-1.002	0.115	-1.1060	0.095	-0.5614	0.117	-0.8408	0.119	-0.3999	0.143	-0.9900	0.099
<i>lane=2</i>	-0.371	0.029	-0.166	0.027	0.0121 [‡]	0.027	-0.1216	0.028	-0.5574	0.032	-0.2745	0.032	0.0258 [‡]	0.026
<i>lane≥3</i>	-0.325	0.044	-0.195	0.046	-0.1713	0.052	-0.3215	0.045	-0.3974	0.052	-0.2511	0.054	0.0391 [‡]	0.046
<i>nhs</i>	-0.442	0.079	-0.407	0.172	0.7454	0.138	-0.2811	0.072	-0.2639	0.150	-0.4300	0.085	1.4320	0.138
<i>stp</i>	-0.802	0.094	-0.487	0.180	0.6121	0.153	-0.3452	0.091	-0.1346 [‡]	0.131	-0.3379	0.104	1.1350	0.151
<i>f_class=2</i>	0.521	0.115	0.0940 [‡]	0.165	-0.9383	0.136	0.4118	0.091	0.7050	0.140	0.9474	0.115	-1.5090	0.149
<i>f_class=3</i>	0.382	0.100	0.3377 [‡]	0.174	-0.8551	0.139	0.3270	0.080	0.3107	0.148	0.4025	0.101	-1.3770	0.136
<i>f_class=4</i>	0.619	0.112	0.2964 [‡]	0.181	-0.6925	0.151	0.2945	0.094	0.0189 [‡]	0.137	0.5468	0.117	-1.0330	0.146
<i>f_class=5</i>	0.618	0.113	-0.0703 [‡]	0.182	-0.6471	0.152	-0.7763	0.098	-0.6250	0.143	0.3015	0.119	-1.5590	0.148
<i>f_class=6</i>	0.472	0.118	-0.419	0.188	-1.2170	0.159	-0.6566	0.108	-0.7955	0.145	0.0582 [‡]	0.124	-1.7880	0.154
<i>f_class=7</i>	0.554	0.204	0.416	0.199	-1.3840	0.173	-0.1375 [‡]	0.171	NA	0.151	0.7614	0.144	-1.3640	0.169
<i>category=2</i>	-0.306	0.070	-0.0444 [‡]	0.051	0.0762 [‡]	0.056	-0.1300	0.045	-0.6077	0.051	-0.1175 [‡]	0.065	0.0246 [‡]	0.035
<i>category=3</i>	-0.319	0.073	-0.0156 [‡]	0.056	-0.0421 [‡]	0.060	-0.2457	0.051	-0.6000	0.056	-0.2317	0.069	-0.0080 [‡]	0.040
<i>category=4</i>	-0.468	0.077	-0.298	0.064	-0.3665	0.065	-0.1535	0.062	-0.6331	0.066	-0.5050	0.077	-0.3982	0.052
<i>category=5</i>	-0.463	0.077	-0.356	0.064	-0.7165	0.065	-0.1899	0.063	-0.8446	0.067	-0.6001	0.076	-0.4145	0.052
<i>BIC</i>		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.

NA = not applicable.

TABLE 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)									
	Coeff.	Std.Err.	Coeff.	Std.Err.	Coeff.	Std.Err.	Coeff.	Std.Err.	Coeff.	Std.Err.	Coeff.	Std.Err.								
<i>intercept</i>	3.92	0.40	4.95	0.44	14.43	0.60	7.41	0.45	3.70	0.36	4.61	0.40	3.80	0.47	7.95	0.48	6.50	0.42	7.54	0.43
<i>age</i>	-0.04	<0.01	-0.04	<0.01	-0.04	<0.01	-0.05	<0.01	-0.04	<0.01	-0.04	<0.01	-0.05	<0.01	-0.05	<0.01	-0.04	<0.01	-0.03	<0.01
<i>adt</i> [†]	-0.04	<0.01	-0.02	<0.01	0.01	<0.01	-0.04	<0.01	0.00	<0.01	-0.01	<0.01	-0.04	<0.01	-0.01	<0.01	-0.01	<0.01	0.00	<0.01
<i>trucks</i> [†]	0.11	0.02	0.03	0.02	0.13	0.02	0.05	0.02	0.01 [‡]	<0.01	-0.03 [‡]	<0.01	-0.01 [‡]	0.02	0.05	<0.01	0.01 [‡]	0.02	0.03	<0.01
<i>elevation</i> [†]	-0.02 [‡]	<0.01	0.18	<0.01	-0.05 [‡]	<0.01	-0.22	<0.01	0.01 [‡]	<0.01	0.01 [‡]	<0.01	-0.17	<0.01	0.18	<0.01	-0.05 [‡]	<0.01	0.01 [‡]	<0.01
<i>precip</i>	-0.01 [‡]	<0.01	-0.02 [‡]	<0.01	-0.13	<0.01	-0.01 [‡]	<0.01	-0.01 [‡]	<0.01	0.02 [‡]	<0.01	0.04	<0.01	-0.05	<0.01	-0.03	<0.01	-0.05	<0.01
<i>min_temp</i>	0.01 [‡]	<0.01	-0.02 [‡]	<0.01	-0.09	0.02	0.01 [‡]	<0.01	-0.02	<0.01	-0.03	<0.01	-0.02 [‡]	<0.01	0.12	<0.01	-0.01 [‡]	<0.01	0.05	<0.01
<i>max_temp</i>	-0.01 [‡]	<0.01	0.01 [‡]	<0.01	-0.05	<0.01	-0.03	<0.01	0.02	<0.01	0.02 [‡]	<0.01	0.03	<0.01	-0.12	<0.01	-0.01 [‡]	<0.01	-0.06	<0.01
<i>wet_days</i>	0.01 [‡]	<0.01	-0.01 [‡]	<0.01	0.02	<0.01	0.01 [‡]	<0.01	0.01	<0.01	-0.01 [‡]	<0.01	0.01 [‡]	<0.01	-0.02	<0.01	0.01 [‡]	<0.01	0.01	<0.01
<i>freeze_thaw</i> [†]	1.00 [‡]	0.56	-0.86 [‡]	0.50	6.59	0.55	-20.89	0.85	0.24 [‡]	0.59	-1.38	0.44	0.78 [‡]	0.52	1.45	0.59	-2.19	0.59	-1.93	0.59
<i>rut_depth</i>	-0.64	0.15	-1.00	0.14	-1.22	0.13	-1.03	0.17	-0.27	0.13	-0.46	0.13	-0.76	0.13	-1.02	0.13	-1.80	0.15	-0.88	0.14
<i>lane=2</i>	0.02 [‡]	0.04	-0.16	0.03	-0.35	0.04	-0.10	0.04	0.03 [‡]	0.04	-0.06 [‡]	0.03	-0.26	0.03	-0.22	0.03	-0.53	0.04	-0.28	0.04
<i>lane≥3</i>	0.22	0.06	-0.03 [‡]	0.06	-0.43	0.06	-0.31	0.06	0.31	0.08	-0.43	0.06	-0.48	0.05	0.01 [‡]	0.07	-0.33	0.06	-0.27	0.06
<i>nhs</i>	1.39	0.15	0.28 [‡]	0.17	0.51	0.16	-0.12 [‡]	0.16	1.20	0.14	-0.25	0.08	-0.41	0.08	0.80	0.18	-0.70	0.15	-0.50	0.17
<i>stp</i>	1.33	0.17	0.18 [‡]	0.19	0.42	0.17	-0.27 [‡]	0.18	0.67	0.17	-0.38	0.11	-0.50	0.10	0.82	0.19	-0.64	0.15	-0.65	0.19
<i>f_class=2</i>	-1.42	0.23	-0.39 [‡]	0.25	-0.24 [‡]	0.17	0.44	0.18	-0.83	0.24	0.32	0.09	0.52	0.10	-0.93	0.17	1.43	0.20	-0.25 [‡]	0.17
<i>f_class=3</i>	-1.36	0.18	-0.55	0.18	-0.38	0.15	0.12 [‡]	0.18	-0.88	0.15	0.15 [‡]	0.09	0.33	0.12	-0.99	0.17	0.93	0.14	0.43	0.16
<i>f_class=4</i>	-1.25	0.18	-0.49	0.20	-0.39	0.16	0.22 [‡]	0.20	-0.56	0.18	0.27	0.11	0.41	0.13	-1.12	0.18	0.89	0.14	0.69	0.18
<i>f_class=5</i>	-1.85	0.19	-0.51	0.20	-0.91	0.16	0.03 [‡]	0.20	-1.43	0.18	0.21 [‡]	0.11	0.18 [‡]	0.13	-2.18	0.18	0.60	0.14	0.43	0.18
<i>f_class=6</i>	-2.07	0.20	-0.71	0.20	-0.77	0.17	-0.72	0.21	-1.71	0.19	0.14 [‡]	0.12	-0.66	0.14	-2.24	0.19	0.24 [‡]	0.15	0.56	0.19
<i>f_class=7</i>	NA	-0.63	0.22	-1.53	0.25	0.42 [‡]	0.29	-0.68	0.19	0.09 [‡]	0.16	0.18 [‡]	0.16	-2.28	0.25	NA	0.72	0.25	NA	NA
<i>category=2</i>	0.02 [‡]	0.09	0.07 [‡]	0.06	-0.36	0.05	0.22	0.10	-0.27	0.07	0.06 [‡]	0.05	-0.20	0.09	-0.07 [‡]	0.06	-0.66	0.06	-0.13	0.05
<i>category=3</i>	0.02 [‡]	0.10	0.01 [‡]	0.07	-0.43	0.06	0.14 [‡]	0.11	-0.17	0.08	-0.05 [‡]	0.06	-0.29	0.10	-0.08 [‡]	0.06	-0.76	0.07	-0.20	0.06
<i>category=4</i>	-0.24	0.11	-0.41	0.08	-0.69	0.07	0.02 [‡]	0.12	-0.22	0.09	-0.13	0.06	-0.42	0.10	-0.15	0.07	-1.01	0.08	-0.74	0.07
<i>category=5</i>	-0.38	0.11	-0.33	0.07	-0.80	0.07	-0.07 [‡]	0.12	-0.18	0.09	-0.08 [‡]	0.06	-0.38	0.10	-0.18	0.07	-1.15	0.08	-1.29	0.07
<i>BIC</i>	125	97	420	553	314	226	88	103	337	233	276									

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

[†] variable value in thousands.

[‡] coefficient with *p* value > 0.05.

NA = not applicable.

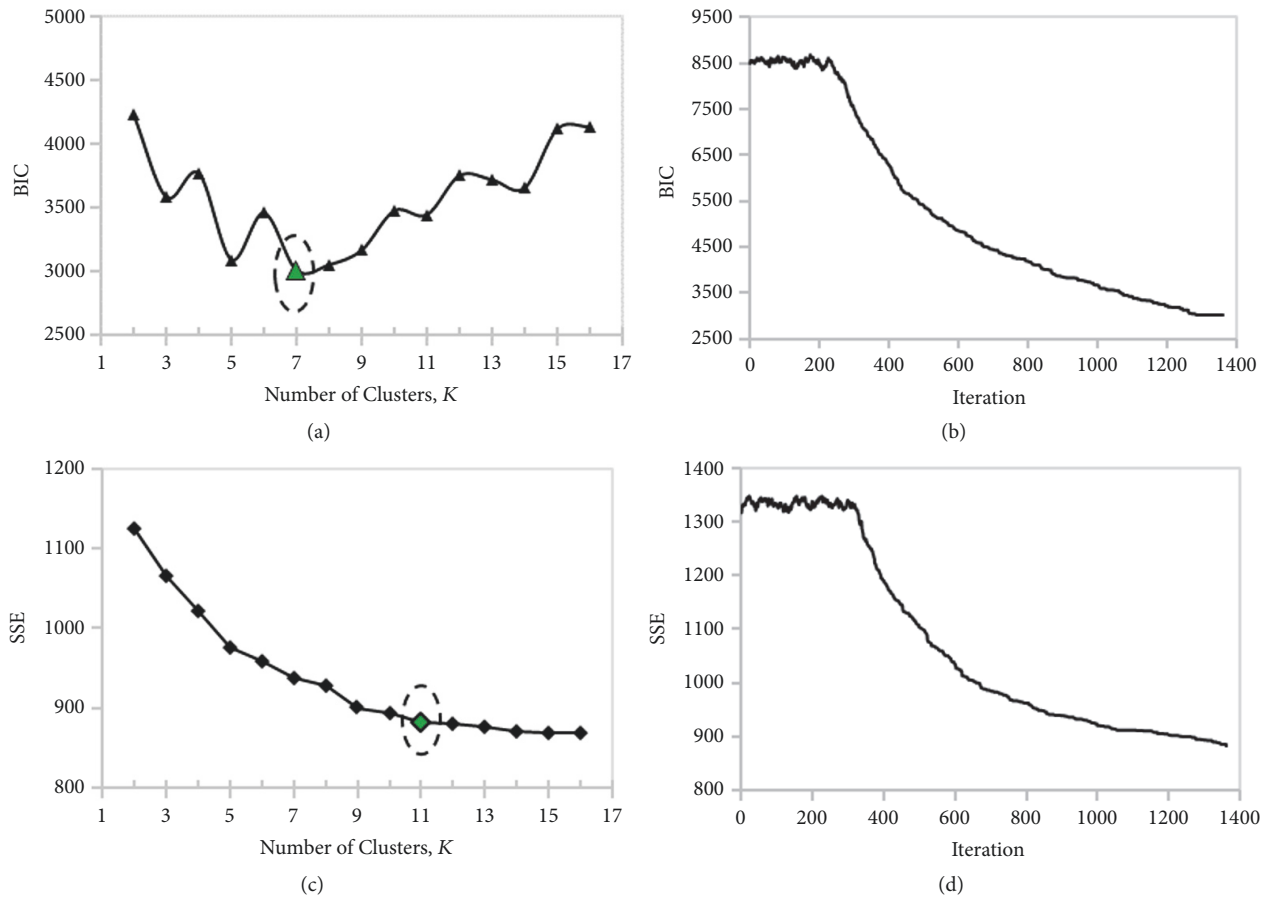


FIGURE 5: BIC trend over the number of clusters (a); trajectory of the BIC during optimization for seven-cluster models (b); SSE trend over the number of clusters (c); and trajectory of the SSE during optimization for eleven-cluster models (d).

that both CLR models overestimated the PSI. A possible reason might be the existence of multicollinearity among explanatory variables.

The RMSE for 2011 and 2012 were calculated for models obtained using both approaches. The RMSE were, respectively, 0.429 and 0.439 for the 11-cluster models estimated using the existing state-of-the-art and the seven-cluster models estimated using the proposed CLR approach. The prediction accuracy of models estimated by the existing state-of-the-art was slightly higher than that of the model estimated by the proposed approach. However, this difference in accuracy is very small. The additional four clusters required 100 parameters corresponding to twenty-five explanatory variables including the intercept. Hence, seven-cluster models estimated using the proposed approach were more parsimonious and preferred over the 11-cluster models.

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, and freeze and thaw cycles. The proposed approach enabled consideration of other types of variables, such as economic and social factors. Formulation of the mathematical program developed in this study supports a number of explanatory variables, multiple observations per pavement segment, and user-defined constraints on cluster characteristics.

A 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 [10, 20–23].

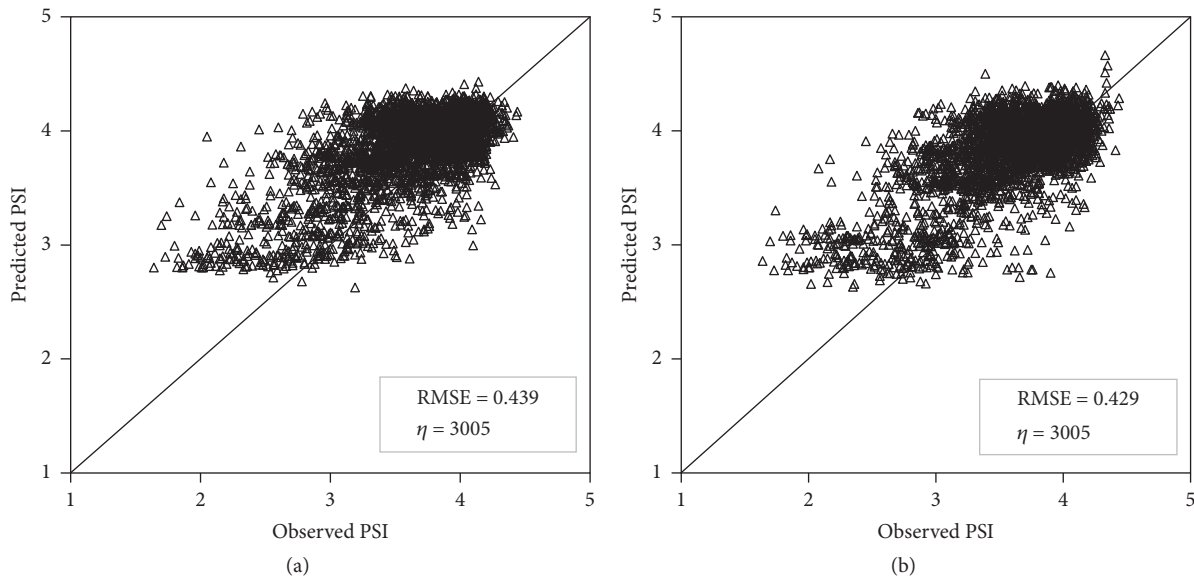


FIGURE 6: PSI predictions using the test dataset with (a) the proposed CLR approach and (b) the state-of-the-art approach.

5. Future Work

This study did not address all limitations of the state-of-the-art CLR approach that were discussed in the Introduction of this paper. The following are potential extensions.

One limitation in this study is related to the error that likely is caused by including insignificant explanatory variables during clustering and regression analyses. The proposed formulation needs to be extended to include only significant variables. Hence, the CLR models would not be restricted only to prespecified explanatory variables. Instead, the models could include cluster-specific significant explanatory variables. In addition, the multicollinearity among explanatory variables should be investigated to exclude highly correlated variables during the model estimation process.

The proposed mathematical formulation was limited to a linear functional form for PPMs. Luo and Chau [1] implemented a CLR approach using an exponential form. However, their study used only pavement age as an explanatory variable. An interesting aspect worthy of investigation would be to explore utilizing the proposed CLR approach by allowing nonlinear relationships between pavement performance measures and multiple explanatory variables.

The proposed SA with the OLS algorithm was designed to search for a global minimum. However, a large amount of computational time was required. Hence, another avenue for future research would be to develop faster and more efficient combinatorial algorithms that could guarantee global optimality.

Data Availability

The Nevada Department of Transportation (NDOT) provided the data that was used for analysis. These data is confidential and should be requested to NDOT.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Many thanks are due 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, Julie Longo, for her help reviewing this manuscript. Special thanks are due to Mr. Antoine Bretecher and Fabien Sortais for their help with the implementation of the solution algorithm.

References

- [1] Z. Luo and E. Y. J. Chou, "Pavement condition prediction using clusterwise regression," *Transportation Research Record*, no. 1974, pp. 70–77, 2006.
- [2] Z. Luo and H. Yin, "Probabilistic analysis of pavement distress ratings with the clusterwise regression method," *Transportation Research Record*, no. 2084, pp. 38–46, 2008.
- [3] M. Steinbach, L. Ertoz, and V. Kumar, "The challenges of clustering high dimensional data," in *New directions in statistical physics*, pp. 273–309, Springer, Berlin, 2004.
- [4] H. Pulugurta, *Development of Pavement Condition Forecasting Models [PhD Dissertation]*, University of Toledo, 2007.
- [5] H. Späth, "Algorithm 39 Clusterwise linear regression," *Computing*, vol. 22, no. 4, pp. 367–373, 1979.
- [6] W. Zhang and P. L. Durango-Cohen, "Explaining Heterogeneity in Pavement Deterioration: Clusterwise Linear Regression Model," *Journal of Infrastructure Systems*, vol. 20, no. 2, p. 04014005, 2014.
- [7] "Highway Research Board, The AASHO road test, Special Rep. No. 61A-E," Tech. Rep., National Academy of Science, National Research Council, Washington, 1962.

- [8] M. Khadka and A. Paz, "Comprehensive Clusterwise Linear Regression for Pavement Management Systems," *Journal of Transportation Engineering, Part B: Pavements*, vol. 143, no. 4, 2017.
- [9] T. M. Kodinariya and P. R. Makwana, "Review on determining number of cluster in kmeans clustering," *International Journal of Advanced Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90–95, 2013.
- [10] J. M. Wooldridge, *Introductory econometrics: A modern approach*, Mason, OH: Thomson/South-Western, 2006.
- [11] J. Wu, *Model-based clustering and model selection for binned data [PhD Dissertation]*, SUPELEC, 2014.
- [12] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 63, no. 2, pp. 411–423, 2001.
- [13] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- [14] J. B. Kadane and N. A. Lazar, "Methods and criteria for model selection," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 279–290, 2004.
- [15] M. Forster and E. Sober, "Why likelihood?" in *The nature of scientific evidence*, pp. 153–190, Univ. Chicago Press, Chicago, IL, 2004.
- [16] C. Fraley and A. E. Raftery, "How many clusters? Which clustering methods? Answers via model-based cluster analysis," *Computer Journal*, vol. 41, pp. 578–588, 1998.
- [17] J. Geweke and R. Meese, "Estimating regression models of finite but unknown order," *International Economic Review*, vol. 22, no. 1, pp. 55–70, 1981.
- [18] R. W. Katz, "On some criteria for estimating the order of a Markov chain," *Technometrics. A Journal of Statistics for the Physical, Chemical and Engineering Sciences*, vol. 23, no. 3, pp. 243–249, 1981.
- [19] A. B. Koehler and E. S. Murphree, "A comparison of the Akaike and Schwarz criteria for selecting model order," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 37, no. 2, pp. 187–195, 1988.
- [20] C. R. Rao and Y. H. Wu, "A strongly consistent procedure for model selection in a regression problem," *Biometrika*, vol. 76, no. 2, pp. 369–374, 1989.
- [21] Y. Yang, "Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation," *Biometrika*, vol. 92, no. 4, pp. 937–950, 2005.
- [22] A. Maydeu-Olivares and C. García-Forero, "Goodness-of-fit testing," *International Encyclopedia of Education*, pp. 190–196, 2010.
- [23] S. I. Vrieze, "Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)," *Psychological Methods*, vol. 17, no. 2, pp. 228–243, 2012.
- [24] G. Galimberti, A. Manisi, and G. Soffritti, *A unified framework for model-based clustering, linear regression and multiple cluster structure detection*, 2015, <https://arxiv.org/abs/1510.03245v1>.
- [25] A. J. Hand, P. E. Sebaaly, and J. A. Epps, "Development of performance models based on department of transportation pavement management system data," *Transportation Research Record*, no. 1684, pp. 215–222, 1999.
- [26] F. Hong and J. A. Prozzi, "Estimation of pavement performance deterioration using Bayesian approach," *Journal of Infrastructure Systems*, vol. 12, no. 2, pp. 77–86, 2006.
- [27] S. W. Haider, K. Chatti, N. Buch, R. W. Lyles, A. S. Pulipaka, and D. Gilliland, "Effect of design and site factors on the long-term performance of flexible pavements," *Journal of Performance of Constructed Facilities*, vol. 21, no. 4, pp. 283–292, 2007.
- [28] H. Khraibani, T. Lorino, P. Lepert, and J.-M. Marion, "nonlinear mixed-effects model for the evaluation and prediction of pavement deterioration," *Journal of Transportation Engineering*, vol. 138, no. 2, pp. 149–156, 2011.
- [29] S. N. Shoukry, D. R. Martinelli, and J. A. Reigle, "Universal pavement distress evaluator based on fuzzy sets," *Transportation Research Record*, no. 1592, pp. 180–186, 1997.
- [30] S. Terzi, "Modeling the pavement present serviceability index of flexible highway pavements using data mining," *Journal of Applied Sciences*, vol. 6, no. 1, pp. 193–197, 2006.
- [31] N. Attoh-Okine and S. Mensah, "MEMS Application in Pavement Condition Monitoring-Challenges," in *Proceedings of the 19th International Symposium on Automation and Robotics in Construction*, Washington, DC, USA, September 2002.
- [32] W. R. Hudson, R. Haas, and E. Perrone, "Measures of Pavement Performance must consider the Road User," in *Proceedings of the 9th International Conference on Managing Pavement Assets*, Alexandria, VA, 2015.
- [33] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.
- [34] W. S. DeSarbo, R. L. Oliver, and A. Rangaswamy, "A simulated annealing methodology for clusterwise linear regression," *Psychometrika*, vol. 54, no. 4, pp. 707–736, 1989.
- [35] R. Román-Román, D. Romero, M. A. Rubio, and F. Torres-Ruiz, "Estimating the parameters of a Gompertz-type diffusion process by means of simulated annealing," *Applied Mathematics and Computation*, vol. 218, no. 9, pp. 5121–5131, 2012.
- [36] *Nevada Department of Transportation, Pavement Management System Overview*, Material Division, Nevada Department of Transportation, USA, 2011.
- [37] M.-W. Park and Y.-D. Kim, "A systematic procedure for setting parameters in simulated annealing algorithms," *Computers & Operations Research*, vol. 25, no. 3, pp. 207–217, 1998.
- [38] S. Babajanzade Roshan, M. Behboodi Jooibari, R. Teimouri, G. Asgharzadeh-Ahmadi, M. Falahati-Naghbi, and H. Sohrabpoor, "Optimization of friction stir welding process of AA7075 aluminum alloy to achieve desirable mechanical properties using ANFIS models and simulated annealing algorithm," *The International Journal of Advanced Manufacturing Technology*, vol. 69, no. 5–8, pp. 1803–1818, 2013.
- [39] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.
- [40] N. E. Collins, R. W. Eglese, and B. L. Golden, "Simulated annealing—an annotated bibliography," *American Journal of Mathematical and Management Sciences*, vol. 8, no. 3–4, pp. 209–307, 1988.
- [41] J. Rose, W. Klebsch, and J. Wolf, "Temperature Measurement and Equilibrium Dynamics of Simulated Annealing Placements," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 9, no. 3, pp. 253–259, 1990.
- [42] S. Z. Selim and K. Al-Sultan, "A simulated annealing algorithm for the clustering problem," *Pattern Recognition*, vol. 24, no. 10, pp. 1003–1008, 1991.
- [43] J. Q. Guo and L. Zheng, "A modified simulated annealing algorithm for estimating solute transport parameters in streams

- from tracer experiment data,” *Environmental Modeling and Software*, vol. 20, no. 6, pp. 811–815, 2005.
- [44] S. Anily and A. Federgruen, “Simulated annealing methods with general acceptance probabilities,” *Journal of Applied Probability*, vol. 24, no. 3, pp. 657–667, 1987.
- [45] C. Cobos, C. Daza, C. Martínez et al., “Calibration of microscopic traffic flow simulation models using a memetic algorithm with solis and wets local search chaining (MA-SW-Chains),” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 10022, pp. 365–375, 2016.
- [46] C. Cobos, C. Erazo, J. Luna et al., “Multi-objective memetic algorithm based on NSGA-II and simulated annealing for calibrating CORSIM micro-simulation models of vehicular traffic flow,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 9868, pp. 468–476, 2016.
- [47] M. Khadka and A. Paz, “Limitations of Existing Pavement Performance Models and a Potential Solution,” in *Proceedings of the World Conference on Pavement and Asset Management*, Baveno, Italy, 2017.
- [48] M. Khadka and A. Paz, “Estimation of optimal pavement performance models for highways,” in *Proceedings of the The 10th International Conference on the Bearing Capacity of Roads, Railways and Airfields (BCRRA 2017)*, pp. 1489–1494, Athens, Greece.
- [49] A. Paz, V. Molano, and J. Sanchez-Medina, “Holistic Calibration of Microscopic Traffic Flow Models: Methodology and Real World Application Studies,” in *Engineering and Applied Sciences Optimization*, vol. 38 of *Computational Methods in Applied Sciences*, pp. 33–52, Springer International Publishing, Cham, 2015.
- [50] A. Paz, V. Molano, E. Martinez, C. Gaviria, and C. Arteaga, “Calibration of traffic flow models using a memetic algorithm,” *Transportation Research Part C: Emerging Technologies*, vol. 55, pp. 432–443, 2015.
- [51] M. Khadka, *Generalized CLUsterwise Regression for Simultaneous Estimation of Optimal Pavement CLUsters and PERformance Models*, ProQuest LLC, Ann Arbor, MI, 2017.
- [52] H. K. Salama, K. Chatti, and R. W. Lyles, “Effect of heavy multiple axle trucks on flexible pavement damage using in-service pavement performance data,” *Journal of Transportation Engineering*, vol. 132, no. 10, pp. 763–770, 2006.
- [53] M. J. Brusco, J. D. Cradit, D. Steinley, and G. L. Fox, “Cautionary remarks on the use of clusterwise regression,” *Multivariate Behavioral Research*, vol. 43, no. 1, pp. 29–49, 2008.

