Springer

Chinese Society of Pavement Engineering

International Journal of
Pavement Research and Technology

Journal homepage: www.springer.com/42947

Upgradation of pavement deterioration models for urban roads by non-hierarchical clustering

Rejani V. U.*, Sunitha Velayudhan, Samson Mathew

Department of Civil Engineering, NIT Tiruchirappalli, Tamil Nadu 620015, India

Received 13 April 2020; received in revised form 4 August 2020; accepted 4 August 2020; available online 18 August 2020

Abstract

Pavements, being a significant component of urban infrastructure, their maintenance and rehabilitation to the desired serviceability level is a challenging problem faced by engineers. The development of a reliable pavement deterioration model is essential to devise proper maintenance policies. This exploratory paper presents the development of network-level pavement performance prediction models for the selected arterial and sub-arterial roads of Tiruchirappalli city, India. Road inventory, traffic volume, maintenance history, pavement condition, and roughness data of the study area are collected periodically for seven years. The Pavement Condition Index (PCI) is determined from the data collected through visual evaluation of the type, severity, and amount of pavement distress. Roughometer is deployed to obtain the International Roughness Index. The parameters which influence pavement deterioration vary widely for different roads within the same network. The pavement sections are assembled into three homogeneous clusters using k-means clustering, which is a nonhierarchical clustering algorithm, so that they can be modeled with better acceptability. Pavement performance prediction models are generated for different clusters using multiple linear regression analysis, and comparison is made with that developed for non-clustered data. The error in prediction is found to be less for clustered models. While the pavement sections in cluster 2, when left unmaintained, deteriorates from a PCI value of 100 to 77 in 5 years, those belonging to cluster 3 are found to deteriorate from 100 to 13. The variation in the deterioration process and the significance of clustering pavement sections for efficient pavement maintenance management is established.

Keywords: Urban pavements; Pavement maintenance; Pavement condition index; Pavement deterioration models; Clustering

1. Introduction

The up keeping of road infrastructure requires methodical tactics involving condition evaluation, performance prediction, program optimization, and development of maintenance strategies. In the 1960s, the idea of Pavement Management Systems (PMS) was used for the first time to design the methodical approach to pavement design and management [1]. Developments in associated technologies took place in the 1970s, and the acquired knowledge was recorded in the book "Pavement Management Systems" [2]. Implementing properly designed PMS depends on factors like credible data, rational models for performance prediction, and user-friendly software for data management.

Data pertaining to pavement condition is a vital element of PMS. The data gathered regularly for a sufficient duration facilitates the representation of the present status of the network, selection of appropriate maintenance strategies, and estimation of future conditions of pavement [3]. The localized data required for network and project level pavement management differ considerably. The network-level pavement management mainly requires road inventory data along with pavement condition data. Inventory data represent the comparatively enduring pavement features, whereas the pavement condition data represents the pavement's serviceability [4]. Highway agencies with lack of skilled workforce and budget constraints are forced to limit the in-depth quantitative assessment of existing pavement conditions, and in such cases, a subjective but methodically executed evaluation could be considered as a more feasible alternative. In most PMSs, indices are formulated from the pavement condition data collected. The Pavement Condition Index (PCI), introduced by the U.S. Army Corps of Engineers [5], is a more intricate index. It is a numerical value in the range 0 to 100 calculated from the visual assessment of distresses on a road network. Osorio et al. [6] presented guidelines for evaluation of distress in flexible and cement concrete pavements for urban networks, deploying manual and automated surveys and rating by experts. Equations were derived for formulating condition indices for flexible pavements using the data. Loprencipe and Pantuso [7] presented deduct value curves to be appended to the ASTM D6433 Distress Identification Catalogue to evaluate the surface distress in urban pavements by including distresses such as tree roots and artificial factors like

* Corresponding author
E-mail addresses: rejanivu@yahoo.com (Rejani V. U.); sunitha@nitt.edu (Sunitha V.); sams@nitt.edu (S. Mathew).
Peer review under responsibility of Chinese Society of Pavement Engineering.

ISSN: 1997-1400 DOI: https://doi.org/10.1007/s42947-020-0105-1

catch basins and manholes which were not contemplated in the ASTM catalogue.

The pavement performance prediction model is the basis for making maintenance policies and budget allocation in the PMS. A precise and reliable pavement deterioration model (PDM) is vital to obtain an optimized PMS model. Different techniques have been used by researchers to develop a PDM which can be broadly categorized as deterministic and stochastic approaches. In deterministic models, values of the dependent variables are entirely governed by the parameters of the model and their initial values. On the other hand, the results of stochastic or probabilistic models possess some randomness and are probability distributions instead of a unique value. The deterministic models include mechanistic models, mechanistic-empirical models, and regression models whereas the stochastic pavement performance models consist of probability-based approaches. Deterministic models have the benefit of being amenable to mathematical analysis whereas stochastic models are probabilistic and can accommodate randomness in pavement performance. Normally, a stochastic performance prediction model is expressed by the Markov transition process [8-11]. Transition probabilities may be determined using either past data or engineering judgments.

A statistical performance prediction model was put forward by Prozzi and Madanat [12] based on the AASHO road test and the Minnesota Road Research Project (MnROAD). The deterministic model predicted the roughness corresponding to the pavement thicknesses, frost gradient, and traffic increment. Kırbas and Karasahin [13] developed deterioration models using artificial neural networks (ANN), deterministic regression analysis, and multivariate adaptive regression splines (MARS) for the performance prediction of urban HMA roads. Even though all the three approaches were accurate and had good R-squared values, the ANN model was found to make predictions more precisely than others. Pantuso et al. [14] presented a model which was developed at network-level on a negative binomial regression to calculate pavement deterioration as a function of its age. A comparison of these models was then done with non-linear regression models. The deterioration anticipated by the model combined with the observed values was used to modify the predictions using a linear empirical Bayesian (LEB) approach. Gogoi et al. [15] tried to employ fuzzy logic for determining the maintenance prioritization of Interlocking concrete block pavements (ICBP). The distress density of ICBP in terms of rutting, depression and impaired paver blocks is specified as input in the fuzzy prioritization model and the pavement condition index (PCI) and the maintenance treatment to be applied are obtained as outputs. In case of many of the road agencies, the resource allocation process is significantly compromised due to limited data accessible for developing performance models. Ramachandran et al. [16] proposed a transfer learning approach founded on a boosting algorithm to generate performance prediction models when pavement data availability is limited. The Highway Development and Management "HDM-4" is an effective tool for pavement maintenance management. HDM-4 performance prediction models are deployed, to predict the pavement condition of urban road network, by many researchers. Jorge and Ferreira [17] presented a pavement maintenance optimisation system using HDM4 models for the municipality of Viseu (Portugal). A global deterministic pavement performance prediction model, which instated the flexible pavement design approach by AASHTO, was used by the PMS. Hong and Wang [18] established a probabilistic modeling approach using a non homogenous continuous Markov chain where the transition probability matrix depends on the model parameters influencing the transition intensity as well as time transformation. Hong [19] proposed non-destructive testing (NDT) tools to collect data to generate probabilistic inputs in pavement reliability analysis. The overlay performance with regard to fatigue was explored. Analytical and simulation methods were adopted to find out the service life reliability of overlays over a 20 year analysis period. A Poisson hidden Markov model was suggested by Lethanh et al. [20] to mathematically establish the interlinkage between deterioration processes. A numerical assessment method using Bayesian statistics with a Markov chain Monte Carlo simulation was introduced to obtain the parameters of the model from its past data. A staged-homogenous Markov model was put forward by Abaza [21] for the prediction of pavement performance at the project level. Individual transition probability matrices were used corresponding to each division of the equal staged-time periods. Transition probability matrices were derived by Pérez-Acebo et al. [22] for the roads of the Republic of Moldova deploying IRI values gathered half yearly. These models were suggested for regions in relatively identical positions; a network without any major additions to the road network in recent years, sections with unspecified pavement structure, pavement maintenance carried out at different time periods, and improper pavement condition data available.

The parameters which influence pavement deterioration may vary widely for different roads within the same network. Undue variability in data affects the accurate prediction of pavement deterioration, and hence, the performance of pavement sections can be predicted more accurately if they are grouped into homogeneous clusters. The cluster analysis approach was adopted for planning the rural road network in Karnataka State, India, by Amarnath et al. [23]. The unconnected villages were grouped and prioritized based on their socio economic status using non-hierarchical clustering. A cluster-wise regression method was put forward by Luo and Yin [24], to evaluate pavement distress and predict pavement performance. The data was grouped according to the severity and extent of distress. Sunitha et al. [25] proposed deterioration models for rural roads based on a clustering approach. The pavement sections were grouped according to the pavement condition data. Only the distress factors relevant to low-volume rural roads were considered for clustering and traffic was found to be insignificant for those roads. A cluster-based linear regression model for predicting pavement deterioration was put forward by Zhang and Durango-Cohen [26]. The model segments a population and depicts the performance with a separate regression model for each segment. Li et al. [27] explored the association between the roughness of pavement and emission from vehicles. The pavement roughness values were clustered using three pattern recognition algorithms, on the basis of vehicle emissions, and impacts on public health. An unsupervised cluster approach termed normalized cuts (NCut) was worked out by Wang et al. [28] to assemble pavement sections into homogenous clusters based on the pavement condition data. Khadka et al. [29] proposed a simplified mathematical programme using cluster-wise regression approach for pavement performance modeling. Pavement segments were clustered using some vital factors, like the type, age, and traffic volume of the pavement.

This investigative paper aims to analyse the performance of selected urban road sections grouped into different homogeneous clusters, develop deterioration models for the clusters, and estimate the effects of clustering in enhancing the performance of the models. The performance models discussed in the literature are

mostly location-specific and influenced by the environment, construction techniques, available materials etc. and hence may not portray the exact deterioration process of urban roads in India. Unlike rural roads, pavement data collection in Indian cities is a tedious and time-consuming process due to the heavy traffic, which is not streamlined. The data collection procedures should, therefore, be fast and simple. In most of the models, only one method of data collection is adopted, ie. either objective or subjective measurement. In this study, the data collection methods include both subjective and objective assessments. Distress data is collected by visual condition survey by trained raters, which is cheap and simple while roughness data is collected using Roughometer III. The methods are suitable for local highway agencies of developing counties. Most of the models for Indian roads are developed on the distress data collected for two or three years [23,25], which is not sufficient to forecast the deterioration patterns of pavements accurately. In the present study, the data was procured during the pre-monsoon and post-monsoon seasons over seven years, and hence, anomalies/outliers can be identified easily. The Pavement Condition Index is calculated considering all the pavement distress parameters relevant to the urban roads of Tamil Nadu, India. With periodic maintenance, the rate of deterioration is retarded, and it is also considered as a major factor in the pavement performance prediction models. The predictive model calibrated to the local conditions will be able to forecast a range of values for the expected deterioration for urban roads.

2. Methodology

The pavement condition is evaluated with respect to distresses in the pavement and the serviceability with respect to the roughness of the pavement. A visual pavement condition survey is performed to gather details of every distress in the pavement, and the pavement condition is presented as Pavement Condition Index (PCI). The PCI is determined using Deduct Curve method from U.S. Army Corps of Engineers Technical Report [5]. As the visual evaluation is subjective, an objective assessment of the pavement in terms of the roughness along the wheel path is carried out by Roughometer III, and the results are presented in terms of International Roughness Index (IRI). Traffic volume surveys are also conducted for the study area through manual counts. The road sections are grouped into homogeneous clusters using the PCI, IRI, traffic volume, and age of the pavement. Multiple Linear Regression Models are developed for each cluster to predict pavement deterioration, after determining the significance of preceding PCI and other independent variables. These models are compared with the single model developed for the same roads without clustering. The methodology is pictorially represented in Fig 1.

3. Study area and data collection

The area selected for the present study is situated in Tiruchirappalli (Trichy) district, which is the geographical centre of the state of Tamil Nadu, South India. Traffic volume surveys are conducted for the arterial and sub arterial road network of Trichy city to the east of NH67. After the analysis of traffic volume, the roads carrying heavy commercial traffic are selected for the study as listed in Table1. Some of the roads are four-lane divided roads, while others are two-lane roads. All the data are collected lane-wise for each road. The road inventory, traffic

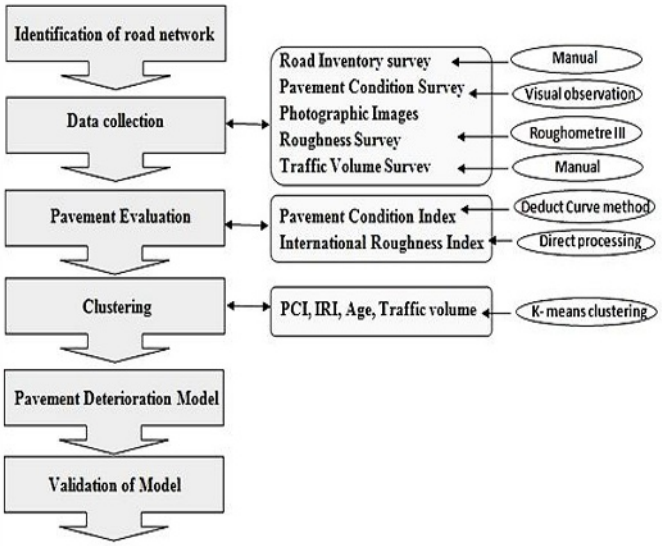

Fig. 1. Methodology.

volume, maintenance history, distress, and roughness data are collected for all the selected roads [30].

3.1. Road inventory

The inventory data is collected during the first season, from which, the length, number of lanes and age of each road is obtained, as given in Table 1. The age is updated using the maintenance data. It is assumed that the age becomes zero, when a major rehabilitation is done on the pavement. So the age shown in the table is the number of years after the last major rehabilitation. For each road starting and ending points are identified and while moving from the starting point to end point, the lanes on left and right sides are taken as left and right lanes respectively. For 4 lane roads, while moving from start to end points, the two lanes on the left side of median are taken as left1 (left most) and left2 and those on the right side of median as right 1 and Right 2 (Right most). Each lane is divided into sections of 25 m length to collect the distress data and the total number of sections including all the lanes is 1488.

3.2. Maintenance history

The pavement maintenance data is collected periodically from the Tiruchirappalli Corporation office. The year of construction, period, and type of maintenance are obtained from the data, and this information is used to assess the age of the pavement. The nature of maintenance work undertaken include pothole filling, shoulder dressing, minor crack sealing, surface treatments, strengthening, etc. The PCI keeps on decreasing every year. After a major rehabilitation (strengthening with Dense Bituminous Macadam + Bituminous Concrete or an overlay of more than 40 mm), the PCI of the road again increases to 100. So it is considered as a newly constructed pavement. The pavement age is assumed to be reduced to zero if a major rehabilitation of the pavement is carried out. Age of the pavement is increased by one every year, if any other treatment or no treatment is done. The age of the pavement, therefore, is the number of years after the last major rehabilitation. Age of the selected roads as on January 2016 is given in Table1.

3.3. Traffic volume

Increased traffic loading is a relevant factor that accelerates the deterioration of urban roads. Traffic volume studies were conducted in 2010 and 2017 for all the 19 roads of the study area to determine the traffic details in Commercial Vehicles per Day (CVPD) which includes vehicles whose laden weight is more than 3 tonnes. A 24-hour manual count was conducted for each road during weekdays, and the CVPD was computed, as given in Table 1. As most of the roads were less than 1 km and there were no major intersection between the start and end points of each road, the variation in traffic volume along the length of road was very less, and hence, the traffic data was collected only at the middle of each road.

3.4. Pavement condition

Pavement distress is a major element in defining the condition of a pavement. Even though roughness, deflection, etc. are also indicators of the pavement condition, conventionally, the term condition survey is used to represent the process of evaluating surface distresses. The type, extent, and severity of distress are generally the aspects considered in distress evaluation [5]. The distress data was collected for thirteen seasons (twice a year, once before and once after the monsoon - from 2010 to 2017) through visual (walk-through) survey. Surveyors were trained with the help of manuals, forms, and photographs, to ensure uniformity and consistency of data collection. Training sessions, data auditing, and follow-up visits to the study roads were done to maximise the accuracy of data. The data was collected from all the lanes separately in both directions. Each lane was divided into sections of 25 m length to collect the distress data which included alligator cracking, transverse cracking, longitudinal cracking, rutting, pothole, depression, patching, and raveling, as per the guidelines of IRC 82-1982 [31]. The code of practice deals with the symptoms, causes and treatments of several types of distresses commonly met with bituminous surfaces and their maintenance planning process. In the present study, the extent of distress is taken as the total distressed area of road in percentage. Severity is noted as low, moderate, or high. The details about the condition of shoulders and longitudinal drains were also collected separately. As most of the drains were closed and the section wise details could not be collected properly, it was not considered for further analysis.

The PCI is determined using Deduct Curve method from U.S. Army Corps of Engineers Technical Report [5]. The Total Deduct Value (TDV) is obtained from the deduct curves for various distress types. The Corrected Deduct Value (CDV) is found out from the Corrected Deduct Value chart. The PCI of the section is determined by using the equation, $PCI = (100 – CDV)$.

PCI values varied from 100 for pavements in very good condition to 0 for some fully damaged road sections. Some of the study roads where proper maintenance is not done had some fully damaged sections with zero PCI due to localized failures. Such data was not deleted as this condition may occur in future also.

3.5. Pavement roughness

Roughness denotes the longitudinal unevenness of a pavement surface, expressed using the International Roughness Index (IRI) which was formulated with the outcomes of the International Road Roughness Test carried out in Brazil in 1982 [32,33]. It is an indicator of the condition of the road as well as its riding comfort. Roughness evaluation has a major role in the network level PMS, as it gives a direct measure of the serviceability of the pavement. In the present study, Roughometer III is used to find the IRI values for all the road segments. Roughness data has been collected once a year for seven years (2011 to 2017). Roughness data was

Table 1
Road network selected for the study.

Sl. No.	Road Name	No. of Lanes	Length (km)	No. of Sections**	Age as on Jan 2016 (years)***	Traffic volume (CVPD)	
						Left lane	Right lane
1	Anna Nagar Main Road	4	1.000	160	2.25	57	128
2	Bharathidasan Road	4	1.300	208	0.08	1694	2191
3	Bishop Road	2	1.050	84	2.08	842	1028
4	Collector Office Road	2	1.675	134	0.92	21	671
5	Convent Road	2	0.525	42	1.08	189	1174
6	Hospital Road	2	0.550	44	1.00	674	508
7	Lawson's Road	2	0.500	40	2.00	895	513
8	Mc Donald's Road	2	0.350	28	0.17	2259	252
9	Pattabiraman Street	2	0.850	68	1.84	34	15
10	Puthur EVR Road	4	0.725	116	3.00	1245	1577
11	Puthur Main Road*	4/2	0.725	84	3.00	788	874
12	Reynold's Road	2	0.375	30	0.17	169	613
13	Rockins Road	2	0.525	42	0.17	2167	1686
14	Royal Road	2	0.625	50	0.17	1086	486
15	Salai Road	2	1.250	100	3.92	1172	830
16	Sastri Road	2	1.025	82	4.33	1181	1034
17	Thillai Nagar Main Road	2	1.050	84	4.33	414	1255
18	Victoria Road	2	0.275	22	3.00	478	478
19	William's Road	2	0.875	70	3.92	1113	175
Total			15.250	1488			

* Puthur Main road is a 4 lane road upto a length of 400m and changes to a two lane road afterwards.
** A single lane of road stretch, 25m long, is mentioned as one section. (For example, a two lane road of one km length has 1000/25*2= 80 sections)
*** The age shown in the table is the number of years after the last major rehabilitation. The age in years is calculated as the age in months divided by 12.

recorded continuously for the entire stretch of road and later while processing the data, length of section was given as 25 m which will give the IRI value for each 25 m section. The wheel path maintained was 0.9 m from the pavement edge, as specified in IRC 81-1997 for pavements of wider than 3.5 m [34]. Roughness values varied from 0 m/km for pavements in very good condition to 12 m/km for some fully damaged road sections with localized failure. IRI was used only for clustering the data and not for modeling.

4. Clustering

Clustering is the process of assorting those elements of a data set which are bound together by certain similarities. K-means clustering is a secure and established unsupervised machine learning algorithm for unlabeled data. It works well with large datasets and is easy to implement and interpret the clustering results. To assort similar elements of a data set together, K-means looks for a fixed number of centroids (center of the cluster) needed in the dataset which is termed as a target number *k*. Each of the data points are attached to one centroid such that the in-cluster sum of squares is kept a minimum.

The standard K-means algorithm is given below:

1. Select a preliminary, even arbitrary partition of objects into k groups.
2. Calculate the centroid (i.e., the mean for all variables) for each cluster.
3. a. Determine the Euclidean distance of each object from the respective centroid.
   b. Assign each data element to its closest centroid.
   c. Compute the new cluster centre for each cluster as the mean value of the elements in it.

Repeat step 3 until the cluster center calculation becomes constant.

The goodness of the partition is measured as the sum of squared distances. For two n-dimensional data elements, $i = (x_{i1}, x_{i2}, ..., x_{in})$ and $j = (x_{j1}, x_{j2}, ..., x_{jn})$, a popular distance function is as follows:

Euclidean Distance,

$$d(ij) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{in} - x_{jn}|^2)} \quad (1)$$

As the various attributes are given in different units, attributes measured on larger scales of measurement may subdue the attributes with a smaller scale when using the Euclidean distance function. To overcome this problem, the data set is normalized so that all the values lie between 0 and 1[35].

In the present study, the pavement sections are grouped into homogeneous clusters, and separate models are proposed to predict future pavement performance. Separate regression models for different clusters give a better fit to the data set than a single equation. Each cluster contains a fraction of the data set that exhibits uniform characteristics.

The 1438 road sections of the study area are clustered into 3 groups using data regarding PCI, IRI, traffic volume, and age of the pavement sections. 50 sections where continuous data could not be taken are excluded from the analysis. The grouping of pavement sections is done considering PCI, IRI, traffic volume, and age of the pavement sections for the year 2016. PCI gives a subjective score of the pavement condition while IRI checks whether the pavement is functionally good. Traffic volume and age vary considerably for different roads in the study area and affect the deterioration pattern. K-means clustering is done using XL-STAT, which is a user-friendly data analysis add-in for Microsoft Excel, altering the number of clusters from 2 to 10. The within-class and between-class variance for various iterations are given in Table 2.

An optimum number of clusters is found satisfying the two conditions:

1. Within class variance should be minimum (keeping clusters as tight as possible) or between class variance should be maximum (maximize the separation between the clusters).
2. Each cluster should contain at least 5% of the entire road sections [25].

The total number of 1438 data points and four attributes are considered. The minimum number of data points required in each cluster is 72. The optimum number of clusters is found to be three by the elbow method [36], as shown in Fig. 2. The optimum values for within-class variance and between class variance are found to be 1320 and 1661, respectively.

The cluster composition is shown in Table 3. The least number of sections in clusters is 289, which is greater than 72 (5% of the total sections).

The properties of 3 clusters in terms of traffic (Commercial Vehicle per Day, which includes vehicles whose laden weight >3T), Age, PCI, and IRI are shown in Table 4.

While analysing the roads in various clusters, it is found that the majority of roads with heavy traffic fall into cluster 2 category, whereas the roads with poor maintenance history, ie. higher age, belong to cluster 3. Cluster 1 consists of roads with moderate traffic and moderate maintenance history. Even though age is the major factor determining the formation of clusters, it can be seen that traffic also has influence on the formation of clusters. For example, there are 150 sections belonging to four roads with same

Table 2
Variance of clusters.

No. of clusters	Within class variance	Between class variance	Minimum no. of sections
1	2981	0	1438
2	1950	1031	573
3	1320	1661	289
4	1139	1842	196
5	997	1983	163
6	885	2096	99
7	804	2178	95
8	746	2235	83
9	668	2313	77

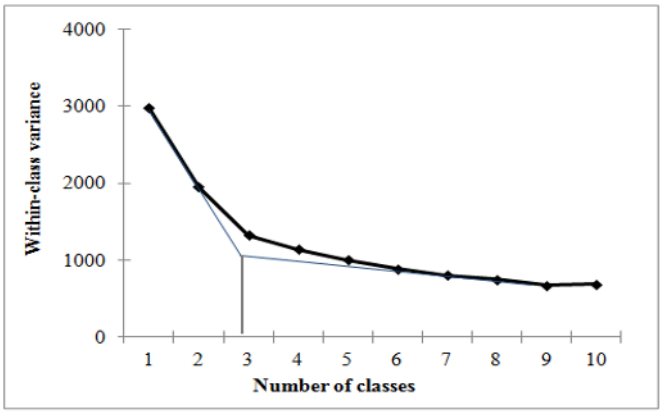

Fig. 2. Number of clusters.

age (0.17 years). 71 sections with low traffic belong to cluster 1 and 79 sections with comparatively heavier traffic belong to cluster 3.

5. Pavement deterioration model

Pavements belong to a group of intricate structures which respond to the various environmental and load related factors in a complex manner. Performance prediction models, to evaluate and predict the impending performance of pavements, are key elements of any PMS. After a review and analysis of several prediction models, the authors concluded that a multiple linear regression model is best suited for the current research study, including a database of the pavement condition data, traffic volume, and age of each road section. Pavement deterioration models are formulated separately for each cluster using the data regarding age and PCI, collected from 2011 to 2016. A single model is also formulated for the entire network using the non-clustered data, to evaluate the influence of clustering. Scheduling of maintenance activities and budget allocation for a stipulated performance level may be done with the help of these models. Though the age of the pavement is the major factor that influenced the formation of clusters, it is found that the clustering resulted in having segments of heavier traffic volume into a separate group.

The correlation between the change in PCI and traffic volume within each cluster is checked. The Pearson's correlation coefficients obtained are 0.079, 0.059 and 0.114 for cluster1, cluster2 and cluster3 respectively which shows a very weak correlation. Hence the traffic volume is not used in modeling the data. Rehabilitations are done on many sections during the analysis period, making the final PCI greater than the initial PCI for that year. Such cases are excluded for that year while modeling. The pavement deterioration model is developed using multiple linear regression analysis with one dependent variable, viz. the condition of the pavement for a year, and two independent variables, viz. pavement condition in the preceding year, and the age of the pavement.

The general structure of the model is,

$$PCI_n = a + b * PCI_{n-1} + c * Age_n \quad (2)$$

where, PCI_n and PCI_{n-1} are the pavement condition indices of year n and n-1 respectively, Age_n = age of the pavement in the n^{th} year, a = constant, b = coefficient of PCI of the pavement in the preceding year and c = coefficient of the age of the pavement section.

The coefficients of (2) obtained for the models for different clusters and that for the non-clustered group are shown in Table 5.

It can be seen that the R^2 value of the model for clustered data is above 0.7, making it a good model. The R^2 value for clustered models is higher than the R^2 value of 0.578 for non-clustered sections. An ANOVA test is carried out on each model and the p values are found to be much lower at a significance level of 0.05 in all cases, indicating the goodness of fit of the models. The test results are shown in Table 6. The t-test was carried out and the results show that the coefficients are significant in all the models.

5.1. Validation of the model

The road condition in terms of PCI for the year 2017 is predicted from that for the year 2016 using the pavement deterioration models developed using clustered data (Clustered model) and non-clustered data (Non-clustered model). The predicted values and the actual measured values are compared for the year 2017.

Graphs are plotted between the predicted and observed values of PCI for each cluster and non clustered model as shown in Figs. 3(a) to 3(d). The figure shows that the model fits well to the data.

The Mean Absolute Percentage Error (MAPE), or the Mean Absolute Percentage Deviation (MAPD), is an indicator of the degree of accuracy of prediction by a model. The Mean Absolute Percentage Error (MAPE) is calculated for all the clusters to assess the differences between the values observed and values predicted by the clustered and non-clustered models, as shown in Table 7.

The Mean Absolute Percentage Error is found to be lesser for the models with clustered data indicating that the prediction is more accurate when the road sections are grouped into homogeneous clusters compared to that without clustering.

5.2. Significance testing

It can be inferred from the above discussions that the clustered and non-clustered models differ from each other. Typically, the clustered model fits better to the data when compared to the non-

Table 3
Composition of clusters.

Class	1	2	3
Class size	601	289	548
Class size (%)	41.79	20.10	38.11
Within-class variance	1407.15	838.27	1478.25
Minimum distance to centroid	13.722	8.824	9.482
Average distance to centroid	35.239	25.105	35.641
Maximum distance to centroid	100.466	93.352	98.245

Table 4
Properties of clusters.

Class		Traffic (CVPD)	Age (Years)	PCI	IRI (m/km)
1	Max	1174	2.25	100	12
	Min	15	0.17	0	0
2	Max	2259	0.17	100	12
	Min	1086	0.08	0	0
3	Max	1577	4.33	100	12
	Min	175	3.00	0	0

Table 5
Coefficients of deterioration models for K-means clusters.

	a	b	c	R^2	Sig.			
					Model	a	b	c
Cluster 1	-17.089	1.116	-0.753	0.715	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Cluster 2	-30.120	1.286	-0.507	0.744	< 0.0001	< 0.0001	0.009	< 0.0001
Cluster 3	-23.502	1.182	-0.742	0.727	< 0.0001	< 0.0001	0.006	< 0.0001
Non-clustered	-31.41	1.220	-0.134	0.578	< 0.0001	< 0.0001	0.039	< 0.0001

Table 6
Results of the ANOVA test.

Cluster	Source	DF	Sum of squares	Mean squares	F	Sig.
1	Regression	2	316150.95	158075.47	1542.76	< 0.0001
	Residual	1230	126029.58	102.46		
	Total	1232	442180.53			
2	Regression	2	144778.66	72389.33	1304.37	< 0.0001
	Residual	896	49725.90	55.50		
	Total	898	194504.56			
3	Regression	2	494807.18	247403.59	2075.43	< 0.0001
	Residual	1560	185961.48	119.21		
	Total	1562	680768.66			
NC	Regression	2	1175865.58	587932.79	2807.77	< 0.0001
	Residual	4099	858308.99	209.40		
	Total	4101	2034174.57			

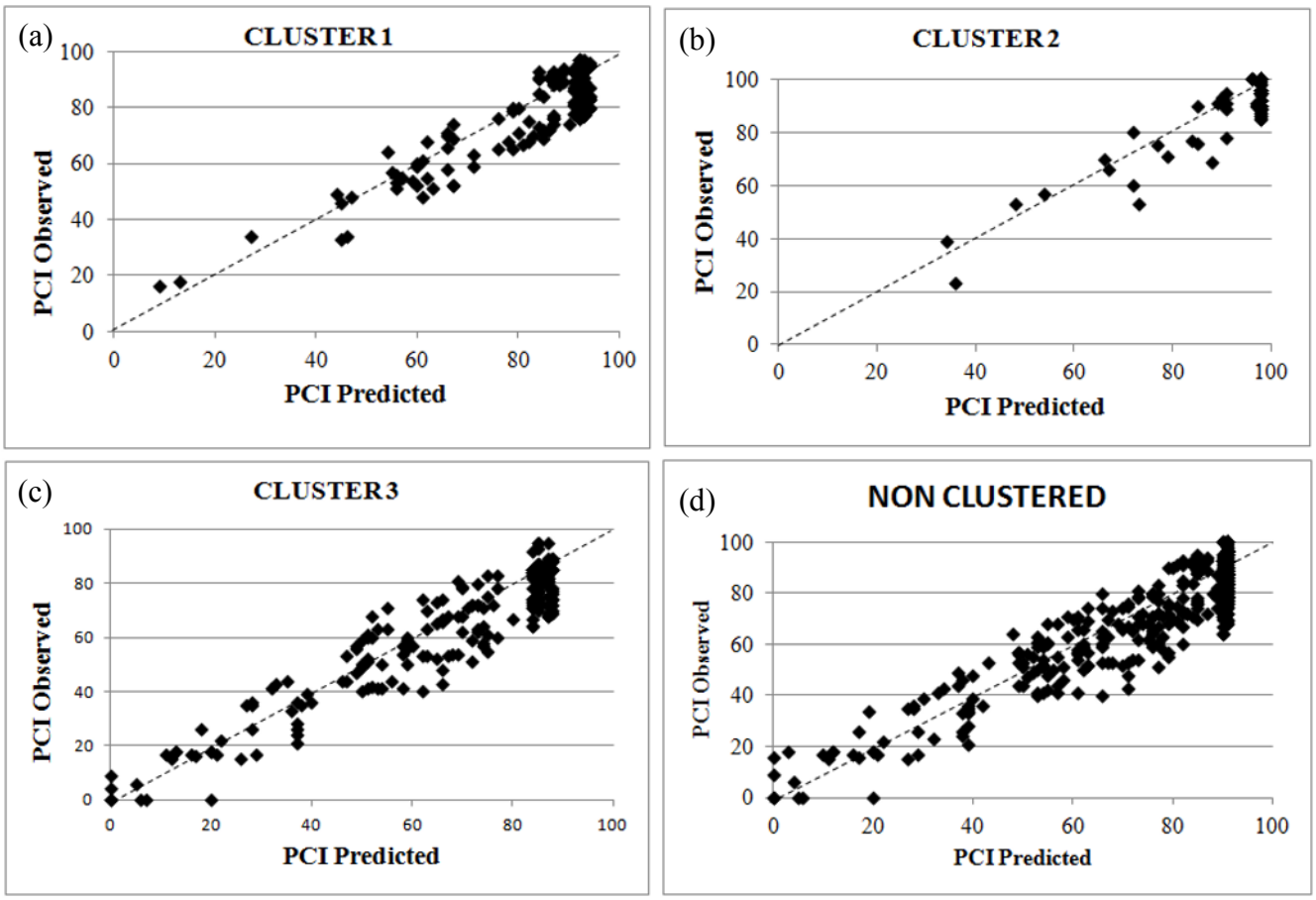

Fig. 3. (a) Observed vs Predicted values - Cluster 1, (b) Observed vs Predicted values - Cluster 2, (c) Observed vs Predicted values - Cluster 3, (d) Observed vs Predicted values - Non Clustered model.

Table 7
Validation of the models.

Class	MAPE	
	Clustered model	Non-clustered model
Cluster 1	6.579	7.222
Cluster 2	5.81	7.712
Cluster 3	9.538	14.000

clustered one. But it has to be determined whether this improvement is significant. It may be done using a partial F-test [37,38].

The F statistic can be calculated by,

$$F = [(RSS_{nc} - RSS_c) / (p_c - p_{nc})] / [RSS_c / (n - p_c)] \quad (3)$$

where, RSS_{nc} is the residual sum of squares of the non-clustered model, RSS_c is the residual sum of squares of the clustered model, p_{nc} is the number of attributes used in the non-clustered model, p_c is the number of attributes used in the clustered model and n is the number of observations. The null hypothesis adopted for the study is that the clustered model does not offer a considerably better fit to the data than the non-clustered model for

an F- distribution with $(p_c-p_{nc}, n-p_c)$ degrees of freedom. The null hypothesis is rejected if the F value obtained for the data is higher than the critical value for some false-rejection probability (e.g., 0.05). RSS obtained for the various clusters, and non-clustered model is given in Table 8.

Here, the calculated F value is 117, and the critical F value, $F_{cri\ (2,\ 790,0.05)}$ is 3.03. F is greater than $F_{cri,}$ which implies that the K - means cluster models are significantly better than the non-clustered ones.

As an example, a comparison of the predicted PCI values for a section having an initial PCI of 100 during an analysis period of 5 years using clustered and non-clustered models is pictorially represented in Fig 4.

It can be seen from Fig. 2 that the prediction of the pavement condition with the non-clustered model gives higher or lower values than those using the clustered models in all the cases. The trend lines conform to the conventional "concave down" shape indicating the slow deterioration during the early life followed by a significant surge in the rate of deterioration. From Fig. 4, it can also be seen that the pavements in cluster 3, which are older compared to others, deteriorate faster. Though the traffic is heavier for cluster 2, the rate of deterioration is slower than the other two clusters. This is because the age of roads in cluster 2 are lesser compared to others, as the maintenance works are carried out regularly on these roads. This is found to be true from the observed data also. It shows that proper and timely maintenance and rehabilitation of urban roads is essential for retarding the deterioration rate and prolonging the life of road infrastructure.

Table 8
RSS obtained for the various clusters and the non-clustered model.

Cluster	No. of independent parameters	RSS
1	4	15478
2	4	5912
3	4	18163
Total		39553
Non - cluster	2	51329

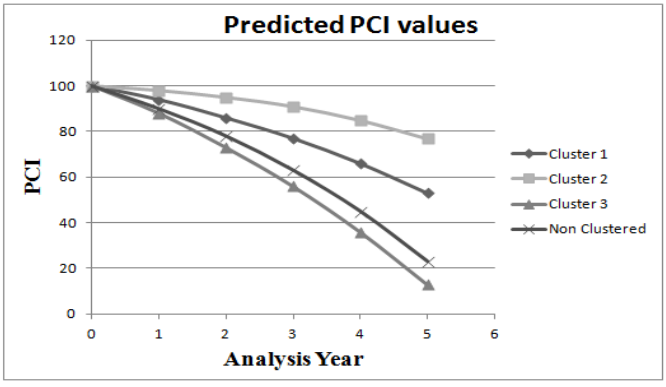

Fig. 4. Prediction of pavement condition using clustered and non-clustered models.

6. Conclusions

Pavement condition survey is done for every 25 m section of all the lanes in the selected roads of the study area, and Pavement Condition Indices (PCIs) are calculated. The International Roughness Index (IRI) values in m/km are measured. The pavement sections are grouped using the method of K-means clustering. The parameters considered are PCI, IRI, traffic volume, and age of the pavement. Three clusters are chosen satisfying the condition of least number of pavement sections in each cluster and the between-class variance of clusters. Separate multiple linear regression models are formulated for each of these clusters to represent the pavement condition deterioration. The model is developed with one dependent variable which is the condition of the pavement for a year, and two independent variables: pavement condition in the preceding year and the age of the pavement. For comparison, a common model is developed for the whole network without clustering. Validations of the models are done using the actual data obtained in 2017 and it is found that the predicted values and the observed values do not show significant variation in any of the cases. The models developed can be used for predicting the maintenance measures for the next year. It is found that the clustered model fits better to the data when compared to the non-clustered one and the improvement is found to be significant. In other words, the prediction is more accurate when the roads are grouped into homogeneous clusters compared to that without clustering. From the study, the necessity and the significance of clustering of pavement sections for maintenance and rehabilitation are established. The vital role of proper and timely maintenance of urban roads in reducing the rate of deterioration is also highlighted.

Future research opportunities exist in grouping the section using other methods of clustering like latent class clustering and comparing the results. Other modeling techniques like HDM-4 and Artificial Neural Networks may also be included for comparison purpose. Also, the maintenance strategies planned according to the predicted PCI needs to be optimized based on the funds available, minimum PCI required after maintenance etc.

Acknowledgement

The authors are thankful to the Centre of Excellence in Urban Transport, Dept. of Civil Engineering, IIT Madras, and Centre of Excellence in Transportation Engineering, Dept. of Civil Engineering, NIT, Trichy for sponsoring this research endeavor.

References

[1] W. R. Hudson, B. F. McCullough, F. N. Finn, K. Nair, B. A. Vallerga, Systems Approach to Pavement Design, Interim Report. NCHRP Project 1-10. Washington DC, USA, 1968.
[2] R. Haas, W. R. Hudson, Pavement Management Systems, McGraw-Hill, New York, USA, 1978.
[3] L. M. Pierce, G. McGovern, K. A. Zimmerman, Practical Guide for Quality Management of Pavement Condition Data Collection. U.S. Department of Transportation Federal Highway Administration, SE Washington DC, USA, 2013.
[4] R. Haas, W. R. Hudson, L. C. Falls, Pavement Asset Management, John Wiley & Sons Inc. New Jersey, Scrivener Publishing, Massachusetts, 2015.
[5] Pavement Maintenance Management, Technical Manual TM 5-623. Department of the Army, Washington DC, USA, 1982.
[6] Osorio A, Chamorro A, Tighe S, Videla C, Calibration and Validation of Condition Indicator for Managing Urban Pavement Networks, Transp. Res. Rec. 2455 (2014) 28–36.
[7] G. Loprencipe, A. Pantuso, A specified procedure for distress identification and assessment for urban road surfaces

based on PCI, Coatings 7 (5) (2017) 65 https://doi.org/10.3390/coatings7050065

[8] N. Li, R. Haas, W.C. Xie, Development of a New Asphalt Pavement Performance Model, Can. J. Civ. Eng. 24 (4) (1997) 547-559.

[9] K. A. Abaza, Empirical Approach for Estimating the Pavement Transition Probabilities used in Non-homogeneous Markov Chains, Int. J. Pavement Eng. 18 (2) (2017) 128-137.

[10] S. P. Soncim, I. C. S. de Oliveira, F. B. Santos, C. A. D. S. Oliveira, Development of Probabilistic Models for Predicting Roughness in Asphalt Pavement, Road Mater. Pavement 19 (6) (2018) 1448-1457.

[11] H. Perez-Acebo, N. Mindra, A. Railean, E. Rojí, Rigid Pavement Performance Models by Means of Markov Chains with Half-Year Step Time, Int. J. Pavement Eng. 20 (7) (2019) 830-843.

[12] J. A. Prozzi, and S. M. Madanat, Development of Pavement Performance Models by Combining Experimental and Field Data, J Infrastruct. Syst. 10 (1) (2004) 9-22.

[13] U. Kırbas, M. Karasahin, Performance Models for Hot Mix Asphalt Pavements in Urban Roads, Constr Build Mater. 116 (2016) 281-288.

[14] A. Pantuso, G. W. Flintsch, S. W. Katicha, G. Loprencipe, Development of Network-level Pavement Deterioration Curves using the Linear Empirical Bayes Approach, Inter. J. Pavement Eng. (2019) https://doi.org/10.1080/10298436.2019.1646912.

[15] R. Gogoi, B. Dutta, Maintenance prioritization of interlocking concrete block pavement using fuzzy logic, Inter. J. Pavement Res. Technol. 13 (2) (2020) 168–175.

[16] S. Ramachandran, C. Rajendran, V. Amirthalingam, Decision Support System for the Maintenance Management of Road Network Considering Multi-Criteria, Inter. J. Pavement Res. Technol. 12 (3) (2019) 325–335.

[17] D. Jorge, A. Ferreira, Road network pavement maintenance optimisation using the HDM-4 pavement performance prediction models, Inter. J. Pavement Eng. 13 (1) (2012) 39-51.

[18] H. P. Hong, S. S. Wang, Stochastic modeling of pavement performance. Inter. J. Pavement Eng. 4 (4) (2003) 235-243.

[19] F. Hong, Asphalt pavement overlay service life reliability assessment based on non-destructive technologies, Struct. Infrastruct. Eng. 10 (6) (2014) 767-776.

[20] N. Lethanh, K. Kaito, K. Kobayashi, Infrastructure deterioration prediction with a Poisson hidden Markov model on time series data, J. Infrastruct. Syst. 21 (3) (2015).

[21] K. A. Abaza, Simplified staged-homogeneous Markov model for flexible pavement performance prediction, Road Mater. Pavement 17 (2) (2016) 365-381.

[22] H. Perez-Acebo, S. Bejan, H. Gonzalo-Orden, Transition Probability Matrices for Flexible Pavement Deterioration Models with Half-year Cycle-time, Int. J. Civ. Eng. 16 (9) (2018) 1045-1056.

[23] M. S. Amarnath, V. U. Rejani, A. K. Raji Rural Road Connectivity Using CLUSTAL Algorithm, J. Indian Highways 39 (6) (2011) 43-54.

[24] Z. Luo, H. Yin, Probabilistic Analysis of Pavement Distress Ratings with the Cluster wise Regression Method, 87th Annual Meeting of the Transportation Research Board, Washington DC, USA, 2008.

[25] V. Sunitha, A. Veeraragavan, K. Karthik Srinivasan, Samson Mathew, Cluster-based Pavement Deterioration Models for Low-volume Rural Roads, ISRN Civ. Eng. (2012) https://doi.org/10.5402/2012/565948

[26] W. Zhang, P. Durango-Cohen, Explaining Heterogeneity in Pavement Deterioration: Clusterwise Linear Regression Model, J Infrastruct. Syst. (2014). https://doi.org/10.1061/(ASCE)IS.1943-555X.0000182

[27] Q. Li, F. Qiao, L. Yu, Clustering Pavement Roughness Based on the Impacts on Vehicle Emissions and Public Health, J. Ergonomics 5 (4) (2016) 1- 4.

[28] W. Wang, S. Wang, D. Xiao, S Qiu, J. Zhang, An Unsupervised Cluster Method for Pavement Grouping Based on Multidimensional Performance Data, J. Transp. Eng. B-Pave. 144 (2) (2018) https://doi.org/10.1061/JPEODX.0000030

[29] M. Khadka, A. Paz, and A. K. Singh Generalised Clusterwise Regression for Simultaneous Estimation of Optimal Pavement Clusters and Performance Models, Int. J. Pavement Eng. (2018) https://doi.org /10.1080/ 10298436.2018.1521970.

[30] K. R. Karthik, V. U. Rejani, V. Sunitha, Samson Mathew, and A. Veeraragavan, Urban Pavement Maintenance Management System for Tiruchirapalli City, Eighth International Conference on Maintenance and Rehabilitation of Pavements (MAIREPAV-8), Singapore, 2016 https://doi.org/10.3850/978-981-11-0449-7-109-cd

[31] Indian Roads Congress, Code of Practice for Maintenance of Bituminous Surfaces of Highways. IRC 82:1982. Journal of Indian Roads Congress, New Delhi , India, 1982.

[32] M. W. Sayers, T. D. Gillespie, C. A. V. Queiroz, Guide Lines for the Conduct and Calibration of Road Roughness Measurements, World Bank technical paper no. 46, Washington DC, USA, 1986.

[33] M. W. Sayers, T. D. Gillespie, C. A. V. Queiroz, The International Road Roughness Experiment. Establishing Correlation and a Calibration Standard for Measurements, World bank technical paper; no. WTP 45, Washington DC, USA,1986.

[34] Indian Road Congress IRC: 81:1997, Guidelines for Strengthening of Flexible Road Pavement Using Benkelman Beam Deflection Technique, Journal of Indian Roads Congress, New Delhi, India (1997).

[35] J. Han, M. Kamber, J. Pei Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Waltham, USA, 2012

[36] T. M. Kodinariya, P. R. Makwana, Review on Determining Number of Cluster in K-Means Clustering, Int. J. Adv. Res. Comput. Sci. 1 (6) (2013) 90-95.

[37] R. A. Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis. 5th edn. Pearson Prentice Hall, New Jersey, 2006.

[38] M. Jamshidian, R. I. Jennrich, W. Liu, A Study of Partial F Tests for Multiple Linear Regression Models, Comput. Stat. Data Anal. 51 (12) (2007) 6269–6284.