U.S. Department of Transportation
Federal Highway Administration

Errata

Date:	December 3, 2018
Issuing Office:	Federal Highway Administration—Office of Research, Development, and Technology: Insert Office R&D
Address:	Turner-Fairbank Highway Research Center 6300 Georgetown Pike, McLean, VA 22101
Name of Document:	Using Data Analytics for Cost-Effective Prediction of Road Conditions: Case of the Pavement Condition Index
FHWA Publication No.:	FHWA-HRT-18-065

The following changes were made to the document after publication on the Federal Highway Administration website:

Location	**Incorrect Values**	**Corrected Values**
Page 1, top of page	Missing researcher information	Researchers: S. Madeh Piryonesi and Tamer El-Diraby, Department of Civil and Mineral Engineering, University of Toronto
Page 17	Missing a space: University ofToronto	University of Toronto

SUMMARY **REPORT**

Using Data Analytics for Cost-Effective Prediction of Road Conditions: Case of the Pavement Condition Index

FHWA Publication No.: FHWA-HRT-18-065
FHWA Contact: Deborah Walker, HRDI-30, 202-493-3068, deborah.walker@dot.gov
*Researchers: S. Madeh Piryonesi and Tamer El-Diraby, Department of Civil and Mineral Engineering, University of Toronto

Abstract

Municipalities and transportation departments devote considerable effort to collecting data—particularly in relation to road conditions. Many small municipalities do not have sufficient resources to collect data regularly. In larger municipalities, on the other hand, collecting field-based data may have negative impacts in terms of crew safety and traffic interruptions; data analytics could help reduce these negative impacts. In this study, data analytics is used to test if affordable and easy-to-collect data can be used to predict future values of the Pavement Condition Index (PCI). North American transportation departments frequently use the PCI to assess road conditions. To calculate the PCI, transportation departments and municipalities must collect distress data and their severity levels.

In this study, the Long-Term Pavement Performance (LTPP) database was used as the source of data.[1] Because the LTPP database does not include the PCI values of its road sections, the first step in the study was to develop a program to calculate the PCI from the distress values in the LTPP database. Next, a set of pavement attributes was selected—mainly based on the ease of collection and cost effectiveness. The researchers tested the potential importance of these attributes in predicting PCI using seven ranking algorithms and a heuristic feature-selection algorithm.

Two types of decision trees were trained based on 942 examples of asphalt roads. Using combinations of 14 attributes, a set of decision trees was developed to predict the level of PCI deterioration with an accuracy of more than 70 percent. Finally, the accuracy and confusion matrices of different decision trees were compared to test the impact of each attribute on prediction accuracy. This method can help municipalities and transportation departments identify the most significant attributes to accurately predict road performance indicators (PIs).

Introduction

Understanding and tracking PIs of roads, especially the physical PIs, is critical to a successful asset-management plan. A better understanding of PIs helps decisionmakers schedule remedial actions, increase customer satisfaction, and be proactive in budget planning and risk assessment.

Different PIs are used to assess the condition and remaining life of roads. Some of the most popular PIs include the PCI, International Roughness Index (IRI), Structural Condition Index (SCI), and Present Serviceability Index (PSI). Collecting data for these indices could be a challenge for smaller municipalities, which are usually restricted in terms of human and financial resources. For larger municipalities, in addition to costs, collecting field-based data can have negative impacts on crew safety or traffic flow.[2]

Data analytics can provide valuable support for the data-collection and prediction processes. Recently, the availability of increased amounts of

U.S. Department of Transportation
Federal Highway Administration
https://highways.dot.gov/research
Research, Development, and Technology Turner-Fairbank Highway Research Center 300 Georgetown Pike, McLean, VA 22101-2296

*Revised December 3, 2018

data and computational power and, on the other side, the variety of available analytics algorithms have enabled engineers to move from descriptive statistics and simplistic correlation analyses to more sophisticated analytics. Data-mining and machine-learning techniques can detect patterns in large datasets, hence the growing use of analytics for different purposes in a variety of industries.[3]

This paper demonstrates how machine-learning models can help municipalities predict the PCI values of roads using easy-to-collect and cost-effective attributes. Therefore, the rationale behind choosing attributes was not a conventional mechanistic or engineering reasoning. Rather, it was to find out if affordable and accessible data can do the same job. The scope of this paper is not limited to predicting the conditions of roads using data analytics. The authors also investigated the relative significance of a road's attributes in its deterioration. This type of analysis can guide municipalities and transportation departments in crafting a more efficient data-collection and -management policy.

In this study, the PCI was chosen because it is commonly used by municipalities and transportation departments in North America. However, the same methodology can be used for analyzing other PIs, such as the IRI, SCI, and PSI. PCI values vary between 0 and 100. A PCI of 100 represents the best possible condition, and 0 represents the worst. Both ASTM and the Ontario Ministry of Transportation have produced detailed guidelines for calculating the PCI. Both sets of guidelines require collecting distress data, such as fatigue cracking, bleeding, edge cracking, rutting, longitudinal and transversal cracking, and raveling.[4,5]

Related Work: Data and Analytics in Asset Management

Without clear understanding of the value and role of datasets in analyzing asset conditions, municipalities might invest in collecting data that are not generating much value or relevant information. After surveying 50 transportation departments in the United States, Pantelias et al. reported that, in many cases, transportation agencies have created vast databases that do not necessarily supply useful information for decisionmaking.[6] Furthermore, they discovered that, in most transportation departments, data collection is still highly subjective and conventional. In other words, data are collected based on past practices and staff experience rather than solid rational analysis of relevance and value added.

Limited research is available about how to define informative data to collect. Pantelias et al. proposed a framework for data collection that aims to support project selection for rehabilitation.[6] The study provided general guidelines and a framework based on literature reviews and survey results. The framework suggests that decisionmakers must study available and missing data and identify data that are necessary to collect. Another study by Woldesenbet et al. used a social-network-analysis approach to model the use of data in generating information and supporting decisionmaking in road management.[7] Using surveys and interviews to create networks of data interrelationships, they assessed how frequently a specific piece of data was used in decisionmaking.

One of the areas of road asset management that could be improved by data analytics is deterioration modeling. Although deterioration modeling is an integral part of asset-management planning, many municipalities overlook it or use generic models. For instance, a recent study in Canada revealed that most small municipalities in Ontario did not incorporate a deterioration model in their asset-management analyses.[8] The same study reported that municipalities that paid attention to deterioration modeling mostly depended on deterministic deterioration curves to predict the conditions of their assets.[8] These deterioration curves have several pitfalls. First, they are deterministic—users have no guidelines on how to add variability to their values when conducting a probabilistic risk analysis. Second, these models are context insensitive; i.e., PCI deterioration curves predict future PCI values merely based on the length of time. These curves overlook other road attributes, such as pavement type, traffic volumes, and climate.

Stochastic deterioration models do not have the disadvantages standardized deterioration curves have.[9] Markovian models, for example, study

and estimate deterioration based on probabilistic analysis. Nonetheless, they often disregard the history of deterioration and previous maintenance actions.[9–11] Additionally, they require longitudinal data, which are not easily available.[8,9] Data-analytics tools that learn or detect patterns from a large dataset can be a suitable alternative. Data analysis is a broad term that has been used to refer to a range of methods, from simple statistical analysis to machine-learning and data-mining techniques. In this summary report, data analytics specifically refers to machine learning and data mining only. Machine-learning and artificial-intelligence algorithms have become popular in civil engineering, including analytics to predict the condition of roads. For instance, Yang et al. used neural networks to predict variations in the crack index of asphalt roads over a short term.[12] Neural networks have a good learning capability; however, large amounts of data are needed for their training and calibration.[12] Furthermore, the black-box nature of neural networks does not help in understanding the relative importance of attributes.[9] In many cases, other algorithms such as decision trees are preferred due to their ease of interpretation and implementation, although algorithms such as neural networks might result in slightly higher accuracy.[3,13,14]

Decision trees have been used to analyze and predict PIs. Chi et al. trained decision trees based on data from the Texas Department of Transportation.[2] They stated that transportation departments can use the results of their models for parts of their networks when falling weight deflectometer data are not available. The accuracy of their models in predicting five levels of SCI was approximately 60 percent, which is satisfactory. They trained models using attributes such as amount of distress and ride score, which are not the cheapest data to collect. Moreover, the attributes were averaged out over a period of 5 yr (e.g., 5-yr average of distress). Using data of multiple years in one attribute to train the model may be one of the limitations of their work because most municipalities and transportation departments do not have updated data for several consecutive years.[2] Furthermore, the size of the training set used by Chi et al., which was 354 road sections, may raise some questions regarding the reliability and robustness of their models.[2]

Researchers have conducted data analysis on data from the LTPP database to model PIs.[2,15,16] Using the historical distress data in the LTPP database and Minnesota road database, Wu developed a methodology to predict the PCI of asphalt roads over time by calculating the current PCI from distress values and predicting future PCI values using PCI master curves.[17] In another study, Meegoda and Gao developed a quantitative relationship between roughness progression and accumulative traffic load, structural number, annual precipitation, and freezing index.[15] Moreover, they used a Weibull distribution to investigate the reliability of roughness progression models.

Objectives: Predictions Using Cost-Effective Data

The first objective of this research was to train a machine-learning model that could adequately predict PCI deterioration within 3 yr through easy-to-collect and affordable data for use by municipalities, especially ones with limited financial resources. The rationale behind using a 3-yr span is that most municipalities in Ontario conduct a comprehensive survey on their road network every 5 yr. A 3-yr prediction can provide a suitable interim estimate.

The decision was made to adopt classification algorithms, particularly decision trees. These algorithms provide an open-box approach whereby decisionmakers can test the role of every attribute at different stages of the analysis. There are three additional reasons for choosing the decision-tree approach for this study. First, training a decision tree (e.g., a C4.5) requires almost no prerequisites or assumptions about the data. In other words, there is almost no limitation on the type of attributes that are used to train a decision tree, which is not the case for some other algorithms; for instance, the attributes to train a naive Bayes classifier must be independent. Second, decision trees are intuitive and easy to interpret. Third, they can be easily implemented and reused for new data. Unlike other classification algorithms, such as the k-nearest neighbors or the naive Bayes classifier, decision trees result in an explicit model that can

easily classify new examples.[3,13] In this paper, the decision trees were developed and validated through mining LTPP data. Municipalities with no longitudinal data can benefit from these models.

The second objective was to determine which attributes have the largest impact on predicting PCI. This objective was accomplished by developing different decision trees using combinations of 14 attributes. Later, the frequency of appearance by each attribute in all trees and its relevant position in each tree were studied.

Methodology

This section summarizes the steps of this study. Since there is no PCI in the online LTPP database, a tool was needed to calculate PCI values of asphalt roads from the distress data. The ASTM methodology was adopted for calculating the PCI values from distresses.[3] To automate the process of calculation, all ASTM curves first needed to be digitized and expressed mathematically via curve fitting. Next, a Python™ program was developed to extract distress data for each road segment and use the formulas to calculate the PCI for each segment. The generated PCI values (with a 3-yr spread) presented the initial and target values for the predictive models. In the next step, a literature review and a set of interviews were conducted to identify possible relevant attributes. A list of 14 attributes was prepared, and data were retrieved from the online LTPP database using queries based on Structured Query Language (SQL). After completing data preparation and cleansing, seven ranking algorithms and a heuristic feature-selection algorithm were applied to the retrieved datasets to identify those attributes that have the largest potential in predicting future PCI values.

After identifying the most relevant and informative attributes, two types of decision trees were trained to predict the PCI of roads after 5 yr. The accuracy of both trees was tested for unseen data using cross validation. To test the effect of the size of the training set on model accuracy, models were trained by different numbers of examples (i.e., 250, 550, and 942). The effect of the size of the training set boosted the accuracy. Finally, the confusion matrices of the decision trees were studied to understand whether the wrong predictions of the models were overestimating or underestimating the PCI. Therefore, in this study, model evaluation was not limited to a one-number evaluation (correlation) as it is practiced by way of traditional regression analysis. This ability to investigate the nature of wrong predictions is a major advantage of machine-learning approaches over simplified statistical analysis methods.

Data Preparation and Cleansing

Data were retrieved using SQL-based queries. It is worth mentioning that Specific Pavement Studies (SPS) sections were not used in the training set because the SPS sections that are colocated have identical traffic or climatic data.[18] A large portion of data preparation focused on generating the PCI values from distresses, which is explained in the next section. Since data were stored in different tables, different fields were collated by SQL join queries. Some attributes were created from a combination of two or more attributes. For instance, the attribute "last remedial action" is a combination of both major rehabilitation and maintenance actions, which are stored in different tables. Data cleansing included removing erroneous records. An example of an erroneous record is when, without maintenance, the PCI increases after 3 yr.

Generating PCI Values to Train Models

The training set must include the PCI value for each road section, which is not included in the LTPP database. The distress data and the dimensions of road sections were retrieved from the LTPP program's online platform.[1] A Python™ program was developed to generate the PCI values from distress data according to the ASTM methodology. For this purpose, all deduct value graphs and correction curves were digitized and mathematically represented. After finding the mathematical functions of curves, the formulas were implemented in a Python™ program. The required steps for the digitization of graphs and the extraction of formulas, all shown in figure 1 and figure 2, follow.

Figure 1 shows the curves proposed by ASTM D6433 - 07 for calculating deduct values of potholes with different levels of severity in metric units. A large number of points, which are shown by red dots, were picked on the curve. The points were then drawn on a scatter chart with a logarithmic scale on the x-axis, and a polynomial curve was fitted to the points, as shown in figure 2. The coefficient of correlation was very close to 1. Considering the coefficients of figure 2, the formula for calculating the deduct value of potholes at a high severity (metric units) is shown in figure 3.

Figure 1. Graph. Digitized graph for calculation of deduct values for high-severity potholes (graph regenerated based on guidelines of ASTM D6433 - 07).[4]

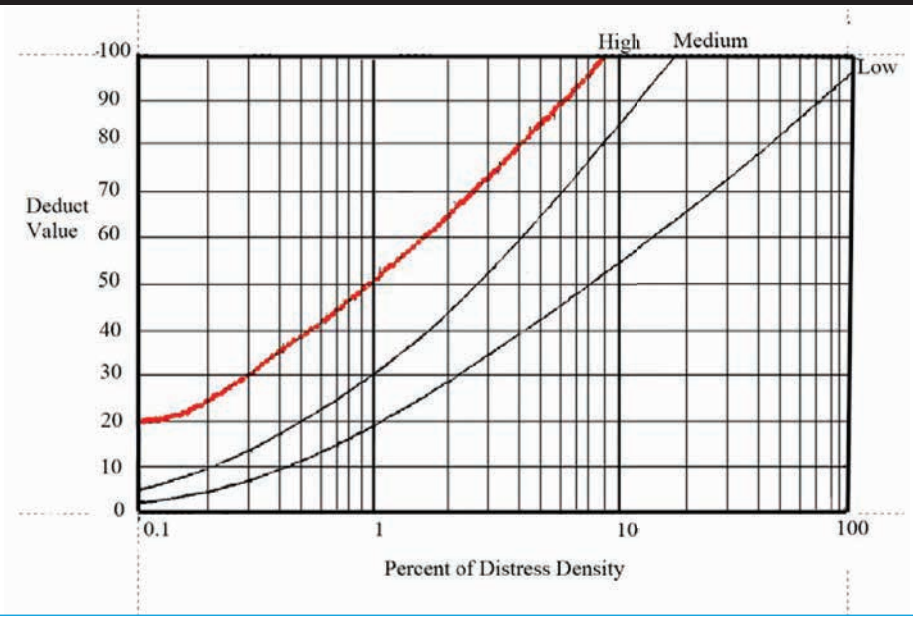


Figure 2. Graph. Fitting a polynomial to the ASTM deduct value curve highlighted in figure 1 and extracting the mathematical formula for potholes with a high severity (metric units).

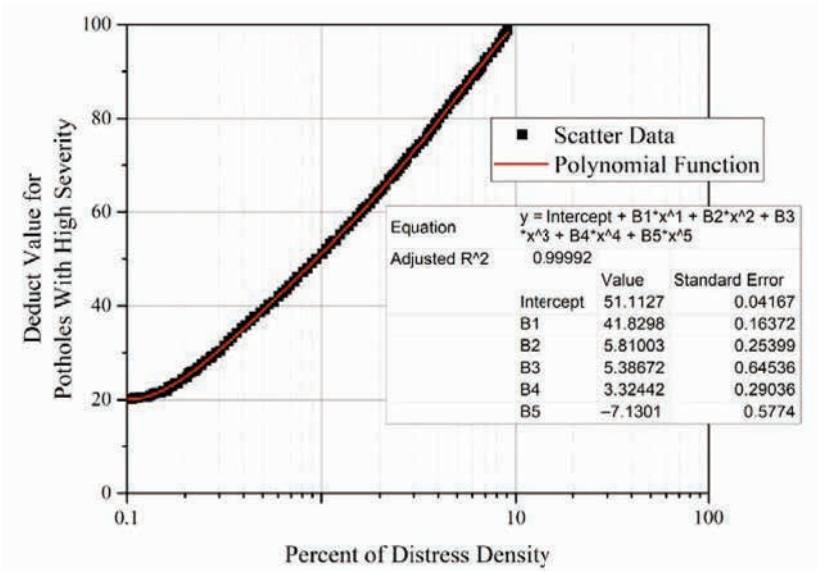


Figure 3. Mathematical formula extracted for the deduct value of high-severity potholes.

$$y = -7.13x^5 + 3.32x^4 + 5.38x^3 + 5.81x^2 + 41.83x + 51.11; \quad R^2 = 0.9999$$

Where:

y = the deduct value.
x = the logarithm of distress density.
R^2 = the coefficient of correlation.

Altogether, 31 deduct values for distress density curves and 8 correction curves were digitized and embedded into a spreadsheet and a Python™ script to calculate the PCI values automatically. Extracted polynomial functions were similar to the formulas reported by Wu, with minor differences as a result of adopting metric units.[17]

Choosing Attributes

Predictive attributes were chosen after conducting a literature review, interviewing 3 experts, studying asset management plans developed in Ontario, and investigating the data collected by 10 small Ontario municipalities. Because most small municipalities typically do not have sufficient funding for data collection, the attributes with relatively low cost of acquisition were selected. Unlike most related previous works, these attributes were chosen based on cost rather than mere engineering reasoning. Table 1 shows the initial 14 attributes chosen to train models.

Selecting the Most Informative Attributes

All 14 attributes in table 1 cannot be used simultaneously to train a model due to overfitting. Overfitting usually happens when a model is too complicated and is fitted to the noise. Such a model has a very low training error and high testing error. In other words, it classifies training data very well, but it fails to classify unseen and new data satisfactorily.[3] Therefore, seven different ranking algorithms were used to identify the relative importance of each initial attribute in predicting the PCI. The algorithms used for screening attributes were information gain, information gain ratio, correlation-based feature selection, chi-squared, Gini index, weighting by rule, and symmetrical uncertainty. Each algorithm assigns a weight or rank to the attributes depending on their contribution to the prediction of the PCI. For instance, the correlation-based feature selection gives a higher rank to attributes that are more correlated with the PCI, while the information gain ranks attributes based on the reduction that each attribute can create in the entropy of the system.[13,19] Table 2 summarizes the ranks of the initial attributes in predicting the PCI as calculated by each algorithm and the average rank of each attribute. According to table 2, PCI0 and FUNC_CLASS are, respectively, the most and the least informative attributes in predicting the PCI value after 3 yr.

The results of ranking algorithms are instrumental in any predication of the PCI. However, the final decision needed more investigation regarding the attributes that had the largest impact on accuracy. Therefore, the Optimize Selection operator of the software RapidMiner™ was also applied to the dataset. This operator selects the most relevant attributes of a dataset. It has a heuristic approach that applies two deterministic greedy feature-selection algorithms: forward selection and backward elimination.[3,13,20] In simple words, in a dataset with n attributes, this operator selects m features ($m < n$) such that they maximize the accuracy of learned models (i.e., decision trees), where n is the total number of attributes in the training set, and m is the number of most informative attributes that maximize the accuracy of the model. An attribute with a weight of 1 has a role in increasing the accuracy of a decision tree, while a weight of 0 means it is possible to train a tree with the same accuracy without using that attribute. The Weight in Heuristic Algorithm column in table 2 confirms the results of average rankings of the attributes. Most attributes with a high average rank received a weight of 1 by the Optimize Selection operator. It is worth noting that this operator does not guarantee finding a global optimum due to its heuristic nature.[20]

Table 1. Initial list of attributes.

Field Name	Description
PCI0	The initial value of the PCI or the value in the current year
AGE	Age of road (since the construction date)
PAVEMENT_TYPE	Type of asphalt pavement
FREEZE_INDEX_YR	Calculated freeze index for year
MAX_ANN_TEMP_AVG	Average of daily maximum air temperatures for year
MIN_ANN_TEMP_AVG	Average of daily minimum air temperatures for year
TOTAL_ANN_PRECIP	Total precipitation for year
FUNC_CLASS	Functional class of road
FREEZE_THAW_YR	Number of freeze–thaw cycles per year
OVERLAY_THICKNESS	Thickness of the placed layer in rehabilitation
AADT_ALL_VEHIC_2WAY	Average annual daily traffic
REMED_TYPE	Type of last remedial action
REMED_YEARS	Number of years since the last remedial action
CONSTRUCTION_NO	Number of conducted remedial actions
PCI (target variable)	PCI after 3 yr (as categorized by the ASTM)

Table 2. Identifying the most informative attributes using feature-selection algorithms

Attribute	Information Gain	Information Gain Ratio	Correlation Based	Chi-Squared	Gini Index	Weight by Rule	Uncertainty	Average Rank	Weight in Heuristic Algorithm
PCI0	1	1	1	1	1	1	1	1	1
REMED_YEARS	2	11	2	2	2	7	2	4	1
FREEZE_INDEX_YR	3	9	6	8	3	12	4	6.4	0
MAX_ANN_TEMP_AVG	6	8	14	4	7	9	7	7.8	0
AGE	5	6	8	6	5	2	6	5.4	1
FREEZE_THAW_YR	4	4	5	5	4	3	5	4.2	1
MIN_ANN_TEMP_AVG	7	7	11	7	6	4	9	7.2	0
OVERLAY_THICKNESS	8	12	13	13	9	8	13	10.8	0
CONSTRUCTION_NO	9	10	9	12	8	10	12	10	0
AADT_ALL_VEHIC_2WAY	10	13	3	9	11	14	8	9.7	1
TOTAL_ANN_PRECIP	11	3	12	10	10	5	10	8.7	0
FUNC_CLASS	12	5	10	14	12	11	14	11.1	0
PAVEMENT_TYPE	13	8	7	11	13	13	11	10.8	0
REMED_TYPE	14	14	4	3	14	6	3	8.2	1

Models

Decision trees were chosen for this study, with the PCI value in 3 yr being the target variable. For training a decision tree, the target value must be discrete. Therefore, PCI values were discretized according to the ASTM rating scale illustrated in figure 4. As shown in the right side of the figure, ASTM divides the PCI into seven classes.[4] A conceptual representation of the implemented models demonstrates that, after training and implementing the models, users can input the values of selected attributes and get the level of PCI after 3 yr.

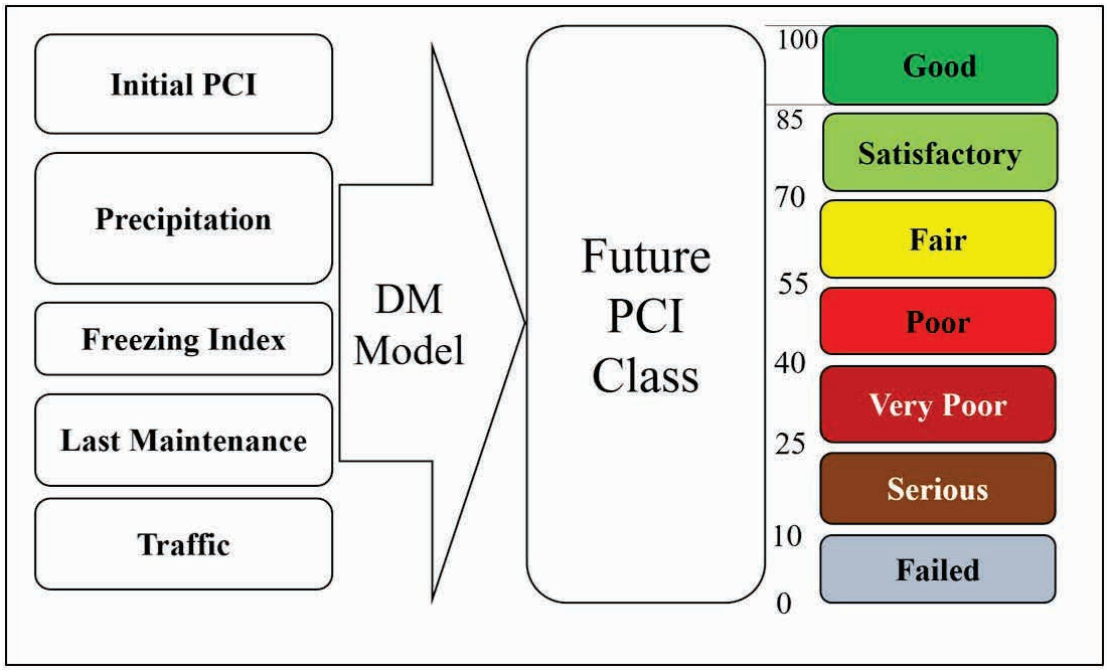
Training Decision Trees

Two decision trees were learned from the prepared training set, which contained 942 examples of road sections. The two trees are the default decision tree of RapidMiner™ (decision tree I) and a C4.5 (decision tree II). These models, in contrast to old decision trees such as ID3, are capable of learning from both categorical and continuous attributes—especially a C4.5, which is a descendant of CLS and ID3 and has a high learning capability.[2,21]

Figure 5 shows a snapshot of decision tree I learned from four attributes: PCI0, REMED_YEAR, FREEZE_THAW_YR, and REMED_TYPE. Similarly, figure 6 shows the tree learned from a C4.5 algorithm with the same examples and attributes.

As mentioned, decision trees are easy to interpret. For instance, figure 6 suggests that, when the current PCI value is larger than 85.1 and smaller than 91.6 (i.e., 85.1 < PCI0 < 91.6) and the road

Figure 4. Illustration. Conceptual representation of implemented models.



DM = data-mining.

Figure 5. Illustration. Decision tree I trained by 942 examples and 4 attributes.

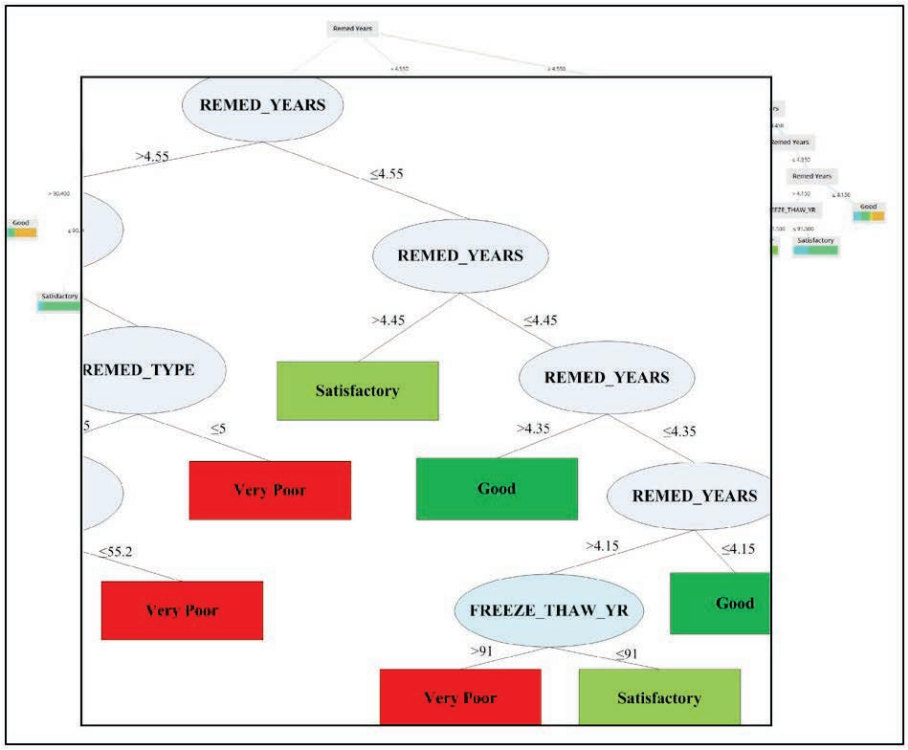
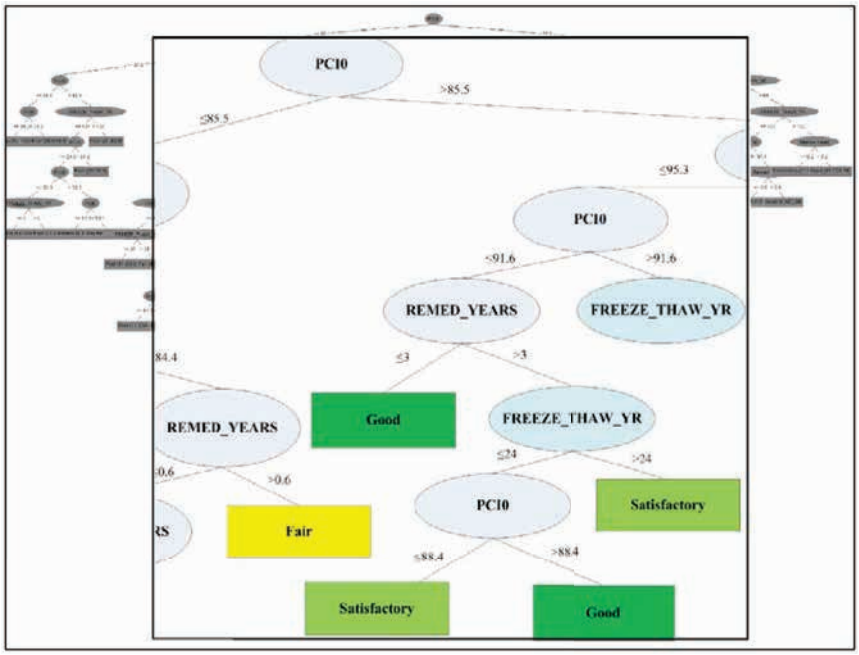


Figure 6. Illustration. Decision tree II (a C4.5) trained by 942 examples and 4 attributes.



has been maintained within the last 3 yr (REMED_YEARS ≤ 3), it will stay in good condition for 3 yr. However, if the last remedial action occurred more than 3 yr ago and the road experiences more than 24 freeze–thaw cycles per yr (FREEZE_THAW_YR > 24), its condition will fall to satisfactory.

Both models were tested multiple times with a similar number of examples and parameters; decision tree II (i.e., a C4.5) showed a higher accuracy (the next section, Model Evaluation, provides further detail). Figure 7 and figure 8 compare the mean and standard deviation of accuracy of the 2 decision trees, both trained by the same training set with 3 different numbers of examples: 250, 550, and 942. Figure 7 clearly demonstrates that decision tree II is outperforming its rival in all three cases. Higher accuracy of a C4.5 decision tree supports the results of previous research.[2] Additionally, figure 7 and figure 8 show that, by increasing the number of examples, the mean accuracy increases and its standard deviation decreases. Therefore, models trained by larger datasets are more accurate and robust.

The hierarchy of the attributes within a decision tree reflects how informative the attributes are. This ease of interpretation is a great advantage of decision trees over other classification techniques. For the convenience of users, the decision trees in figure 5 and figure 6 were implemented using MATLAB®. Users can input the values of the attributes of their roads and climatic data and receive assessment about the deterioration of their roads after 3 yr. This assessment is context sensitive and is not a standard curve.

Figure 7. Graph. Comparing the mean accuracy of the two trained decision trees for different sizes of training sets.

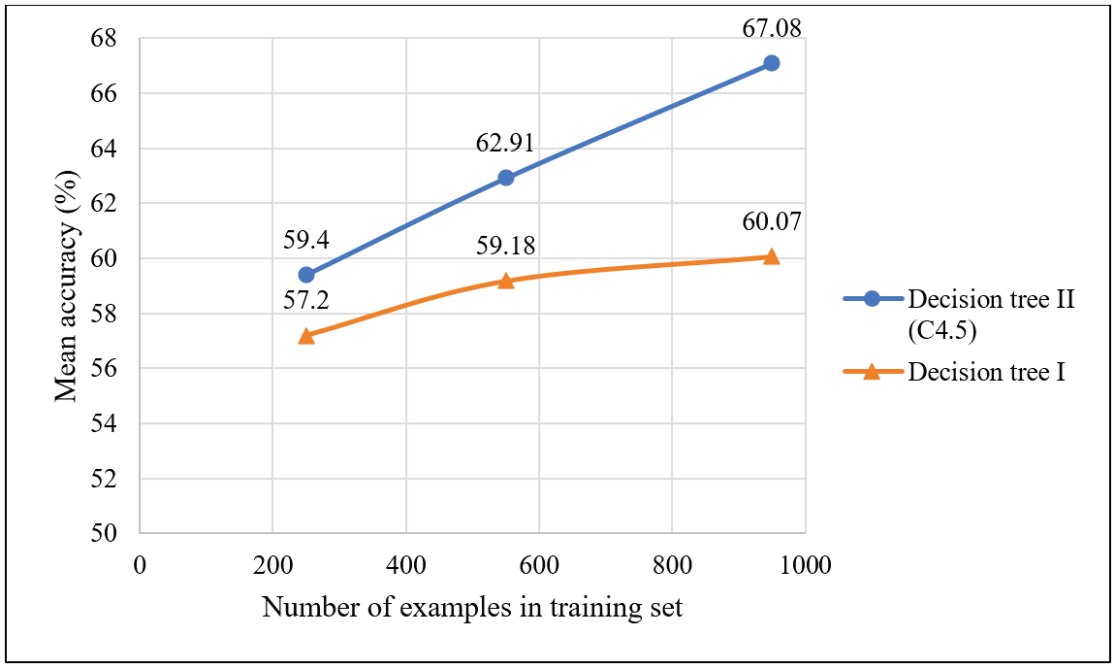
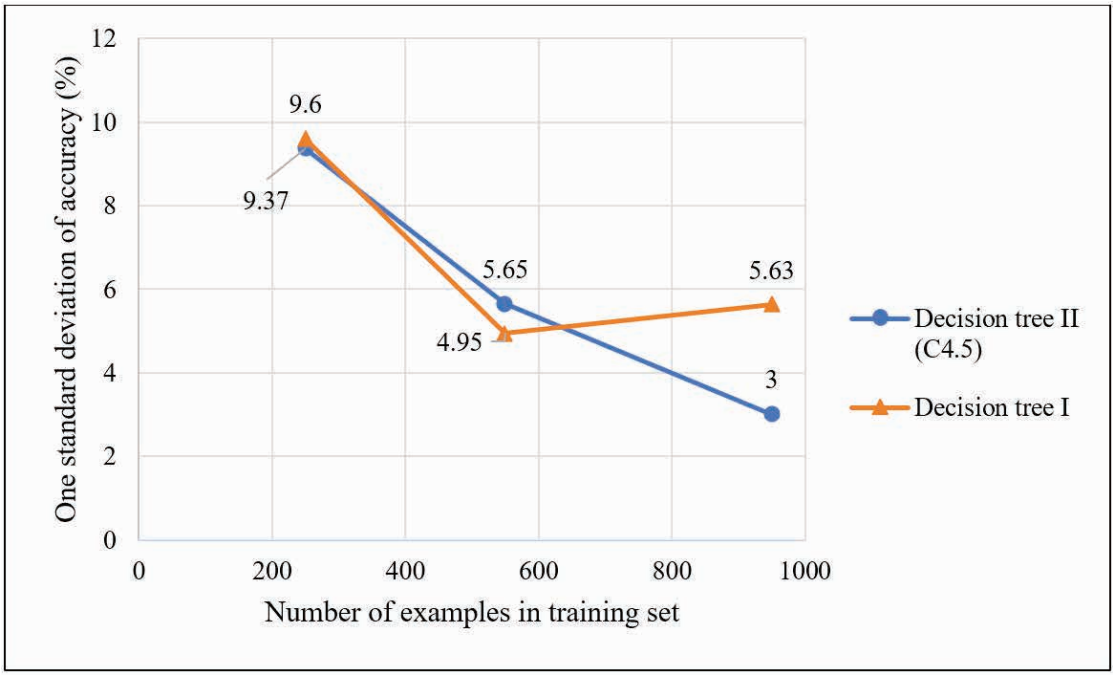


Figure 8. Graph. Comparing one standard deviation of cross-validation accuracy of the two trained decision trees for different sizes of training sets.



Model Evaluation

The accuracy of the developed models was tested using cross validation. The training data were divided into 10 subsets. The models were trained based on 9 of them and tested using the 10th. This process was iterated 10 times. The accuracy of the model is the average of these 10 iterations. The best overall accuracy of the learned decision tree trained by 942 examples was 69.2 ± 4.7 percent. This means that, on average, approximately 70 percent of predictions were correct. This number is calculated by dividing the sum of the elements on the main diagonal of the confusion matrix (table 3), which represents the number of correct predictions, by the sum of all elements, which represents all predictions (942). Note that, in this case, the basic odds of making a correct prediction by wild guessing is 1/7 (or 14.3 percent) because the target variable has seven labels. This positively reflects a high level of performance of the proposed model. The ±4.7-percent value represents one standard deviation of the accuracy.

Studying confusion matrices provides insights beyond one-number evaluations. In contrast to the measures of correlation-based and descriptive statistical techniques, a confusion matrix can reveal how fatal incorrect predictions are.[3] Table 3 shows a confusion matrix that resulted from testing the accuracy of a C4.5. As mentioned, the general accuracy of this model for classifying unseen data is 69.2 ± 4.7 percent. The columns of a confusion matrix show the number of actual examples in each class, and the rows represent the predictions of the model. Therefore, the class recall of a specific class describes the performance of a model in predicting the label of that class. According to table 3, the developed model has a high recall for good and a low recall for failed and serious. From a data analytics perspective, this difference is due to the large number of good and small number of failed roads in the training set. From a practical perspective, this difference in class recalls could be an advantage of the model because the percentage of roads

Table 3. Confusion matrix of a C4.5 trained by 942 examples.

Pred./Act. PCI	Actual Good	Actual Satisfactory	Actual Fair	Actual Poor	Actual Very Poor	Actual Serious	Actual Failed	Class Precision
Predicted good	180	38	18	1	1	0	0	75.6%
Predicted satisfactory	16	98	26	7	2	1	0	65.3%
Predicted fair	7	23	161	14	10	6	0	72.9%
Predicted poor	2	3	22	91	19	6	0	63.6%
Predicted very poor	0	2	10	14	92	11	3	69.9%
Predicted serious	0	1	4	2	11	30	4	57.7%
Failed	0	0	0	0	0	5	1	16.6%
Class recall	87.8%	59.4%	66.8%	70.5%	67.6%	50.8%	20.0%	—

Note: Bolded cell borders indicate main diagonal of matrix.
—No data.

in a failed condition is very low in real-world networks. This observation leads to a practical solution for increasing the accuracy of this model. The solution is to merge three lower classes (i.e., very poor, serious, and failed). Merging classes is a common approach to increasing the accuracy of classifiers.[3] In this case, it is quite practical for two reasons. First, these three classes constitute only 21 percent of examples. Second, real-world roads with a PCI lower than 40 usually need similar treatments. Therefore, the number of classes (labels) was reduced to five. All PCI numbers lower than 40 were labeled very poor. As expected, reducing the number of classes resulted in increased accuracy. The cross-validation accuracy of the new model was 72.5 ± 5.07 percent.

When working with decision trees, three criteria can be considered to determine the attributes with the largest impacts. First, which attributes maximize the accuracy? Second, which attributes appear in a higher position in the tree hierarchy? Third, which attributes result in a more cost-effective confusion matrix? In this context, a more cost-effective confusion matrix is defined as a matrix with fewer cases of overestimations. The answer to the first two criteria may be inferred from table 2, but answering the third question requires analyzing the confusion matrix. The idea of a cost-effective confusion matrix is explained using table 3. In the Actual Satisfactory column of this table, 98 examples were correctly predicted as satisfactory. Among incorrect predictions, 38 examples were class-ified as good, which is an overestimation of the PCI. The rest of the predictions for satisfactory class under-estimated the PCI. Since overestimating the PCI can result in a faster-than-predicted deterioration of roads (hence a reduction in the levels of service and customer satisfaction) its secondary costs are larger. Therefore, in the case of PCI prediction, it can be interpreted as a false-negative prediction. Accordingly, between two models with the same accuracy, the model with the smaller number of false-negative predictions is more cost-effective.

Table 4 compares the results of a C4.5 learned from different combinations of four attributes. The decision trees in table 4 have similar accuracies. However, if an overestimation of the PCI is defined as false negative, the number of false-negative predictions of decision tree I (i.e., the tree trained based on PCI0, REMED_YEARS, FREEZE_THAW_YR and REMED_TYPE) is considerably lower than others. This quality could be interpreted as a strength of decision tree I. In other words, this study recommends that engineers not rely on one-number evaluations of models but rather take into consideration all aspects of different models.

It should be noted that the large number of false negatives of decision tree II could be a result of missing values of average annual daily traffic (AADT) data. The impact of data quality on accuracy needs further research, especially for algorithms such as a C4.5 that cannot handle missing values of predictor attributes. The LTPP database includes a lot of missing values for those attributes that are reported by local agencies. The AADT is an example of such variables.

Table 4. Comparing models using their number of false-negative predictions.

Decision Tree	Attributes	Number of False Negatives	Accuracy (%)
1	PCI0 REMED_YEARS FREEZE_THAW_YR REMED_TYPE	153	67.08 ± 3.00
2	PCI0 REMED_YEARS FREEZE_THAW_YR AADT_ALL_VEHIC_2WAY	198	67.93 ± 3.96
3	PCI0 REMED_YEARS FREEZE_THAW_YR AGE	188	67.19 ± 4.28
4	PCI0 FREEZE_THAW_YR FUNC_CLASS MAX_ANN_TEMP_AVG MIN_ANN_TEMP_AVG	193	64.96 ± 6.21

Conclusion and Recommendations

In this study, two decision trees were trained to predict the PCI value of roads after 3 yr. A machine-learning approach was adopted to overcome the weaknesses of previous PCI prediction models—mainly their use of deterministic curves. Careful consideration was given to selecting the most accessible and economical attributes. First, a provisional list of 14 easy-to-collect and relevant attributes was prepared. Since the LTPP database does not include PCI values, a program was developed to calculate the PCI from distress data. After calculating PCI values and adding them to the training set, seven ranking algorithms and a heuristic feature-selection algorithm were applied to data to identify the most relevant attributes.

The accuracy of the decision trees reached approximately 75 percent for unseen data. Considering the results of ranking algorithms, the accuracy of models, and the number of false negatives in the confusion matrices, several recommendations were made about the most informative data. The analysis showed that the current PCI and the time since the last remedial actions are among the most informative attributes for predicting future PCI values. On the other hand, the functional class of road and the pavement type were the least informative features of this dataset. It was also

demonstrated that one-number summaries cannot represent the performance of a model properly. Further observations, such as studying confusion matrices, are necessary to assess the performance of models and the value of data. Finally, the researchers emphasize that the findings reported here are based on the desire to use easy-to-collect data. It is possible that adding more technical and engineering attributes could enhance the accuracy of the models. In that case, a cost-accuracy tradeoff analysis could be considered.

References

1. Federal Highway Administration. (2017). "LTPP InfoPave™." (website) Washington, DC. Available online: https://infopave.fhwa.dot.gov/, last accessed July 27, 2017.
2. Chi, S., Murphy, M., and Zhang, Z. (2014). "Sustainable Road Management in Texas: Network-Level Flexible Pavement Structural Condition Analysis Using Data-Mining Techniques." *Journal of Computing in Civil Engineering*, 28(1), pp. 156–165, American Society of Civil Engineers, Reston, VA.
3. Provost, F. and Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc., Sebastopol, CA.
4. Way, N.C., Beach, P., and Materials, P. (2015). *ASTM D6433–07: Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys*, ASTM International, West Conshohocken, PA.
5. Chong, G.J., Phang, W.A., and Wrong, G. (1982). *SP-024, Manual for Condition Rating of Flexible Pavements*, Ministry of Transportation of Ontario, Ontario, Canada.
6. Pantelias, A., Flintsch, G.W., Bryant, J.W., and Chen, C. (2008). "Asset Management Data Practices for Supporting Project Selection Decisions." *Public Works Management & Policy, 13*(3), pp. 239–252, SAGE Publications, Thousand Oaks, CA.
7. Woldesenbet, A., Jeong, H.D., and Park, H. (2015). "Framework for Integrating and Assessing Highway Infrastructure Data." *Journal of Management in Engineering, 32*(1), 04015028, American Society of Civil Engineers, Reston, VA.
8. Kinawy, S.N., El-Diraby, T.E., and Piryonesi, S.M. (2017). "A Comprehensive Review of Approaches Used by Ontario Municipalities to Develop Road Asset Management Plans." Presented at the 96th Annual Meeting of the Transportation Research Board, Washington DC, United States. Available online: https://trid.trb.org/view/1437181, last accessed September 14, 2018.
9. Ens, A. (2012). *Development of a Flexible Framework for Deterioration Modelling in Infrastructure Asset Management*, Master of Science thesis, Department of Civil Engineering, University of Toronto, Toronto, Canada. Available online: https://tspace.library.utoronto.ca/bitstream/1807/33410/1/Ens_Abra_M_201211_MASc_thesis.pdf, last accessed February 5, 2017.
10. Piryonesi, S.M. and Tavakolan, M. (2017). "A Mathematical Programming Model for Solving Cost-Safety Optimization (CSO) Problems in the Maintenance of Structures." *KSCE Journal of Civil Engineering*, *21*(6), pp. 2,226–2,234, Korean Society of Civil Engineers, Springer, Berlin, Germany.
11. Kleiner, Y. (2001). "Scheduling Inspection and Renewal of Large Infrastructure Assets." *Journal of Infrastructure Systems*, *7*(4), pp. 136–143, American Society of Civil Engineers, Reston, VA.
12. Yang, J., Lu, J., Gunaratne, M., and Xiang, Q. (2003). "Forecasting Overall Pavement Condition With Neural Networks: Application on Florida Highway Network." *Transportation Research Record: Journal of the Transportation Research Board*, *1853*, pp. 3–12, Transportation Research Board, Washington, DC.

13. Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York, NY.

14. Shahin, M.A., Jaksa, M.B., and Maier, H R. (2009). "Recent Advances and Future Challenges for Artificial Neural Systems in Geotechnical Engineering Applications." *Advances in Artificial Neural Systems*, *2009*, pp. 1–9, Hindawi, London, UK.

15. Meegoda, J.N., Asce, F., and Gao, S. (2014). "Roughness Progression Model for Asphalt Pavements Using Long-Term Pavement Performance Data." *Journal of Transportation Engineering*, *140*(8), pp. 1–7, American Society of Civil Engineers, Reston, VA.

16. Dong, Q. and Huang, B. (2014). "Evaluation of Influence Factors on Crack Initiation of LTPP Resurfaced-Asphalt Pavements Using Parametric Survival Analysis." *Journal of Performance of Constructed Facilities*, *28*(2), pp. 412–421, American Society of Civil Engineers, Reston, VA.

17. Wu, K. (2015). *Development of PCI-Based Pavement Performance Model for Management of Road Infrastructure System*, Master of Science thesis, Arizona State University, Tempe, AZ.

18. Elkins, G.E., Schmalzer, P., Thompson, T., and Simpson, A. (2003). *Long-Term Pavement Performance Information Management System, Pavement Performance Database User Reference Guide*, Report No. FHWA-RD-03-088, Federal Highway Administration, Washington, DC.

19. Hall, M. (1999). *Correlation-Based Feature Selection for Machine Learning*, Doctorate of Philosophy thesis, University of Waikato, Hamilton, New Zealand.

20. RapidMiner. (2017). "Optimize Selection—RapidMiner Documentation" (website) Boston, MA. Available online: https://docs.rapidminer.com/studio/operators/modeling/optimization/feature_selection/optimize_selection.html, last accessed July 27, 2017.

21. Wu, X., Kumar, V., Ross, Q.J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., et al. (2008). "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems*,*14*(1), pp. 1–37, Springer-Verlag, Berlin, Germany.

Researchers—This study was conducted by S. Madeh Piryonesi and Tamer El-Diraby at the Department of Civil and Mineral Engineering, University of Toronto, with the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). The report is based on the LTPP data and research cited within the document. The authors would like to thank Mr. Joe Tierney, Dr. James Smith. The efforts of Mr. Wei Yu Zhao, University of Toronto, are acknowledged as well.

Distribution—This summary report is being distributed according to a standard distribution.

Availability—This summary report may be obtained from the FHWA Product Distribution Center by email to report.center@dot.gov, by fax to 301-577-1421, by phone to 301-577-0818, or online at https://highways.dot.gov/research/.

Key Words—Machine learning, data analytics, Pavement Condition Index, data collection, asset management, deterioration modeling.

NOVEMBER 2018

FHWA-HRT-18-065
HRDI-30/11-18(WEB)

*Revised December 3, 2018