Pavement Condition Prediction for Communities: A Low-Cost, Ubiquitous, and Network-Wide Approach

Tao Tao, Ph.D.[1]; and Sean Qian, Ph.D., M.ASCE[2]

Abstract: Effective prediction of pavement deterioration is critical to forecast infrastructure performance and make infrastructure investment decisions under escalating environmental and traffic change. However, most communities often struggle to undertake such predictive tasks due to limited sensing capacity and lack of granular data. With the pavement condition rating (PCR) data generated from artificial intelligence (AI)-powered computer vision technologies and multiple openly available data sets, we propose a low-cost and ubiquitous approach to predict system-level pavement conditions using nine communities across the US as an example. In addition to predicting absolute PCRs as was done in classical models, we develop another set of models to predict the change in PCRs over any time increment (i.e., time lapse between a PCR observation and retrofit decision point) and compare the results. The findings showed that the proposed low-cost prediction approach yields results comparable to existing studies, demonstrating its promising application in supporting pavement management. Furthermore, the PCR change model indicates that, besides current PCR, weather, road classification, socioeconomics, and built environment attributes are important to predicting PCR change. The interactive impacts also show salient interactive effects between variables and current PCR, offering suggestions on better allocating the limited resources in pavement maintenance projects. Finally, the proposed model could enhance climate resiliency and transportation equity during the pavement management process. **DOI: 10.1061/JITSE4.ISENG-2378.**

Introduction

Effective prediction of pavement conditions plays a critical role in maintaining the performance of road networks and enhancing the resilience of communities. Accurate predictions enable transportation agencies to proactively manage pavement assets, schedule timely maintenance, and allocate resources more efficiently (Piryonesi and El-Diraby 2021). This leads to improved serviceability of roads, extended pavement life, and reduced life-cycle costs. In the context of increasing environmental challenges, accurate pavement condition predictions also support efforts to enhance the resilience of communities. By anticipating and mitigating pavement distresses associated with changing climate and extreme weather events, communities can better prepare for and adapt to these challenges, ensuring the continuity and reliability of their transportation systems.

Although some large metropolitan areas may have the capacity and resources to undertake pavement condition prediction tasks and utilize the results to inform their pavement asset management programs, most communities face more significant challenges in doing so, mainly because they cannot afford to collect and maintain data sets related to pavement conditions and relevant built/natural environment measures, oftentimes depending on costly sensors and lab experiments. It is a common challenge, for both large and small communities, to measure pavement ratings for more than a small fraction of road segments on a yearly basis.

Although studies have been using several data sets maintained by either federal or local transportation agencies to carry out pavement condition predictions, these data sets have several limitations. First, these data sets mainly cover a small fraction of highways and arterial road sections in a community-wide area, which provides limited insights for small or underserved communities because most of their roads are local roads. Second, these data sets only encompass a limited number of road sections for each community. Most communities do not have access to a complete set of road inventory data for every road segment, such as pavement materials, pavement ratings, and so on. Thus, it is unclear if the resultant predictive model does a comprehensive examination on all the road sections in the whole network. Third, these data sets often contain missing value for variables that are important to predict pavement conditions, such as pavement characteristics and traffic activities. In addition to the challenges raised by the quality and availability of related data sets, most studies only split the sample once when they train and test models, making the model performance heavily dependent on the specific data partitioning.

In this paper, we propose a practical approach to address these challenges faced by communities. With the pavement condition data generated from artificial intelligence (AI)-powered computer vision technologies and multiple openly available data sets, we carry out network-wide pavement condition predictions using a complete coverage of all roads in nine communities across the US. We use a rigorous approach known as repeated nest cross-validation to tune parameters and select the best modeling method among multiple machine learning methods. In addition to predicting absolute value of pavement conditions as was done in classical models, we develop another set of models to predict the change of pavement conditions over any time increment (i.e., time lapse between two observations of pavement conditions for road segments). We compare the results between these two sets of models.

Our study has several important contributions to pavement predictions in practice. First, we propose a low-cost and ubiquitous approach to predict system-level pavement conditions in communities

[1]Postdoctoral Researcher, Dept. of Civil and Environmental Engineering, Carnegie Mellon Univ., Pittsburgh, PA 15213. ORCID: https://orcid.org/0000-0001-9865-5985. Email: taot@andrew.cmu.edu

[2]Professor, Dept. of Civil and Environmental Engineering, Carnegie Mellon Univ., Pittsburgh, PA 15213 (corresponding author). ORCID: https://orcid.org/0000-0001-8716-8989. Email: seanqian@cmu.edu

Note. This manuscript was submitted on June 27, 2023; approved on September 30, 2024; published online on November 26, 2024. Discussion period open until April 26, 2025; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Infrastructure Systems***, © ASCE, ISSN 1076-0342.

with open-source data sets available from a full coverage of geographical locations and a long time span, and the results showed that the model results are comparable to those found using the models for roads/pavements in other studies. Second, modeling the whole network with large-scale pavement condition data and other openly available data sets enables better model performance and outcomes, as well as better decision-making when it comes to prioritizing maintenance and retrofit efforts. Third, we innovatively use two types of large-scale and open-source data sets: socioeconomics and built environment variables to account for the impacts of traffic activity and pavement characteristics to address the issue that these two types of variables are often not available.

Fourth, to reduce the influence of randomness in sample splitting potentially brought by the large-size data set, we introduce repeated nested cross-validation to tune and select best model. Finally, to provide more insights into how different factors affect pavement conditions individually and interactively, we introduce two interpretation tools [i.e., Shapley additive explanations (SHAP) values and accumulated local effect plots] to estimate the contributions of independent variables and their main and interactive effects on estimating pavement conditions. The results offer suggestions on efficiently allocating the limited resources in pavement maintenance projects.

The rest of the study is organized as follows. We review the studies on pavement condition predictions and related research gaps in the section of "Literature Review." We introduce our data sets and methods the section of "Data and Method." We present and discuss the results in the section of "Results." In the section of "Conclusions," we conclude this study and offer implications for pavement condition prediction in communities.

Literature Review

Pavement condition is an important index to help policymakers plan maintenance projects on the road segments and ensure the roads are reliable and safe for their users. Modeling and predicting of the pavement condition is an important but also challenging task (Sidess et al. 2020).

The collection of pavement condition data primarily employs two distinct approaches: manual walking surveys and automated surveys (Attoh-Okine and Adarkwa 2013). A manual walking survey involves trained professionals visually identifying and cataloging pavement distresses using specialized tools. This method, although reliable, can be time-consuming and subject to human error. Conversely, automated surveys involve the use of sophisticated technology to directly record pavement condition data. This method is faster and more accurate than manual surveys, offering more objective results while requiring fewer human resources.

The evolution of computer vision and deep learning technologies has significantly advanced automated surveys. These technologies not only increase the accuracy of distress identification but also reduce the overall cost of surveys by optimizing data collection and analysis processes (Xu and Zhang 2022; Sholevar et al. 2022). With a large data set of images as training set, deep learning techniques can accurately recognize the specific patterns of pavement distresses and offer an objective rating. A number of vendors offer this low-cost solution for municipalities where a regular smartphone can be mounted on the dashboard of a municipal vehicle to collect video data for processing. For instance, RoadBotics charges around 50,000 USD to measure pavement conditions for all roads in the city of Savannah, Georgia [around 1,100 km (700 mi)], which is 80,000 USD cheaper than the traditional approach of visual inspection (City of Savannah 2023; RoadBotics 2023b). This is done in 3 months, which is faster than the traditional approach (usually in 3 years) (RoadBotics 2023b).

Pavement conditions are affected by multiple factors that studies need to consider in their predictive models. Table 1 summarizes these factors used in selected studies on pavement condition predictions focusing on the areas in the North America. These variables mainly include two distinct types: internal and external factors (Xu and Zhang 2022). Internal factors include initial/current pavement condition, pavement characteristics, and age of the pavement. External factors include weather/climate, road characteristics, traffic activity, and maintenance history.

Studies usually utilized two types of data sets: the Long-Term Pavement Performance (LTPP) data set and locally collected data sets, which have several limitations. First, most of these data sets only contain information for roads of a higher classification, such as freeways and major arterial roads. They are often the focus due to their higher traffic volumes and the more significant consequences should they fail or become excessively deteriorated. However, this leaves gaps for lower-classification roads like local or rural roads. Second, many of these data sets are limited in their geographical scope, typically encompassing only a few road segments from a single location (Table 1). Although these samples provide valuable information, their limited coverage presents challenges when attempting to extrapolate findings for network-wide pavement performance predictions.

Third, some of the factors that are important for pavement condition predictions are missing in these data sets. For example, annual average daily traffic (AADT) is not available for a large portion of the road sections in LTPP (Gong et al. 2018). Finally, the availability of these data sets is often limited in smaller communities, such as townships and rural areas. These communities may have different traffic patterns, climate conditions, and maintenance practices from large metropolitan areas. The lack of comprehensive pavement condition data sets can pose significant challenges for local authorities in effectively managing their pavement infrastructures.

Empirical models, which predict the pavement conditions based on experimental and observed data, are most widely applied in literature (Sidess et al. 2020). Among the literature, empirical models could be categorized into two types: statistical models and machine learning models. Statistical models apply probability distributions, such as normal (Sadeghi et al. 2015) and ordinal logistic (Wang 2013), to estimate the relationships between factors and pavement conditions. Statistical models are easy to interpret because they provide p-values based on statistical tests. However, statistical models are not efficient to model the complex relationships between factors and pavement conditions.

Instead of probability distributions, machine learning models utilize advanced approaches to fit the data, such as decision trees (Piryonesi and El-Diraby 2020) and neural networks (Ziari et al. 2021). Machine learning models can estimate irregular nonlinear impact of factors. However, machine learning models are subject to overfitting the data. In addition, machine learning models need additional tools to interpret the results (Molnar 2020).

It is worth noting that although machine learning models provide better performance in fitting the data than statistical models, statistical models could also achieve a good performance if the studies select appropriate models and include important factors. For example, Anastasopoulos et al. (2012) applied random parameters to account for the unobserved heterogeneity across observations. In addition, Sarwar and Anastasopoulos (2016) considered previous pavement condition, which is strongly correlated with future pavement conditions and explain much of the variance. In this study, we focused on machine learning models given that they

Table 1. Model information for selected studies on future pavement condition prediction with advanced machine learning approaches in North America

References	Location	Sample size and coverage	Model	Model tuning/ selection approach	Outcome	Initial/current pavement condition	Pavement features	Age	Weather/ climate	Road characteristics	Traffic activity	Maintenance history	Model fitness
Lou et al. (2001)	Florida	Unknown, selected segments	ANN	Train-test-validation split	CI	No	Yes	Yes	No	No	No	Yes	1-year R^2: 0.9–0.91
Yang et al. (2003)	Florida	Unknown, selected segments	ANN	Train-test split	PCR	Yes	No	Yes	No	No	No	Yes	1-year R^2: 0.79–0.88, 2-year R^2: 0.55–0.76, 3-year R^2: 0.52–0.59, 4-year R^2: 0.4–0.48, 5-year R^2: 0.2–0.38
Kargah-Ostadi et al. (2010)	North America (LTPP)	1,072, selected segments	ANN	Train-test split	IRI	Yes	Yes	Yes	Yes	No	Yes	Yes	1-year R^2: 0.96
Ziari et al. (2015)	North America (LTPP)	205, selected segments	ANN	Train-test-validation split	IRI	No	Yes	Yes	Yes	No	Yes	No	1-year R^2: 0.96, 2-year R^2: 0.97
Hossain et al. (2017)	North America (LTPP)	Unknown, selected segments	ANN	Train-test split	IRI	No	No	Yes	Yes	No	Yes	No	No
Piryonesi and El-Diraby (2020)	North America (LTPP)	943, selected segments	GBM	Cross-validation	PCI	Yes	Yes	Yes	Yes	Yes	Yes	Yes	3-year classification accuracy: 82%
Barua et al. (2021)	North America (LTPP)	1,091–2,090, selected segments	GBM	Train-test-validation split	PCI	Yes	Yes	Yes	Yes	No	Yes	Yes	Runway R^2: 0.9, taxiway R^2: 0.86
Evers et al. (2023)	Colorado	Unknown, selected segments	RF	Train-test split	IRI	Yes	Yes	No	Yes	No	Yes	Yes	No
This study	Nine US communities	1,938–11,963, network-wide segments	GBM/RF	Repeated nested cross-validation	PCR	Yes	Yes (proxy)	No	Yes	Yes	Yes (proxy)	No	1-year R^2: 0.90, 2-year R^2: 0.84, 3-year R^2: 0.85, 4-year R^2: 0.78

Note: ANN = artificial neural network; GBM = gradient boosting machine; RF = random forest; CI = cracking index; PCR = pavement condition rating; IRI = International Roughness Index; AC = alligator cracking; and PCI = pavement condition index.

can account for the complex impacts of factors on pavement conditions. As indicated in Table 1, several approaches including artificial neural networks (ANN), gradient boosting machines (GBM), and random forests (RF) were used in the literature.

To tune or select the best model, researchers commonly use train-test split, train-test-validation split, and cross-validation (Table 1). Train-test split involves dividing the sample into training and testing sets to assess model performance. An improved approach is train-test-validation split, which adds a validation set for final evaluation. Cross-validation splits the sample into multiple subsets, repeatedly training and testing the model on different fold combinations. However, these methods rely on a single split, making model performance susceptible to specific data partitioning.

All studies enumerated in Table 1 used pavement condition as their dependent variable. This is intuitive because the forecasted pavement condition can be directly applied to inform decision-making processes in pavement management. However, one potential drawback of this research design lies in the disproportionate influence of the initial or current pavement condition in models predicting future pavement states. This factor often plays a dominant role, thereby overshadowing the contributions of other variables. Consequently, this might lead to a skewed perception of the relative importance of these other factors, potentially underestimating their influence on pavement deterioration over time.

For example, Piryonesi and El-Diraby (2020) estimated a GBM model to predict pavement condition index (PCI) and found that initial PCI significantly dominated the model, with all other variables contributing less than 20% of the influence of initial PCI. Evers et al. (2023) applied a RF model to predict the International Roughness Index (IRI) and had a similar finding. Estimating the change of pavement condition could help alleviate this issue when examining the contributions of factors. The change indicates the difference between the future and current pavement condition. By excluding the current pavement condition from the dependent variable in the model, the disproportionate influence of the current pavement condition on the results is minimized.

Several studies have considered temporal changes of pavement condition in their models. For example, a prevalent approach to examining pavement condition deterioration involves using Markov models and their extensions to estimate the transition probabilities that pavement condition changes from one state to a lower state during a specified time unit (Kobayashi et al. 2012; Ortiz-García et al. 2006; Madanat et al. 1995; Mizutani and Yuan 2023; Lethanh and Adey 2013). Additionally, some studies directly applied temporal changes as dependent variables in their models (Swei et al. 2018; Madanat et al. 1995). However, the primary focus of these studies has generally been on enhancing the accuracy of estimating transition probabilities rather than on exploring the impact of various factors on pavement deterioration. Furthermore, stochastic processes, such as Markov models, are limited by the scope of spatiotemporal features they can include for prediction.

Data and Method

Data

The pavement condition rating (PCR) was obtained from RoadBotics. RoadBotics uses AI-powered computer vision technologies to evaluate the performance of road infrastructure (RoadBotics 2023a). They develop mobile device applications to visualize and analyze roads and generate ratings. This approach is low-cost and easy to carry out. The governments of communities collaborate with RoadBotics to collect and manage the PCR data. Communities own their PCR data.

RoadBotics categorizes their PCR into five levels, ranging from one to five. A lower PCR indicates the corresponding road pavement condition is better. Fig. 1 presents two example road segments. Fig. 1(a) shows a road segment with its PCR = 1, which is in good condition. The road in Fig. 1(b) has PCR = 5, which has cracks and small potholes. For each road segment, RoadBotics collects ratings of multiple locations along it. Then, they calculate the average value of these ratings as the rating for the corresponding road segment. Besides PCR, RoadBotics documents the date when the PCR was collected.

The PCR data provided by RoadBotics contains road segments from multiple communities and multiple years. The data include Dormont, Robinson, Shaler, and Upper St Clair in Pennsylvania, Alma in Arkansas, Cumberland in Maryland, Sedona in Arizona, South Bend in Indiana, and Longboat Key in Florida (Fig. 2). In the US, these are in general small communities (e.g., small cities and townships), with population ranging from 5,850 to 103,353 (Table 2).

RoadBotics collected PCRs for most of the road segments for each community. However, from 2018 to 2022, RoadBotics collected data in different years for different communities because the pavement management plans of these communities are different. Dormont and Cumberland have the most temporal coverage of four years. Robinson, Upper St Clair, Alma, South Bend, and Longboat Key have the least coverage, with 2 years of data available. In addition, the coverage of road segments is slightly different across years. In total, all nine communities have 22,918 road segments across 5 years, which encompass 2,759.5 km (1,714.7 mi).

We requested weather data from the Local Climatological Data (LCD) tool hosted by the National Centers for Environmental Information (2023). The stations for data collection are located across the US, mostly in airports. We assigned the weather data

(a) (b)

Fig. 1. Two example road segments with different PCRs: (a) PCR = 1; and (b) PCR = 5. (Images © RoadBotics.)

J. Infrastruct. Syst., 2025, 31(1): 04024034

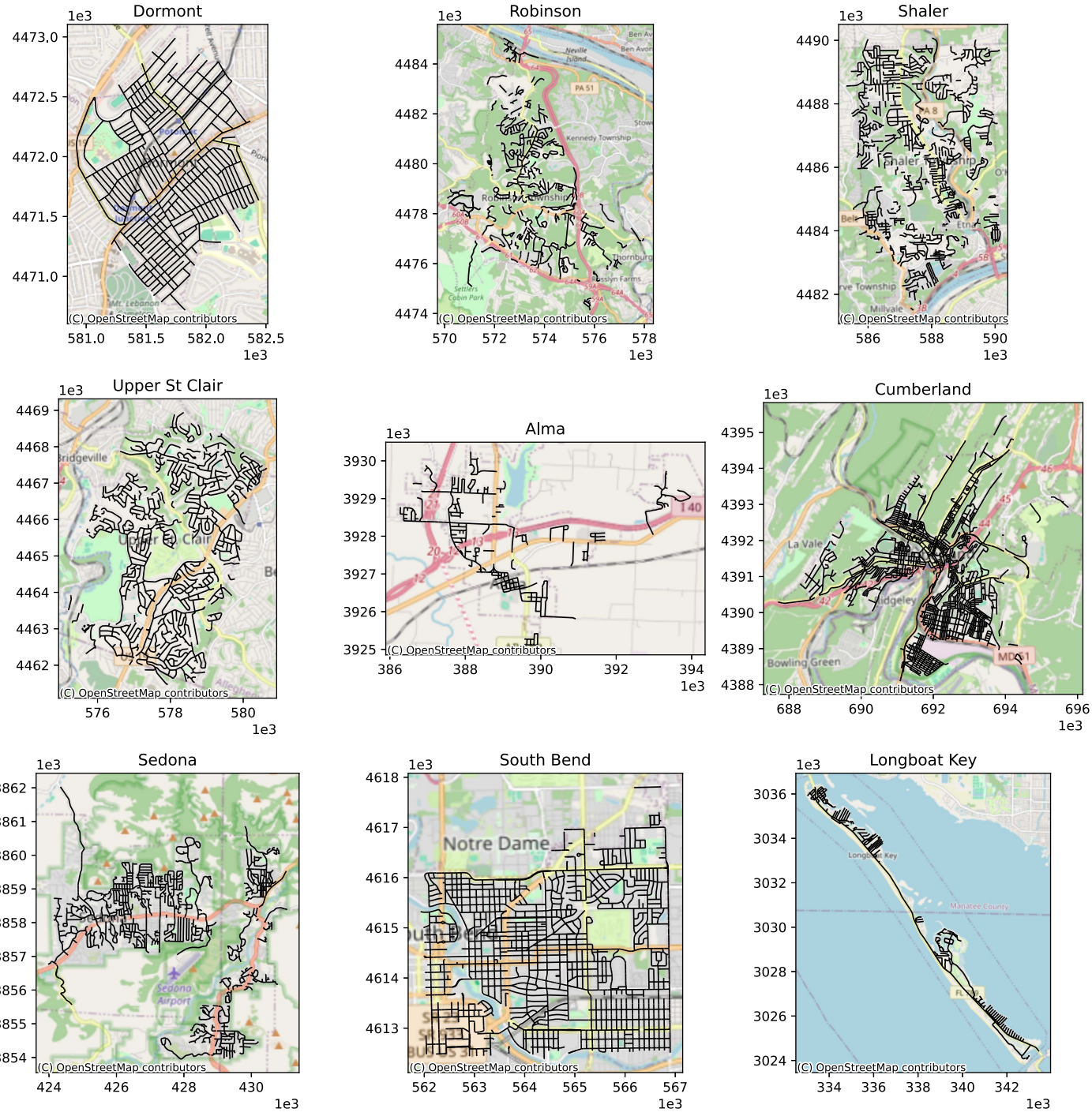

Fig. 2. Study areas. (Base maps © OpenStreetMap contributors.)

from the closest station to each community. The LCD data include hourly records of multiple measurements, such as temperature, wind speed, and precipitation. Among these variables, the impact of wind speed is often indirect compared with other weather elements like temperature and precipitation. Wind speed affects pavement condition indirectly through change the speed of water evaporation and surface temperature of the pavement. Wind speed has been shown to facilitate the water evaporation rate (McVicar et al. 2012), which could reduce the water-related damage. In addition, wind speed can decrease the surface temperature of the pavement (Qin et al. 2022), potentially leading to cracks. These relationships, although difficult to estimate in traditional statistical models, can be learned implicitly by the machine learning models applied in this study.

We extracted the road classification data from the Open Street Map (OSM) road networks (OSM 2023). We aggregated the OSM labels of road classification into three levels: primary road, secondary road, and local and other roads. Primary roads included motorway, primary, trunk, and related links. Secondary roads included secondary and secondary link. Local and other roads included all other types of roads. We then matched the road classification of OSM network to the road segments by RoadBotics.

Another important type of variable that influences pavement condition is traffic activity (or traffic load). Variables such as

J. Infrastruct. Syst., 2025, 31(1): 04024034

Table 2. Number and length of road segments with valid PCR in the communities from 2018 to 2022

Community	State	Population (year)	Number of road segments [length in km (length in mile)]				
			2018	2019	2020	2021	2022
Dormont	Pennsylvania	8,117 (2021)	470 [42.6 (26.5)]	436 [41.0 (25.5)]	467 [42.3 (26.3)]	—	482 [43.3 (26.9)]
Robinson	Pennsylvania	13,673 (2018)	—	677 [115.6 (71.8)]	695 [117.2 (72.8)]	—	—
Shaler	Pennsylvania	27,963 (2018)	1,031 [150.0 (93.2)]	1,054 [153.0 (95.1)]	1,055 [152.1 (94.5)]	—	—
Upper St Clair	Pennsylvania	19,737 (2019)	—	784 [148.7 (92.4)]	780 [148.7 (92.4)]	—	—
Alma	Arkansas	5,850 (2021)	—	264 [43.1 (26.8)]	—	—	306 [47.0 (29.2)]
Cumberland	Maryland	18,736 (2021)	2,041 [184.1 (114.4)]	2,249 [201.8 (125.4)]	—	2,102 [186.0 (115.6)]	230 [19.8 (12.3)]
Sedona	Arizona	9,763 (2021)	—	988 [139.0 (86.4)]	931 [134.5 (83.6)]	966 [136.5 (84.8)]	—
South Bend	Indiana	103,353 (2021)	—	—	2,164 [201.8 (125.4)]	2,022 [192.8 (119.8)]	—
Longboat Key	Florida	7,592 (2021)	—	375 [60.2 (37.4)]	—	349 [58.3 (36.2)]	—

AADT and equivalent single-axle load (ESAL) are frequently utilized in research to account for the impact of traffic activity. However, these types of variables are usually only observed for primary roads in major urban areas. Estimates of AADT for all roads may be available but oftentimes are inaccurate. Most communities do not have the capacity to collect and maintain this information. In addition, the road segments in the communities considered in this study are mostly local roads, which are not covered by the traffic data collection.

To account for the impact of traffic activity more effectively, we incorporated socioeconomic and built environment variables derived from two open data sources. Studies have shown that the strength of traffic activity is closely related to socioeconomics and built environment variables. For example, Anderson et al. (2006) included socioeconomic variables such as population within a half-mi buffer of a traffic count station to estimate the average daily traffic. Ewing and Cervero (2010) found that built environment variables such as street network design and accessibility are strongly correlated with vehicle distances traveled. In this study, socioeconomic variables include total population, car to work percentage, labor percentage, and median household income, which were downloaded from the National Historical GIS database hosted by University of Minnesota (Manson et al. 2023). Socioeconomics are available each year for all census block groups (CBGs) in the US.

Built environment variables include employment density, network density, job accessibility, and land-use mix, which were downloaded from the Smart Location database hosted by the EPA (2021). Built environment features are available in 2019 for all CBGs in the US. In this study, we assumed that socioeconomics changed by year. Furthermore, we assumed that built environment variables remained consistent across the years, which is justifiable considering the relatively short study period of 5 years, spanning from 2018 to 2022.

In addition to traffic activity, socioeconomic factors are also correlated with pavement characteristics. This association arises primarily because an important portion of investment into pavement infrastructure is sourced from local budgets, which are inherently linked to the socioeconomic status of the respective areas. Therefore, incorporating socioeconomic factors into pavement prediction models can help account for variations in infrastructure investment and its subsequent impact on pavement characteristics.

Table 3 lists the variables and their description, data source, and time. Some of the variables, including weather and socioeconomics, were aggregated to certain periods according to the research design, which is covered in the next section.

Method

Research Design

In this study, we predicted absolute PCR and PCR change in short term with variables including current PCR, road classification, weather, socioeconomics, built environment attributes, and several other variables. We used Eq. (1) to carry out the prediction for absolute PCR

$$\mathrm{PCR} = f(\mathrm{PCR}_0, R, W, S, B, O) \tag{1}$$

where PCR = future PCR to be estimated; and PCR_0 = current PCR. For each road segment, it has PCRs from multiple years as indicated in Table 2. Based on the present research design, one PCR was chosen as the current PCR and another PCR from a future year was chosen for prediction. For instance, a road segment in

Table 3. Variable description, source, and time

Variable	Description	Source	Time
PCR	Pavement condition rating of the road segment in the corresponding date	RoadBotics[a]	2018–2022
PCR change	Change of pavement rating condition between current and previous dates when PCR was collected	This study	
	Road classification		
Road classification (dummies)	Road classification of the road segment, including primary road, secondary road, and local/other road	Open Street Map	—
	Weather (aggregate for the days between the two dates of PCRs)		
Maximum temperature	Average maximum daily temperature (°F)	LCD[b]	2018–2022
Minimum temperature	Average minimum daily temperature (°F)		
Average temperature	Average daily temperature (°F)		
Frozen days	Number of days with daily minimum temperature lower than 0°C (32°F)		
Hot days	Number of days with daily maximum temperature higher than 30°C (86°F)		
Freeze–thaw cycles	Freeze–thaw cycles, calculated based on the half of the number that hourly temperature crossed 0°C (32°F)		
Total precipitation	Total precipitation (in.)		
Wind speed	Average daily wind speed (mi/h)		
Windy days	Number of days with daily maximum wind speed larger than 24.1 km/h (15 mi/h)		
	Socioeconomic characteristics [weighted by area percentage of CBGs within 1.61-km (1-mi) buffer of the road segment]		
Total population	Average total population of the years between the dates of the PCRs	NHGIS[c]	2018–2022
Car to work percentage	Average percentage of population commuting with car of the years between the dates of the PCRs		
Labor percentage	Average percentage of population in labor force of the years between the dates of the PCRs		
Change of total population	Difference in total population between the years of the dates of the PCRs		
Change of car to work percentage	Difference in car to work percentage between the years of the dates of the PCRs		
Change to labor percentage	Difference in labor percentage between the years of the dates of the PCRs		
Median household income	Average median household income of the between the years of the dates of the PCRs		
Change of income	Difference in median household income between the years of the dates of the PCRs		
	Built-environment characteristics [weighted by area percentage of CBGs within 1.61-km (1-mi) buffer of the road segment]		
Employment density	Gross employment density (jobs/acre) on unprotected land	EPA[d]	2019
Land-use mix	Land-use entropy calculated based on the five-tier employment (retail, industrial, service, entertainment, and office) numbers on unprotected land		
Network density	Total road network density (mi/sq mi)		
Job accessibility	Jobs within 45 min auto travel time, time-decay (network travel time) weighted		
	Other		
Community (dummies)	Community where the road is located	RoadBotics	—
Number of days	Number of days between the two dates of the PCRs	This study	

[a]RoadBotics (2023c).
[b]Local Climatological Data (National Centers for Environmental Information 2023).
[c]National Historical GIS (Manson et al. 2023).
[d]US Environmental Protection Agency Smart Location Database (EPA 2021).

Shaler has three PCRs collected in 2018, 2019, and 2020. We constructed three observations and incorporated them into the final analysis using (1) the 2018 PCR as the current PCR and the 2019 PCR as the future PCR; (2) the 2019 PCR as the current PCR and the 2020 PCR as the future PCR; and (3) the 2018 PCR as the current PCR and the 2020 PCR as the future PCR. Given this approach, for road segments with PCRs available over N years, each segment can construct C_N^2 (i.e., the number of unique combinations of two items chosen from a set of N items, also known as N choose two) observations. The key strength of this research design lies in its versatility. It can seamlessly accommodate any assortment of road segments and any selection of dates for PCR measurements.

In this study, we focused solely on observations where the current PCR was lower than the future PCR, indicating pavement deterioration. We excluded observations where the current PCR was higher than the future PCR, representing pavement improvement, to minimize the effects of unaccounted-for pavement maintenance and potential inaccuracies in PCR measurements on our estimations.

W indicates weather variables. All of these variables were aggregated over the time period between the two PCR dates (i.e., current and future PCRs) for each constructed observation, as demonstrated in the previous examples. The dates of the PCRs were originally documented by RoadBotics. Weather variables include three main types: temperature, precipitation, and wind speed. Temperature includes average maximum temperature, average minimum temperature, and average temperature, which represent the average-level temperature condition. These variables can depict the diverse climate conditions found in different communities.

However, although the time period between two PCR dates can vary significantly (e.g., ranging from approximately 1 to 4 years), these average-level temperature variables may be similar across different time periods. To address this issue, we supplemented the data set with variables reflecting the length of periods, including number of frozen days, number of hot days, and freeze–thaw cycles. Similarly, we included both average wind speed and number of windy days for wind speed. Additionally, we incorporated total precipitation into the analysis.

S represents socioeconomic attributes, including total population, car to work percentage, labor percentage, and median household income. For each variable, we included their average value, which was calculated through years between the two PCR dates and difference between the two years of the current and future PCRs.

R indicates road classification, and B indicates built environment variables. These two types of variables are assumed as consistent through years. Additionally, we incorporated two further variables into our analysis. One is community, which serves to control for the effects associated with varying contexts of communities. The other variable is the number of days between the two PCR dates, providing a measure of time elapsed between assessments.

Next, f represents the model used to estimate the relationships between future PCR and multiple types of factors. Model selection is an important task because it significantly affects the prediction performance. We illustrate the approach of model selection in the next section.

Besides the set of models to predict absolute values of future PCR, we also estimated another set of models to predict PCR change. These two sets of models included the same types of independent variables. We compared the results between these two models and expected the set of PCR change models could provide new insights into how different factors contribute to pavement condition deterioration. Table 4 presents the descriptive statistics of the variables considered in the modeling process.

Method

Within the existing literature, several models have demonstrated exceptional performance in predicting pavement conditions (Table 1), including RF, GBM, and multiple-layer perception (MLP, i.e., artificial neural network). Therefore, in this study, we employed these methods, in addition to using linear regression model (LM) as a benchmark, to conduct our estimations.

Both RF and GBM are models based on decision trees. Decision trees work by splitting the sample into subsamples according to various decision rules based on the values of independent variables, and then using the average value of the dependent variable in subsamples for predictions. One limitation of decision trees is their relatively weak predictive capability. To address this issue, there are two popular approaches: bagging and boosting. Bagging combines the results of multiple decision trees in parallel, which is applied by RF (Breiman 2001). Boosting is an iterative process that combines decision trees in a sequential order. Boosting is used in GBM (Friedman 2001, 2002). Both RF and GBM are capable of modeling complex relationships between dependent and independent variables and exhibit strong predictive performance.

MLP consists of multiple layers of interconnected nodes (i.e., artificial neurons or perceptions). These layers include an input layer, one or more hidden layers, and an output layer. The neurons of the hidden layers apply nonlinear activation functions to the results from the previous layer. These nonlinear activation functions, including logistic function, hyperbolic tangent function, and rectified linear unit (ReLU), allow MLP to model complex, nonlinear relationships in the data.

We employed repeated nested cross-validation to compare the prediction performance of the selected models. Nested cross-validation, as its name suggests, consists of two nested loops of cross-validation (Fig. 3). The inner loop aims to select the best combination of parameters for the models. The outer loop evaluates the performance of the tuned models. The nested cross-validation is repeated multiple times and the performance of the models is recorded, including R-squared, mean absolute error (MAE), and root-mean squared error (RMSE). Introducing repetition to this process could help mitigate the impact of having a specific random split of the data that can potentially bias the model performance. After the best modeling method is selected, another round of repeated cross-validation is used to search for the best combination of parameters for the selected modeling method.

In this study, we ran 10-time repeated cross-validation on the sample to select the best modeling method among LM, RF, GBM, and MLP (Fig. 3). The outer loop is fivefold cross-validation. The inner loop is threefold cross-validation and chooses the best combination of parameters based on RMSE. After the best modeling method was selected, we used 10-time repeated fivefold cross-validation to search for the best combination of parameters of this method based on RMSE.

Finally, we measured the contributions of independent variables in PCR and PCR change models. We applied the SHAP value (Chen et al. 2022), which is based on the concept of Shapely value from cooperative game theory (Shapley 1953). For each observation, the inclusion of one independent variable could have a marginal impact on the prediction of future pavement condition, and this marginal impact is calculated as the SHAP value. The average absolute SHAP value of one independent variable in all observations of the sample is the corresponding contribution. SHAP value has been widely used in explaining the contributions of independent variables in black-box machine learning methods (Xiao et al. 2021; Wagner et al. 2022).

Furthermore, we used the accumulated local effect (ALE) plots to show the main effect of important independent variables and interactive effect between current PCR and other important variables in PCR change model. ALE could present the marginal impact by one or two independent variables on the dependent variable after controlling for the effects of other independent variables (Apley and Zhu 2020; Molnar 2020). Compared with the other similar interpretation tools [e.g., partial dependence plots in Barua et al. (2021)], it can avoid the influence of multicollinearity among independent variables (Molnar 2020).

Results

Model Performance

Each set of models comprises five distinct models, categorized as all-year, 1-year, 2-year, 3-year, and 4-year models. The all-year model encompasses all observations, regardless of the duration between the current and future PCR dates. In contrast, the other four models consist of specific observation subsets. In the literature, scholars usually estimated models to predict the pavement conditions in specific years. We followed and constructed the other four models. The 1-year model contained observations with a time span of less than 1 year (365 days) between the PCR dates. The 2-year model included observations with a time period between 1 and 2 years (365–730 days). Similarly, the 3- and 4-year models incorporated observations with time spans specific to their respective categories. With these four models, we can compare our model performance with those in the literature.

Table 4. Descriptive statistics of the observations

Variable	Mean	Standard deviation	Minimum	Maximum
Future PCR	2.94	0.95	1	5
PCR change	0.55	0.45	0	3.67
		Current PCR		
Current PCR	2.39	0.9	1	5
		Road classification		
Road classification (dummies)	Primary: 3% Secondary: 6% Local and other: 92%	—	—	—
		Weather		
Maximum temperature [°C (°F)]	18.54 (65.38)	2.51 (4.51)	14.05 (57.29)	28.84 (83.92)
Minimum temperature [°C (°F)]	8.34 (47.01)	2.03 (3.65)	5.13 (41.24)	20.32 (68.59)
Average temperature [°C (°F)]	13.28 (55.91)	2.12 (3.81)	9.62 (49.31)	24.33 (75.8)
Frozen days	154	94	0	392
Hot days	86	64	17	369
Freeze–thaw cycles	275.43	190.4	0	842
Total precipitation [cm (in.)]	203.25 (80.02)	129.29 (50.9)	23.16 (9.12)	527.1 (207.52)
Wind speed [km/h (mi/h)]	10.46 (6.5)	2.33 (1.45)	7.07 (4.39)	13.52 (8.4)
Windy days	246	107	102	519
		Socioeconomic characteristics		
Total population (person)	1,330	342	420	3,017
Car to work percentage	0.83	0.09	0.46	1
Labor percentage	0.59	0.1	0.14	0.83
Median household income (USD)	63,513	22,992	30,099	169,371
Change of total population	−20	263	−1,660	956
Change of car to work percentage	−0.01	0.03	−0.34	0.15
Change of labor percentage	0	0.03	−0.13	0.12
Change of income (USD)	4,031	6,477	−58,544	45,165
		Built environment characteristics		
Employment density [jobs/ha (jobs/acre)]	5.78 (2.34)	5.39 (2.18)	0.07 (0.03)	33.33 (13.49)
Land-use mix	0.67	0.12	0.17	0.91
Network density [km/ha (mi/sq mi)]	0.11 (17.54)	0.05 (7.55)	0.01 (1.14)	0.19 (30.92)
Job accessibility (jobs)	44,052	40,831	1,434	134,332
		Other		
Community (dummies)	Dormont: 16% Robinson: 3% Shaler: 15% Upper St Clair: 4% Alma: 2% Cumberland: 37% Sedona: 11% South Bend: 12% Longboat Key: 1%	—	—	—
Number of days	697	329	325	1,500

In Table 5, we present the results of PCR models based on repeated nested cross-validation. For the all-year model, GBM had a better performance than the other three models in terms of the three measures. RMSE and MAE are presented in negative form to facilitate comparison. Its *R*-squared, RMSE, and MAE were approximately 0.85, −0.37, and −0.28, respectively.

For the 1-year model, RF had the best performance in fitting the data, with its *R*-squared value around 0.90, RMSE as −0.26, and MAE as −0.19. The performance of the 1-year model, as measured by *R*-squared, is comparable to the findings of several prior studies. For instance the ANN models used by Lou et al. (2001) to predict the cracking index in 1 year had a *R*-squared value ranging from 0.9 to 0.91. The ANN models to predict the PCR by Yang et al. (2003) exhibited *R*-squared values ranging from 0.79 to 0.88.

For the 2-year model, GBM was better than the other models. It had an *R*-squared of 0.84, RMSE of −0.36, and MAE of −0.27. The results were mixed when comparing our RF 2-year model with similar models in other studies. For example, the *R*-squared value of our model was much higher than that of the ANN model to predict PCR by Yang et al. (2003), ranging from 0.55 to 0.76. However, the ANN model by Ziari et al. (2015) to predict the international roughness index in 2 years had a larger *R*-squared (0.97) than our model.

GBM also performed better than other models in the 3- and 4-year models. The GBM 3-year model had an R-squared of 0.85, RMSE of −0.39, and MAE of −0.30. The GBM 4-year model had an *R*-squared of 0.78, RMSE of −0.47, and MAE of −0.36. In terms of *R*-squared, the two models had a better fitness than the ANN models estimated by Yang et al. (2003). The *R*-squared values were 0.4 to 0.48 and 0.2 to 0.38 for their 3- and 4-year models, respectively.

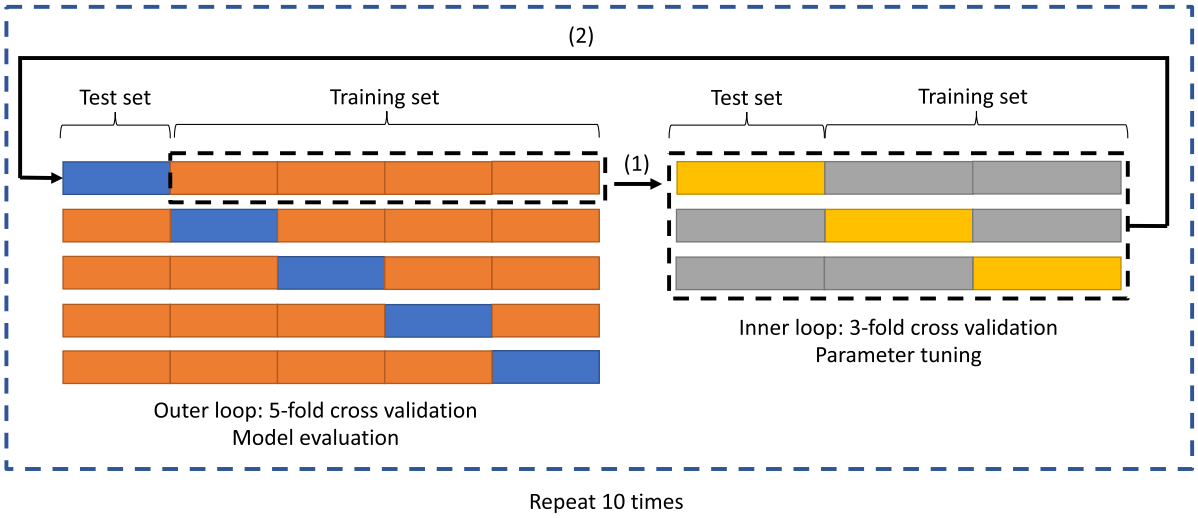

Fig. 3. Repeated nested cross-validation.

Table 6 presents the results of PCR change models. GBM had a better performance than other models for the all-year model. However, RF performed better than other models for the 1-, 2-, 3-, and 4-year models. The GBM all-year model had an R-squared of 0.34, RMSE of −0.37, and MAE of −0.28. The R-squared of the other four models ranged from 0.25 to 0.31. The corresponding RMSE ranged from −0.46 to −0.26. The corresponding MAE ranged from −0.35 to −0.19. In general, the fitness of PCR change models was lower than the PCR models, which is consistent with our expectation. In PCR models, most of the variance in future PCRs could be explained by the current PCRs, resulting in a higher goodness of fit.

Visualization of the errors in the road network is another way to demonstrate the model performance. Due to limitation of space, we used Dormont in Pennsylvania as an example. Figs. 4 and 5 present the distribution of prediction error across selected periods (1 and 4 years) in Dormont for all-year PCR model and PCR change model, respectively.

Consistent with the results in Tables 5 and 6, as the projected period lengthened, the prediction error correspondingly rose. For

Table 5. Model performance based on repeated nested cross-validation (PCR)

	R-squared		RMSE		MAE	
Model	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
			All-year predication (N = 11,963)			
LM	0.8100	0.0002	−0.4149	0.0002	−0.3295	0.0001
RF	0.8391	0.0005	−0.3817	0.0006	−0.2876	0.0005
GBM	**0.8491**	**0.0011**	**−0.3696**	**0.0012**	**−0.2790**	**0.0011**
MLP	0.8308	0.0011	−0.3914	0.0013	−0.3035	0.0011
			One-year predication (N = 1,024)			
LM	0.8636	0.0143	−0.3053	0.0130	−0.2453	0.0022
RF	**0.8980**	**0.0031**	**−0.2644**	**0.0043**	**−0.1875**	**0.0029**
GBM	0.8960	0.0019	−0.2669	0.0024	−0.1923	0.0020
MLP	0.8737	0.0062	−0.2943	0.0055	−0.2353	0.0022
			Two-year prediction (N = 5,763)			
LM	0.8074	0.0003	−0.3901	0.0002	−0.3146	0.0002
RF	0.8362	0.0013	−0.3595	0.0014	−0.2729	0.0013
GBM	**0.8394**	**0.0008**	**−0.3559**	**0.0010**	**−0.2713**	**0.0007**
MLP	0.8247	0.0015	−0.3721	0.0015	−0.2928	0.0013
			Three-year prediction (N = 3,238)			
LM	0.8241	0.0008	−0.4130	0.0007	−0.3251	0.0007
RF	0.8428	0.0016	−0.3904	0.0021	−0.2955	0.0015
GBM	**0.8461**	**0.0015**	**−0.3864**	**0.0017**	**−0.2928**	**0.0014**
MLP	0.8414	0.0007	−0.3923	0.0009	−0.3037	0.0007
			Four-year prediction (N = 1,938)			
LM	0.7407	0.0016	−0.5089	0.0017	−0.4025	0.0014
RF	0.7748	0.0030	−0.4743	0.0027	−0.3643	0.0020
GBM	**0.7811**	**0.0035**	**−0.4676**	**0.0035**	**−0.3602**	**0.0034**
MLP	0.7770	0.0019	−0.4717	0.0020	−0.3636	0.0020

Note: Bold values indicates that the corresponding model has the best performance.

Table 6. Model performance based on repeated nested cross-validation (PCR change)

Model	*R*-squared		RMSE		MAE	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
			All-year predication ($N = 11{,}963$)			
LM	0.1665	0.0008	−0.4149	0.0002	−0.3295	0.0001
RF	0.3135	0.0021	−0.3765	0.0006	−0.2881	0.0004
GBM	**0.3402**	**0.0041**	**−0.3692**	**0.0011**	**−0.2806**	**0.0013**
MLP	0.2674	0.0045	−0.3889	0.0012	−0.3010	0.0018
			One-year predication ($N = 1{,}024$)			
LM	0.0589	0.1049	−0.3053	0.0130	−0.2453	0.0022
RF	**0.3134**	**0.0189**	**−0.2623**	**0.0034**	**−0.1924**	**0.0029**
GBM	0.2918	0.0132	−0.2659	0.0026	−0.2002	0.0024
MLP	0.1580	0.0133	−0.2903	0.0021	−0.2318	0.0015
			Two-year prediction ($N = 5{,}763$)			
LM	0.0881	0.0010	−0.3901	0.0002	−0.3146	0.0002
RF	**0.2537**	**0.0030**	**−0.3529**	**0.0006**	**−0.2717**	**0.0005**
GBM	0.2489	0.0041	−0.3540	0.0010	−0.2731	0.0007
MLP	0.1767	0.0047	−0.3706	0.0010	−0.2915	0.0018
			Three-year prediction ($N = 3{,}238$)			
LM	0.1214	0.0038	−0.4130	0.0007	−0.3251	0.0007
RF	**0.2492**	**0.0037**	**−0.3818**	**0.0010**	**−0.2913**	**0.0009**
GBM	0.2453	0.0086	−0.3827	0.0021	−0.2920	0.0020
MLP	0.2068	0.0076	−0.3922	0.0016	−0.3042	0.0018
			Four-year prediction ($N = 1{,}938$)			
LM	0.0851	0.0070	−0.5089	0.0017	−0.4025	0.0014
RF	**0.2616**	**0.0060**	**−0.4573**	**0.0018**	**−0.3522**	**0.0011**
GBM	0.2519	0.0123	−0.4600	0.0035	−0.3546	0.0027
MLP	0.2229	0.0103	−0.4691	0.0030	−0.3618	0.0026

Note: Bold values indicates that the corresponding model has the best performance.

instance, within the PCR model, the inaccuracies observed for projections set 1 year in the future were significantly less pronounced than those for projections set 4 years ahead. In addition, some road segments in the lower part of the area had larger errors in both PCR and PCR change models, potentially due to certain variables missing in the models.

Variable Contributions

We estimated the contributions of independent variables to predicting future PCR and PCR change in all-year models, respectively. Fig. 6 presents the top 10 independent variables based on their contributions (i.e., average absolute SHAP value) to estimating future PCR. Current PCR was the dominant factor in predicting future PCR, and its contribution was much larger than other variables. This is consistent with previous studies (Piryonesi and El-Diraby 2020; Evers et al. 2023; Gong et al. 2018). Its average absolute SHAP value was close to 0.7, indicating that inclusion of current PCR in the all-year model could have an average marginal impact of 0.7 on the future PCR.

The two variables following current PCR were freeze–thaw cycles and frozen days. Their average absolute SHAP values were 0.05 and 0.04, respectively. The other important variables included number of days, change of income, land-use mix, wind speed, median household income, minimum temperature, and network density. Their contributions were similar, with the average absolute SHAP value ranging from 0.015 to 0.022. As illustrated in Fig. 6, the prominent role of current PCR could bias people's perceptions of the contributions made by other factors. Statistically, current PCR is highly correlated with future PCR, and thus it accounts for most of the variance in future PCR predictions.

We present the top 10 independent variables employed in the GBM all-year model, ranked based on their contributions to predicting PCR changes, in Fig. 7. Current PCR was still the most important variable based on its average absolute SHAP value (approximately 0.09). Nonetheless, in comparison with the GBM all-year PCR model, current PCR did not show the same dominant role in influencing predictions. The following four variables were frozen days (0.06), freeze–thaw cycles (0.03), and number of days (0.03). The other important variables included minimum temperature, land-use mix, local or other road, change of income, wind speed, and network density. They shared a similar average absolute SHAP value of 0.02. Although most of these 10 variables are the same as those in the all-year PCR model, their rankings are different.

Main and Interactive Effects

As shown in the preceding sections, the results of PCR model potentially bias people's impression of other variables' contribution to predicting pavement conditions and PCR change model helps fixing this issue. Therefore, when discussing the main and interactive effects of important variables on pavement conditions, we focused on the PCR change model. Fig. 8 presents the main effect of top seven important variables. Fig. 9 presents the interactive effect between current PCR and other six important variables. The interactive effect does not include the main effect and only indicates the marginal effect resulting from the interaction between two independent variables. To facilitate comparison, we used the same scale for the y-axes of the main effect plots and for the z-axes of the interactive effect plots, respectively. In main effect plots, we used rugs on the x-axes to present the distribution of the corresponding independent variables.

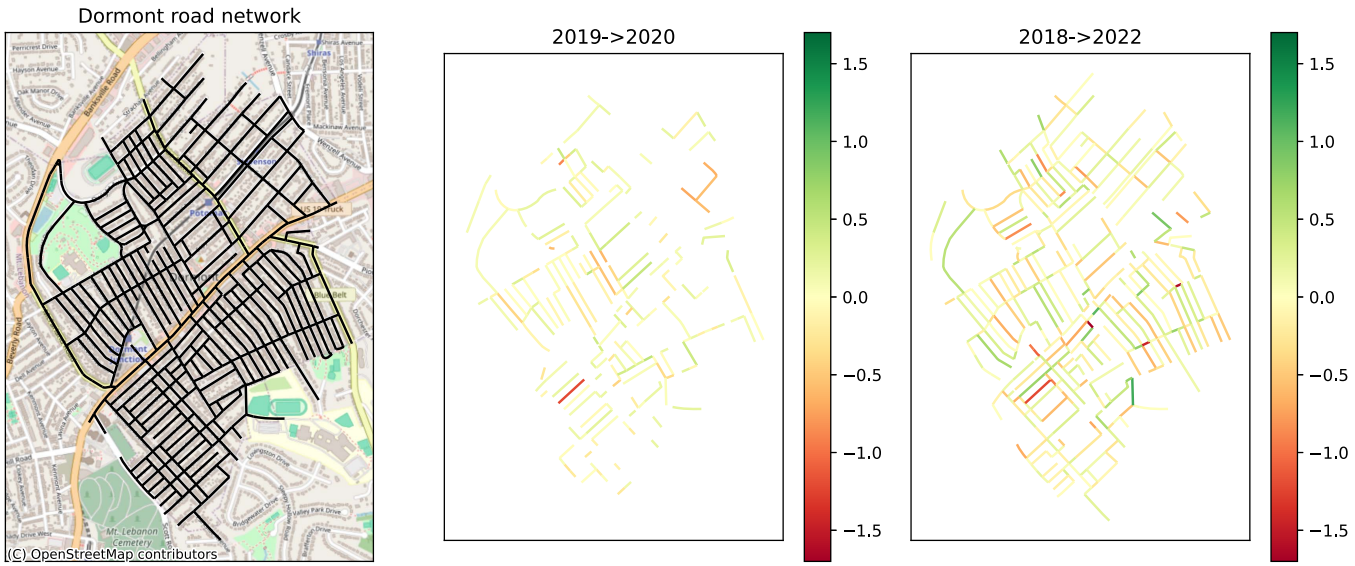

Fig. 4. Distribution of prediction error for the road segments across selected periods in Dormont, Pennsylvania (all-year PCR model). (Base maps © OpenStreetMap contributors.)

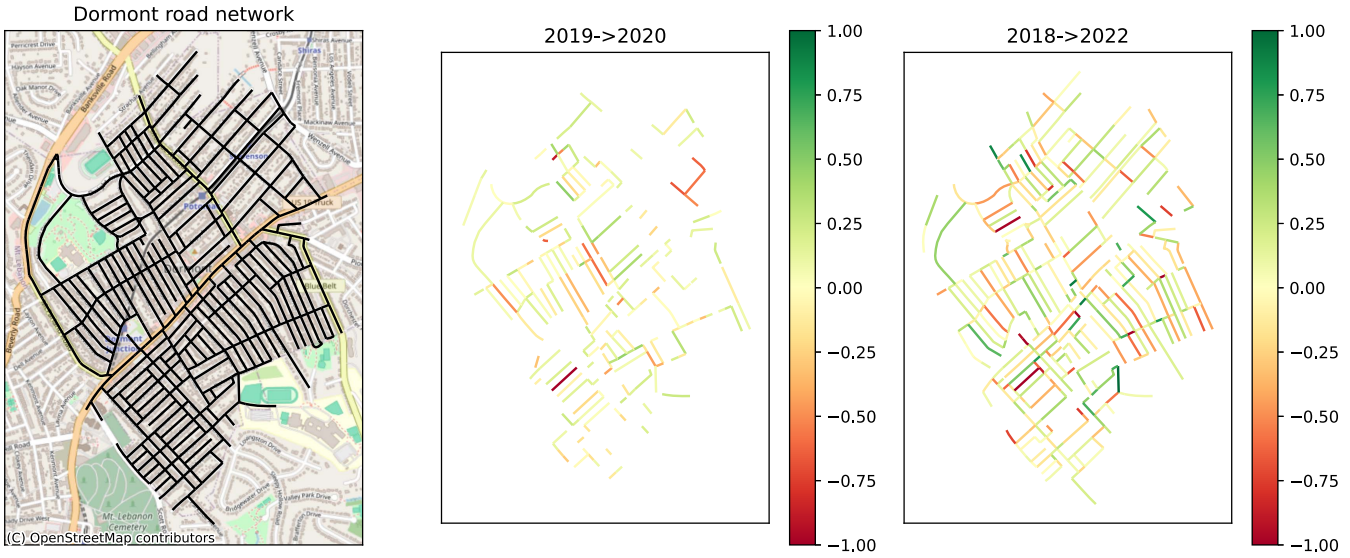

Fig. 5. Distribution of prediction error for the road segments across selected periods in Dormont, Pennsylvania (all-year PCR change model).

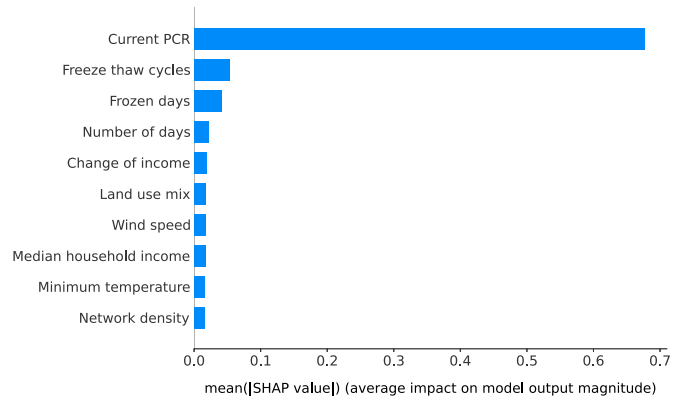

Fig. 6. Average absolute SHAP value of independent variables considered in GBM all-year model to predict PCR.

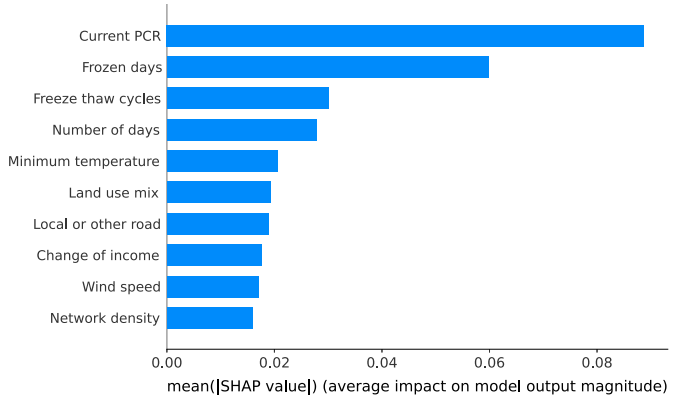

Fig. 7. Average absolute SHAP value of independent variables considered in GBM all-year model to predict PCR change.

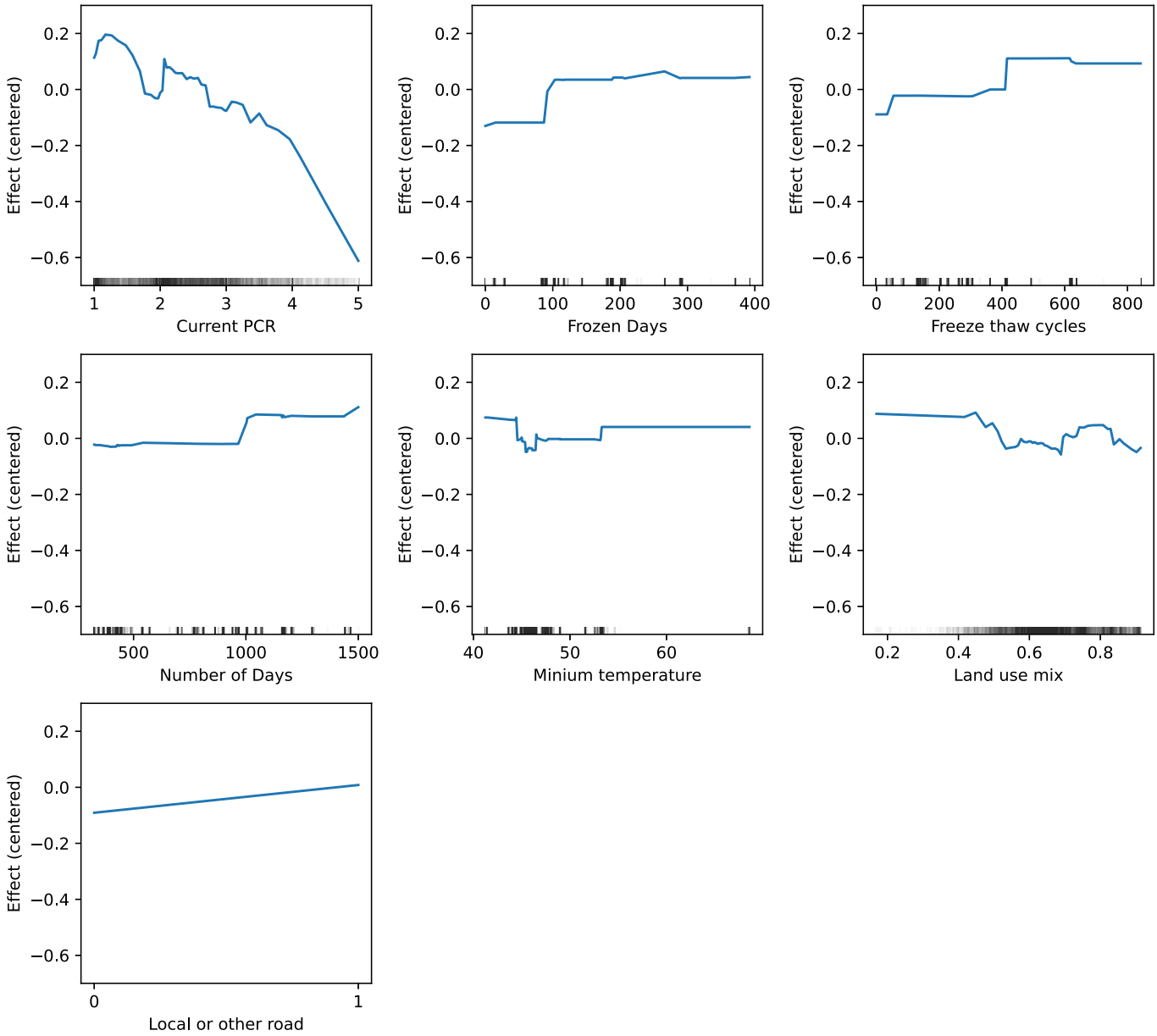

Fig. 8. Main effects of important variables in all-year PCR change model.

As shown in Fig. 8, current PCR was negatively correlated with PCR change. In addition, frozen days, number of days, and freeze–thaw cycles had positive relationships with PCR change. Minimum temperature and land-use mix were negatively correlated with PCR change. Local or other road was associated with higher PCR change than other road classifications. These relationships are consistent with our expectations.

In Fig. 9, current PCR and frozen days showed strong interactive effects. When current PCR and frozen days were low, there was a positive impact on PCR change. When current PCR was higher than four and frozen days was larger than about 100 days, there was a negative impact on PCR change. Regarding the interactive impact between current PCR and freeze–thaw cycles, there was a negative impact on PCR change where both current PCR and freeze–thaw cycles were high. The interactive effect between current PCR and number of days showed a similar pattern. For other variables, their interactive effects were mostly zero or did not suggest a clear pattern.

Conclusions

In this study, we proposed a low-cost, network-wide, and ubiquitous approach to predict the pavement condition in nine communities in the US. Specifically, we applied the PCR data collected by AI-powered computer vision technologies from RoadBotics and several types of variables from open-sourced data sets, including road classification, weather, socioeconomics, and built environment variables. We used several machine learning models, including RF, GBM, and MLP, to estimate two sets of models to predict absolute PCR and PCR change, respectively. The results of this study contribute significantly to literature.

This study utilized several open-sourced data sets that cover most of the areas in the US and multiple years, and the PCR models estimated in this study showed a good performance in predicting future PCRs. Based on the results of repeated nested cross-validation, GBM was selected for the all-year PCR model. RF was selected for the 1-year PCR model, and GBM was selected for the

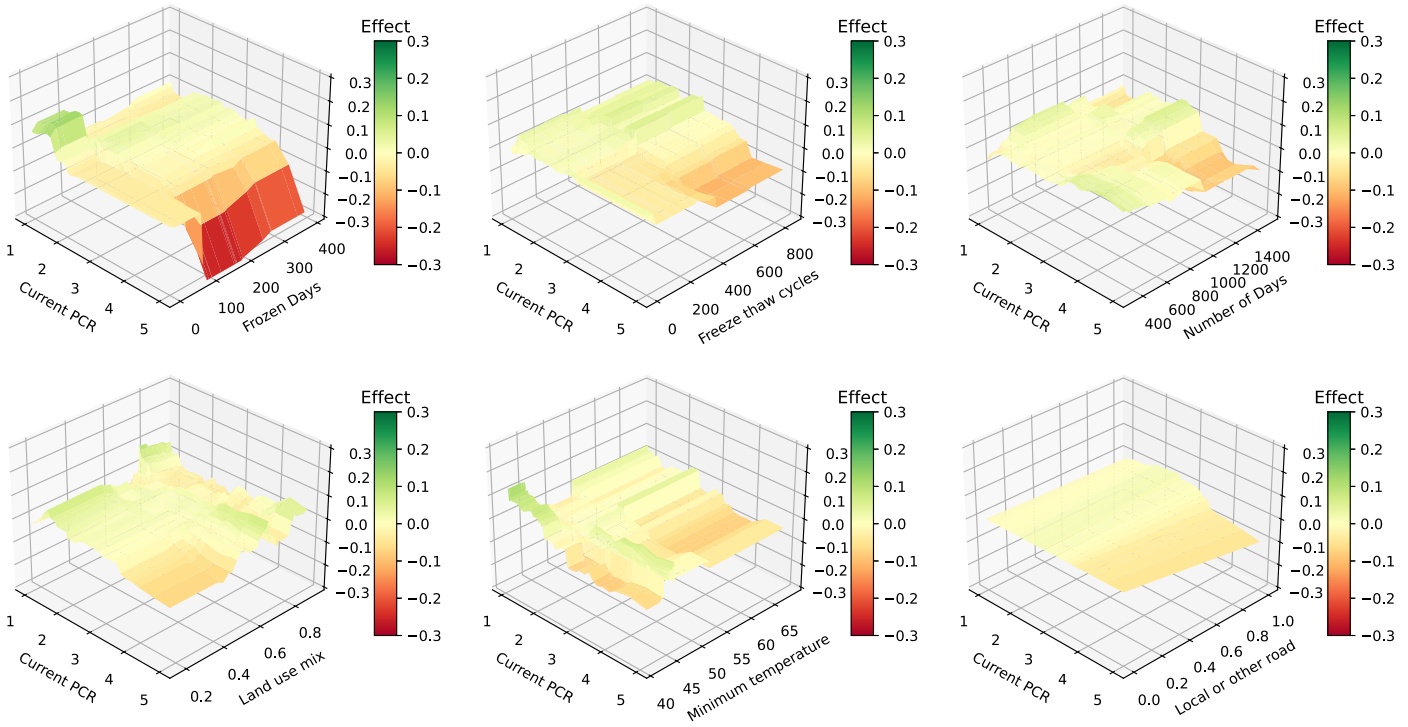

Fig. 9. Interactive effects between current PCR and other important variables in all-year PCR change model.

2-, 3-, and 4-year models. Compared with the similar models in the literature, the 1-year model had a comparable performance, the 2-year model had mixed results, and the 3- and 4-year models had better performance in terms of the values of *R*-squared. In terms of the contributions of single independent variables, our all-year model showed that current PCR has a dominant role in estimating future PCRs. This result is also consistent with the results in other studies.

We included socioeconomic variables from National Historical GIS and built environment attributes from EPA Smart Location Database to account for the impact of traffic activity and pavement characteristics, and some of these variables showed good contributions to model performance. In both all-year PCR and PCR change models, land-use mix, network density, and change of income were among the top 10 independent variables according to their contributions. In addition, median household income was among the top 10 variables in all-year PCR. These results suggest that these variables are effective in estimating either PCR and PCR change in addition to variables usually considered in the models such as current PCR, weather, and road classification.

Besides the PCR models, we estimated another set of models to predict the PCR change. In PCR models, current PCR explains most of the variance in future PCR, and the contributions of other variables become trivial. Modeling PCR change avoids this issue and helps demonstrating the contributions of other variables. The all-year PCR change model indicated that, although current PCR made the most contribution, other variables, including frozen days, freeze–thaw cycles, and number of days, had substantial contributions to predicting PCR change.

The main and interactive effects showed that the variables not only affect PCR change independently but also manifest salient effects when interacting with current PCR. The interaction between current PCR and frozen days increased PCR change when both variables were low and decreased PCR change when both two variables were higher than certain thresholds. There was a positive impact on PCR change when both current PCR and freeze–thaw cycles exceed certain thresholds. The interaction between PCR change and number of days showed a similar effect. The findings suggest that frozen days will be more harmful to pavement conditions when the road segments are in relatively good condition. However, when the road segments are in poor condition, frozen days, freeze–thaw cycles, and number of days seem to have diminished impact.

This study provides four important implications for pavement management in communities. First, leveraging various data sets from open sources could address the data shortage faced by most communities. Open-source data provide a wealth of information to supplement and enhance the pavement monitoring process. Also, by leveraging open-source data, which have large spatial and temporal coverage, communities can tap into a wider range of information and improve the accuracy of their pavement assessments. This, in turn, enables better decision-making when it comes to prioritizing maintenance and repair efforts, ultimately leading to more efficient allocation of resources.

Second, this study demonstrates that the results of modeling PCR and PCR change both have their unique advantages and could complement each other. Modeling pavement condition directly is easy to interpret, and the results could be applied directly to prioritize road segments for maintenance. Modeling the change of pavement conditions, on the other hand, is more sensitive to changes in the factors and provides additional insights on how factors influence pavement conditions. Their results could help understand which factors contribute more to pavement condition deterioration and thus allocate resources to the road sections influenced most by these factors. Decision makers could benefit from both models for pavement management.

Third, pavement management policies need to consider not just the individual effects of factors, but also their interactive impacts, which helps better allocating the limited resources. For example, when pavement is in good condition, more aggressive maintenance

measures could be employed to reduce their exposure to frozen environments. This could involve increasing the frequency of inspections during cold weather periods or utilizing more resistant materials during initial construction. On the other hand, when pavement conditions are poor, our results showed that these variables have a diminished impact. This might suggest a more conservative use of resources. Considering the diminished effect of frozen days and freeze–thaw cycles under poor conditions, these periods might not warrant an immediate intensive repair, but rather a scheduled complete rehabilitation.

Finally, the proposed model could enhance climate resiliency and transportation equity during pavement management process. With this model, we can predict pavement performance with hypothetical climate scenarios, encompassing a wide range of temperatures, freeze–thaw cycles, wind speeds, and precipitation levels. By accurately predicting pavement conditions on a yearly basis, it facilitates more precise estimation of maintenance and operation costs in the future. Consequently, this model not only heightens awareness regarding the need for climate change adaptation but also aids in more informed budget allocation for roadway construction projects. Moreover, considering the pavement condition across the entire community ensures the inclusion of all populations, resulting in more equitable outcomes in pavement management decision-making.

This study presents two limitations. Firstly, factors such as age and maintenance history were not incorporated into the analysis due to the unavailability of related data sets for the selected communities. Because including these factors could further enhance model performance, we encourage future studies to consider them when the necessary data sets become available. Secondly, we combined observations from all communities into one model rather than modeling them individually. This decision was made because the temporal coverage of PCR data varied between communities, making separate models incomparable and unable to provide additional insights. For future research, we recommend developing individual models when the temporal data coverage is consistent across all communities.

Data Availability Statement

All data utilized in this study have been responsibly attributed to their respective sources in the article. At the same time, we shared the processed data sets and codes through the GitHub repository: https://github.com/vtao1989/AssetPerformance. The pavement condition data were generously provided by RoadBotics. It is important to note that these raw data cannot be shared with third parties.

Acknowledgments

This project was funded by Pennsylvania Infrastructure Technology Alliance (PITA). The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. The authors would like to thank Dr. Christoph Mertz at RoadBotics by Michelin to provide data and feedback to this research.

References

Anastasopoulos, P. C., F. L. Mannering, and J. E. Haddock. 2012. "Random parameters seemingly unrelated equations approach to the postrehabilitation performance of pavements." *J. Infrastruct. Syst.* 18 (3): 176–182. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000096.

Anderson, M. D., K. Sharfi, and S. E. Gholston. 2006. "Direct demand forecasting model for small urban communities using multiple linear regression." *Transp. Res. Rec.* 1981 (1): 114–117. https://doi.org/10.1177/0361198106198100117.

Apley, D. W., and J. Zhu. 2020. "Visualizing the effects of predictor variables in black box supervised learning models." *J. R. Stat. Soc.: Ser. B* 82 (4): 1059–1086. https://doi.org/10.1111/rssb.12377.

Attoh-Okine, N., and O. Adarkwa. 2013. *Pavement condition surveys—Overview of current practices*. Newark, DE: Delaware Center for Transportation.

Barua, L., B. Zou, M. Noruzoliaee, and S. Derrible. 2021. "A gradient boosting approach to understanding airport runway and taxiway pavement deterioration." *Int. J. Pavement Eng.* 22 (13): 1673–1687. https://doi.org/10.1080/10298436.2020.1714616.

Breiman, L. 2001. "Random forests." *Mach. Learn.* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Chen, H., S. M. Lundberg, and S.-I. Lee. 2022. "Explaining a series of models by propagating Shapley values." *Nat. Commun.* 13 (1): 4512. https://doi.org/10.1038/s41467-022-31384-3.

City of Savannah. 2023. "Road assessment purchasing summary." Accessed March 8, 2023. https://agenda.savannahga.gov/content/files/road-assessment-purchasing-summary.pdf.

EPA. 2021. "Smart location database." Accessed March 5, 2023. https://www.epa.gov/smartgrowth/smart-location-mapping#SLD.

Evers, E., M. Z. Bashar, and C. Torres-Machi. 2023. *Improving the prediction of full-depth reclamation pavements performance using random forests*. Washington, DC: Transportation Research Board of the National Academies.

Ewing, R., and R. Cervero. 2010. "Travel and the built environment." *J. Am. Plann. Assoc.* 76 (3): 265–294. https://doi.org/10.1080/01944361003766766.

Friedman, J. H. 2001. "Greedy function approximation: A gradient boosting machine." *Ann. Stat.* 29 (5): 1189–1232. https://doi.org/10.1214/aos/1013203451.

Friedman, J. H. 2002. "Stochastic gradient boosting." *Comput. Stat. Data Anal.* 38 (4): 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Gong, H., Y. Sun, X. Shu, and B. Huang. 2018. "Use of random forests regression for predicting IRI of asphalt pavements." *Constr. Build. Mater.* 189 (Nov): 890–897. https://doi.org/10.1016/j.conbuildmat.2018.09.017.

Hossain, M. I., L. S. P. Gopisetti, and M. S. Miah. 2017. "Prediction of international roughness index of flexible pavements from climate and traffic data using artificial neural network modeling." In *Proc., Airfield and Highway Pavements 2017*. Reston, VA: ASCE.

Kargah-Ostadi, N., S. M. Stoffels, and N. Tabatabaee. 2010. "Network-level pavement roughness prediction model for rehabilitation recommendations." *Transp. Res. Rec.* 2155 (1): 124–133. https://doi.org/10.3141/2155-14.

Kobayashi, K., K. Kaito, and N. Lethanh. 2012. "A statistical deterioration forecasting method using hidden Markov model for infrastructure management." *Transp. Res. Part B: Methodol.* 46 (4): 544–561. https://doi.org/10.1016/j.trb.2011.11.008.

Lethanh, N., and B. T. Adey. 2013. "Use of exponential hidden Markov models for modelling pavement deterioration." *Int. J. Pavement Eng.* 14 (7): 645–654. https://doi.org/10.1080/10298436.2012.715647.

Lou, Z., M. Gunaratne, J. J. Lu, and B. Dietrich. 2001. "Application of neural network model to forecast short-term pavement crack condition: Florida case study." *J. Infrastruct. Syst.* 7 (4): 166–171. https://doi.org/10.1061/(ASCE)1076-0342(2001)7:4(166).

Madanat, S., R. Mishalani, and W. H. Ibrahim. 1995. "Estimation of infrastructure transition probabilities from condition rating data." *J. Infrastruct. Syst.* 1 (2): 120–125. https://doi.org/10.1061/(ASCE)1076-0342(1995)1:2(120).

Manson, S., J. Schroeder, D. Van Riper, K. Knowles, T. Kugler, F. Roberts, and S. Ruggles. 2023. *IPUMS national historical geographic information system: Version 18.0 [dataset]*. Minneapolis: IPUMS Center for Data Integration. https://doi.org/10.18128/D050.V18.0.

McVicar, T. R., et al. 2012. "Global review and synthesis of trends in observed terrestrial near-surface wind speeds: Implications for evaporation." *J. Hydrol.* 416 (Jan): 182–205. https://doi.org/10.1016/j.jhydrol.2011.10.024.

Mizutani, D., and X.-X. Yuan. 2023. "Infrastructure deterioration modeling with an inhomogeneous continuous time Markov chain: A latent state approach with analytic transition probabilities." *Comput.-Aided Civ. Infrastruct. Eng.* 38 (13): 1730–1748. https://doi.org/10.1111/mice.12976.

Molnar, C. 2020. "Interpretable machine learning—A guide for making black box models explainable." Accessed March 12, 2023. https://christophm.github.io/interpretable-ml-book/.

National Centers for Environmental Information. 2023. "Local climatological data." Accessed March 5, 2023. https://www.ncei.noaa.gov/products/land-based-station/local-climatological-data.

Ortiz-García, J. J., S. B. Costello, and M. S. Snaith. 2006. "Derivation of transition probability matrices for pavement deterioration modeling." *J. Transp. Eng.* 132 (2): 141–161. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:2(141).

OSM (Open Street Map). 2023. "Geofabrik download server." Accessed March 5, 2023. https://download.geofabrik.de/north-america.html.

Piryonesi, S. M., and T. El-Diraby. 2020. "Data analytics in asset management: Cost-effective prediction of the pavement condition index." *J. Infrastruct. Syst.* 26 (1): 04019036. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000512.

Piryonesi, S. M., and T. El-Diraby. 2021. "Climate change impact on infrastructure: A machine learning solution for predicting pavement condition index." *Constr. Build. Mater.* 306 (Nov): 124905. https://doi.org/10.1016/j.conbuildmat.2021.124905.

Qin, Y., X. Zhang, K. Tan, and J. Wang. 2022. "A review on the influencing factors of pavement surface temperature." *Environ. Sci. Pollut. Res. Int.* 29 (45): 67659–67674. https://doi.org/10.1007/s11356-022-22295-3.

RoadBotics. 2023a. "About RoadBotics by Michelin." Accessed March 13, 2023. https://www.roadbotics.com/company-about/.

RoadBotics. 2023b. "Case study: The City of Savannah, Georgia." Accessed March 8, 2023. https://www.roadbotics.com/city-of-savannah-case-study/.

RoadBotics. 2023c. "RoadBotics is now Michelin Better roads." Accessed March 11, 2023. https://www.roadbotics.com.

Sadeghi, J., E. R. Najafabadi, and M. E. Kaboli. 2015. "Development of degradation model for urban asphalt pavement." *Int. J. Pavement Eng.* 18 (8): 659–667. https://doi.org/10.1080/10298436.2015.1095912.

Sarwar, M. T., and P. C. Anastasopoulos. 2016. "Three-stage least squares analysis of postrehabilitation pavement performance." *Transp. Res. Rec.* 2589 (1): 97–109. https://doi.org/10.3141/2589-11.

Shapley, L. S. 1953. "A value for n-person games." In *Contributions to the theory of games, volume II, annals of mathematics studies*, 307–317. Princeton, NJ: Princeton University Press.

Sholevar, N., A. Golroo, and S. R. Esfahani. 2022. "Machine learning techniques for pavement condition evaluation." *Autom. Constr.* 136 (Apr): 104190. https://doi.org/10.1016/j.autcon.2022.104190.

Sidess, A., A. Ravina, and E. Oged. 2020. "A model for predicting the deterioration of the pavement condition index." *Int. J. Pavement Eng.* 22 (13): 1625–1636. https://doi.org/10.1080/10298436.2020.1714044.

Swei, O., J. Gregory, and R. Kirchain. 2018. "Does pavement degradation follow a random walk with drift? Evidence from variance ratio tests for pavement roughness." *J. Infrastruct. Syst.* 24 (4): 04018027. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000450.

Wagner, F., N. Milojevic-Dupont, L. Franken, A. Zekar, B. Thies, N. Koch, and F. Creutzig. 2022. "Using explainable machine learning to understand how urban form shapes sustainable mobility." *Transp. Res. Part D: Transp. Environ.* 111 (Oct): 103442. https://doi.org/10.1016/j.trd.2022.103442.

Wang, Y. 2013. "Ordinal logistic regression model for predicting AC overlay cracking." *J. Perform. Constr. Facil.* 27 (3): 346–353. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000327.

Xiao, L., S. Lo, J. Liu, J. Zhou, and Q. Li. 2021. "Nonlinear and synergistic effects of TOD on urban vibrancy: Applying local explanations for gradient boosting decision tree." *Sustainable Cities Soc.* 72 (Sep): 103063. https://doi.org/10.1016/j.scs.2021.103063.

Xu, Y., and Z. Zhang. 2022. "Review of applications of artificial intelligence algorithms in pavement management." *J. Transp. Eng., Part B: Pavements* 148 (3): 03122001. https://doi.org/10.1061/jpeodx.0000373.

Yang, J., J. J. Lu, M. Gunaratne, and Q. Xiang. 2003. "Forecasting overall pavement condition with neural networks: Application on Florida highway network." *Transp. Res. Rec.* 1853 (1): 3–12. https://doi.org/10.3141/1853-01.

Ziari, H., H. Fazaeli, S. J. Vaziri Kang Olyaei, and M. A. Ziari. 2021. "Evaluation of effects of temperature, relative humidity, and wind speed on practical characteristics of plastic shrinkage cracking distress in concrete pavement using a digital monitoring approach." *Int. J. Pavement Res. Technol.* 15 (1): 138–158. https://doi.org/10.1007/s42947-021-00016-2.

Ziari, H., J. Sobhani, J. Ayoubinejad, and T. Hartmann. 2015. "Prediction of IRI in short and long terms for flexible pavements: ANN and GMDH methods." *Int. J. Pavement Eng.* 17 (9): 776–788. https://doi.org/10.1080/10298436.2015.1019498.