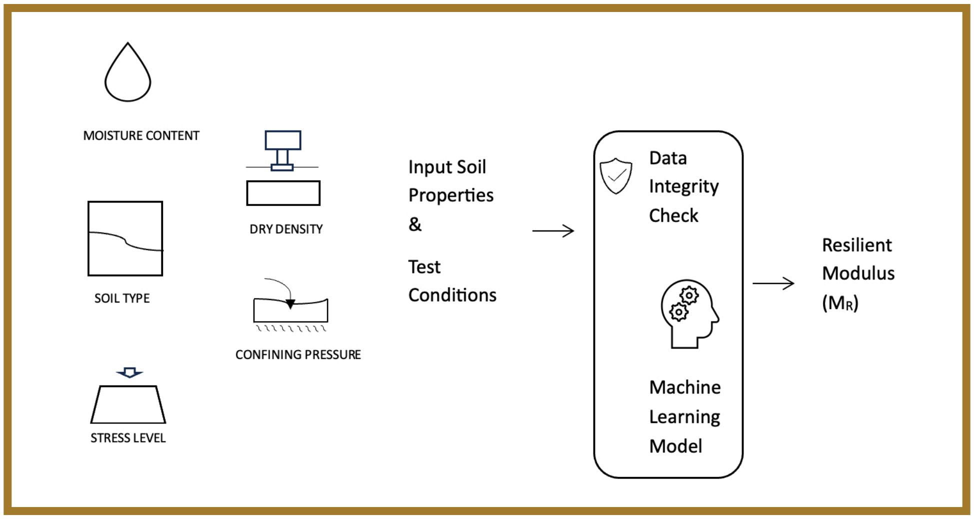
JOINT TRANSPORTATION
RESEARCH PROGRAM
INDIANA DEPARTMENT OF
TRANSPORTATION AND PURDUE UNIVERSITY

Use of Machine Learning Methods to Obtain a Reliable Predictive Model for Resilient Modulus of Subgrade Soil



Sara Khoshnevisan, Mehdi Norouzi, Laith Sadik

SPR-4714 • Report Number: FHWA/IN/JTRP-2024/27 • DOI: 10.5703/1288284317768

RECOMMENDED CITATION

Khoshnevisan, S., Norouzi, M., & Sadik, L. (2024). *Use of machine learning methods to obtain a reliable predictive model for resilient modulus of subgrade soil* (Joint Transportation Research Program Publication No. FHWA/IN/JTRP-2024/27). West Lafayette, IN: Purdue University. https://doi.org/10.5703/1288284317768

AUTHORS

Sara Khoshnevisan, PhD
Assistant Professor of Civil & Architectural Engineering and Construction Management
University of Cincinnati
(513) 556-5456
sara.khoshnevisan@uc.edu
Corresponding Author

Mehdi Norouzi, PhD
Associate Professor Educator
University of Cincinnati

Laith Sadik
Graduate Researcher
Geotechnical Engineering
University of Cincinnati

JOINT TRANSPORTATION RESEARCH PROGRAM

The Joint Transportation Research Program serves as a vehicle for INDOT collaboration with higher education institutions and industry in Indiana to facilitate innovation that results in continuous improvement in the planning, design, construction, operation, management and economic efficiency of the Indiana transportation infrastructure. https://engineering.purdue.edu/JTRP/index_html

Published reports of the Joint Transportation Research Program are available at http://docs.lib.purdue.edu/jtrp/.

NOTICE

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views and policies of the Indiana Department of Transportation or the Federal Highway Administration. The report does not constitute a standard, specification or regulation.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. FHWA/IN/JTRP-2024/27	**2. Government Accession No.**	**3. Recipient's Catalog No.**

4. Title and Subtitle

Use of Machine Learning Methods to Obtain a Reliable Predictive Model for Resilient Modulus of Subgrade Soil

5. Report Date

August 2024

6. Performing Organization Code

7. Author(s)

Sara Khoshnevisan, Mehdi Norouzi, and Laith Sadik

8. Performing Organization Report No.

FHWA/IN/JTRP-2024/27

9. Performing Organization Name and Address

Joint Transportation Research Program
Hall for Discovery and Learning Research (DLR), Suite 204
207 S. Martin Jischke Drive
West Lafayette, IN 47907

10. Work Unit No.

11. Contract or Grant No.

SPR-4714

12. Sponsoring Agency Name and Address

Indiana Department of Transportation (SPR)
State Office Building
100 North Senate Avenue
Indianapolis, IN 46204

13. Type of Report and Period Covered

Final Report

14. Sponsoring Agency Code

15. Supplementary Notes

Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract

This project explores the development and optimization of predictive models for the resilient modulus (M_R) of subgrade soil using advanced machine learning techniques. Comprehensive data from INDOT spanning several years was analyzed to enhance the accuracy of M_R predictions. The study not only refined the modeling approach through statistical methods and validation but also identified crucial soil properties that significantly impact M_R values. Recommendations for future data collection were made to further improve the models. The developed models and these recommendations will be used to guide INDOT in making informed decisions for pavement design and maintenance, which will ultimately lead to more efficient and cost-effective engineering practices.

17. Key Words

clustering, machine learning, model validation, random forest, XGBoost, regression, repeated load triaxial test, resilient modulus

18. Distribution Statement

No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161.

19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 42 including appendices	**22. Price** $63,841.30

EXECUTIVE SUMMARY

Introduction

This project aimed to develop an advanced predictive model utilizing machine learning algorithms to improve the efficiency of estimating the resilient modulus (M_R) of subgrade soils, which is a critical factor in pavement design. Currently, the Indiana Department of Transportation (INDOT) relies on the repeated load triaxial testing method prescribed by AASHTO T307, which, while effective, is resource-intensive, costly, and time-consuming.

The objective of this study was to use existing INDOT resources, including M_R test data and local soil index properties, to develop a machine learning-based model that accurately predicts the M_R of subgrade soils. The goal was to minimize the need for extensive laboratory testing, thereby conserving resources and enhancing the efficiency of pavement design processes. By leveraging advanced machine learning techniques, the project intended to create a reliable tool that efficiently and cost-effectively predicted subgrade soil behavior under various stress conditions.

The project followed a structured approach to develop predictive models for the resilient modulus M_R of subgrade soil that encompassed several key stages. It began with a comprehensive literature review that examined existing research on M_R estimation and prediction, focused on the strengths and limitations of various methods, and identified essential features for model development.

Due to the diverse and often noisy nature of geotechnical data, rigorous data cleaning was performed to ensure data quality. This process included removing outliers, correcting errors, and ensuring data consistency, which was crucial for the accuracy and reliability of subsequent analysis. Following data cleaning, exploratory data analysis was conducted to investigate the intricacies of the data from a geotechnical perspective, with a focus on geological and environmental dynamics affecting M_R. Anomaly detection was then performed to identify points that significantly deviated from the norm. These anomalies were addressed to prevent potential distortions in the predictive model performance.

The development of the models began with simple linear models and progressed to more complex ones as various machine learning algorithms were evaluated to determine the most effective model for predicting M_R based on soil properties.

In addition to ML model development, a curve-fitting method from the SciPy Python library was employed to refine and optimize the coefficients of the constitutive model employed by the repeated load triaxial (RLT) testing equipment at INDOT.

Finally, the models were rigorously tested and validated to ensure their effectiveness. This included training on a portion of the data and validating on a separate set to test the model's ability to generalize to new data. These steps provided a structured approach to handling and analyzing data, developing robust models, and ensuring that the predictions were reliable and valid for practical applications.

Findings

1. This project led to the development of multiple machine learning models that predicted the resilient modulus (M_R) of subgrade soil. These models used existing INDOT data to estimate M_R more efficiently than traditional methods. They were evaluated and validated against actual test results to ensure they provided dependable predictions.
2. The developed models estimated M_R using various approaches and facilitated the identification of similar data points within the existing dataset. This capability provided an informed estimate of the M_R value and aided in the decision making and evaluation of the M_R predictions provided by the machine learning model.
3. The study pinpointed critical soil properties that substantially impact M_R predictions. Understanding these key features can guide more meticulous approaches in laboratory testing to ensure the accurate collection of these properties for future assessments.
4. A list of suggestions was provided to highlight the points that need attention when conducting laboratory tests, ensuring that the M_R values obtained from repeated load triaxial (RLT) testing are reliable and the reports are comprehensive and contain all necessary information.
5. The project provided valuable insights into the use of machine learning in geotechnical applications, thus contributing to professional development and enhancing the skill set within INDOT.
6. Recommendations were provided for future data collections that refine the developed models to increase their prediction accuracy.
7. A fully compiled dataset of Shelby tube samples, encompassing repeated load triaxial (RLT) testing and sieve analysis results, hydrometer tests, Atterberg limits, moisture content, and dry density measurements, is available. This dataset successfully passed all required sanity checks.

Implementation

The project implementation was meticulously planned and executed through well-defined stages. The models were tuned and refined using new data provided by INDOT members during the project. Rigorous testing was then conducted with additional error analysis and geotechnical insights to ensure the model performed as expected.

For the deployment of the models, the Python code was housed in a Google Colab notebook. Google Colab provided a cloud-based environment with popular libraries and frameworks, which eliminated the need for local setup. It also offered free access to GPUs and TPUs to enhance computational power for data-intensive tasks without the need for costly hardware. The platform facilitated easy sharing and seamless collaboration. It integrated with Google Drive for managing large datasets and was compatible with Jupyter notebooks. Moreover, Colab supported interactive data visualization and was highly scalable, allowing work from any internet-connected device.

The regression models were delivered in an Excel file, providing a simpler, more accessible format for users who preferred this application for its straightforward usability and widespread adoption in various professional settings.

By the end of the implementation phase, comprehensive training sessions were conducted to review all steps of model development and to familiarize INDOT members with the developed models and the Python code, which ensured a smooth transition post-deployment.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

Acronym	Meaning
AASHTO	American Association of State Highway and Transportation Officials
ANN	Artificial Neural Networks
dd	Dry Density
DT	Decision Tree
FC	Fines Content
FWD	Falling Weight Deflectometer
INDOT	Indiana Department of Transportation
iTrees	Isolation Trees
LL	Liquid Limit
MAE	Mean Absolute Error
MEPDG	Mechanistic-Empirical Pavement Design Guide
ML	Machine Learning
M_R	Resilient Modulus
MSE	Mean Squared Error
NCHRP	National Cooperative Highway Research Program
OMC	Optimum Moisture Content
p200	Percent Passing Sieve No. 200-Fines Content
PL	Plastic Limit
R	Pearson's coefficient
R^2	R-squared
ReLU	Rectified Linear Unit
RF	Random Forest
RLT	Repeated Load Triaxial
RMSE	Root Mean Square Error
RSS	Residual Sum of Squares
USCS	Unified Soil Classification System
USDA	United States of Department of Agriculture
XGBoost	Extreme Gradient Boosting

1. RATIONALE AND NEED FOR RESEARCH

Soil resilient modulus (M_R) characterizes the stiffness of unbound materials in pavement systems. Specifically, it is defined as the ratio of cyclic stress applied to the recoverable, or elastic, strain after numerous cycles of repeated loading as shown in Equation 1.1. As such, it provides a direct measure of soil stiffness and is an important parameter for the design and analysis of pavement structures. Because of the major impact the resilient modulus has on pavement performance under repeated traffic loadings, it has been adopted in the *Mechanical Empirical Pavement Design Guide* (MEPDG) (AASHTO, 2003).

$$M_R = \frac{\sigma_d}{\varepsilon_r} \qquad \text{(Eq. 1.1)}$$

where σ_d is the deviator stress; and ε_r is the recoverable strain.

The determination of the soil resilient modulus can be achieved via laboratory or field tests. One of the most prevalent laboratory methods is the repeated load triaxial test, which follows the AASHTO T307 procedure. In this method, a cylindrical soil sample (disturbed or undisturbed), is confined using a rubber membrane and subjected to a confining pressure while loaded axially. The specimen is subjected to a predetermined number of loading cycles, and the axial strain and stress responses are measured at each cycle. Subsequently, the M_R is computed by fitting a mathematical model to the measured stress-strain responses.

Alternatively, M_R can also be directly determined in the field via the falling weight deflectometer (FWD). This non-destructive testing device operates by releasing a steel plate from a predetermined height onto the pavement surface, inducing a transient loading scenario. Sensors positioned at various points around the impact site capture the ensuing deflection reaction of the pavement surface. The collected deflection data is subsequently subjected to analysis using a back-calculation technique, facilitating the estimation of the soil's M_R.

Both the laboratory and field-testing approaches tend to be complex, costly, labor-intensive, and time-consuming. Therefore, over the past few decades, significant efforts have been dedicated to developing correlation models that predict M_R based on different properties avoiding the necessity for direct testing. Part of these endeavors has focused on developing M_R models solely reliant on soil properties (e.g., Carmichael, 1978; Hajj et al., 2018; Jackson, 2015; Rahim, 2005). However, these models have some limitations and are area- or material-specific. In another aspect of this endeavor, researchers have developed constitutive equations to predict the resilient modulus in a more comprehensive and efficient manner (ARA, 2004; Hicks & Monismith, 1971; Uzan, 1985; Witczak & Uzan, 1988). These constitutive models are stress state-dependent, and the regression coefficients/model parameters are material specific (e.g., dependent on the material). The most generalized M_R constitutive equation, as depicted in Equation 1.2, is developed by NCHRP 1-37A (ARA, 2004) and further adopted by the MEPDG (AASHTO, 2015, 2020).

$$M_R = k_1 p_a \left(\frac{\theta}{p_a}\right)^{k_2} \left(\frac{\tau_{oct}}{p_a} + 1\right)^{k_3} \qquad \text{(Eq. 1.2)}$$

where τ_{oct} is the octahedral shear stress; θ is the bulk stress; p_a is the atmospheric pressure equal to 101.4 kPa; and k_1, k_2, and k_3 are the regression coefficients.

While generalized models provide a useful starting point, the unique characteristics of local soils, environmental conditions, and regional practices necessitate the development of area-specific models. These tailored models enhance the accuracy and reliability of predictions, leading to better decision-making, optimized resource use, and improved compliance with local standards.

To ensure the selection of the most suitable approach and methodology for the project, an extensive literature review was meticulously conducted. The results of this literature review are presented in the following section.

2. LITERATURE REVIEW

This literature review involved an extensive examination of a wide range of academic papers, research articles, relevant reports, and scholarly works related to the project's subject matter. The objective was to achieve a comprehensive understanding of the existing body of knowledge and methodologies within the field. The results of this review serve as a critical foundation for the project's approach and methodology. The insights gained equipped the research team to make informed decisions about which methodologies are best suited to address the project's unique challenges and contribute meaningfully to the field of study. This thorough review not only enhanced the quality and validity of the research but also positioned the project within the broader context of existing knowledge and advancements in the chosen field.

2.1 Review of Factors Affecting M_R

Initially, a comprehensive literature review was undertaken to explore the factors (parameters) affecting the soil's M_R. It should be highlighted that the focus was only on fine-grained soils, which held the dominant presence in the state of Indiana. This study aimed to enhance understanding of the resilient characteristics of the soil of interest, while also providing valuable support for the "feature selection" process- an essential phase in machine learning approach.

2.2 Stress State

The soil's M_R is a stress-dependent parameter exhibiting sensitivity to both confining and deviator stress (Han & Vanapalli, 2015; Khasawneh, 2019;

Sweere, 1990; Uzan, 1985). Regarding the confining stress, it is shown that there is a slight tendency for the M_R to increase with an increase in the confining stress (Houston et al., 1993; Mohammad et al., 1995). However, according to some studies, this increase is marginal (Pezo & Hudson, 1994; Nguyen & Mohajerani, 2015).

Regarding deviator stress, it has been observed that the M_R is significantly affected by it. As the deviator stress increases, the M_R decreases due to the softening effect (Houston et al., 1993; Khasawneh, 2019; Kim & Siddiki, 2005; Nazzal & Mohammad, 2010).

2.2.1 Soil Density

Many studies have aimed to understand the relationship between M_R and density (e.g., Andrei et al., 2009; Seed et al., 1962; Thom, 1988). However, a complete understanding of this relationship remains elusive (Alnedawi et al., 2022). This lack of clarity could arise from the interplay between density and other factors such as compaction level, sample preparation methodology, fines content, moisture content, relative compaction, and applied stress (Zvonarić et al., 2021). Conversely, Brown and Selig (1994) have stated that the impact of density on M_R is relatively insignificant.

2.2.1.1 Moisture content. The degree of saturation significantly affects the M_R, with M_R decreasing as soil saturation increases (Mitry, 1965; Seed et al., 1962) indicating that M_R decreases when the soil is saturated. Additionally, Cary and Zapata (2011) noted that M_R might increase as moisture content decreases. For moisture contents exceeding the optimum moisture content (OMC), a decrease in M_R value is expected (e.g., Dawson et al., 1996; Ekblad, 2008). Duong et al. (2015) demonstrated that the effect of moisture content on M_R increases with higher fines content in the soil.

2.2.1.2 Matric suction. Studies conducted by Yang et al. (2005), Liang et al. (2008), and Ba et al. (2013) have shown that the use of matric suction in the prediction of M_R of subgrade soils and unbound pavement provides better accuracy than using the moisture content.

2.2.1.3 Freeze and thaw. The M_R is notably influenced by the process of freezing and thawing (Domitrović et al., 2019). Frozen fine-grained soils exhibit an elevated M_R when compared to their unfrozen state. Conversely, the soil that has recently undergone thawing exhibits a significant decrease in M_R compared to the frozen and unfrozen conditions (Hardcastle, 1992).

The study by Qu et al. (2019) showed a notable decrease in the M_R as the number of freeze-thaw cycles increase. According to Çoleri (2007), the most crucial phase for M_R occurs at the end of thawing, at which the M_R is anticipated to reach its minimum level. This critical M_R value should be considered during the pavement design stage to mitigate the risk of failure.

2.2.1.4 Fines content. Several studies have shown that the M_R generally decreases as the fine's contents increase (Qu et al., 2019; Saberian & Li, 2021).

2.2.1.5 Stress history. The stress history is expected to impact M_R due to progressive densification and particle rearrangement under repeated loading conditions (Dehlen & Monismith, 1970). Boyce (1976) has proposed that the impact of stress history can be lessened by preloading, involving a few cycles of the current loading as a conditioning stage. Conversely, Hicks and Monismith (1971) have stated that stress history holds minimum effect and can be nearly eliminated; and that a stable resilient response can be obtained after applying approximately 100 cycles of the same stress amplitude and will stay the same after 25,000 repetitions. However, Allen (1973) recommended subjecting the specimen to approximately 1,000 load repetitions before conducting the repeated load resilient tests.

2.3 Review of Adopted Approaches

2.3.1 Calibration of the Regression Coefficients of the Generalized M_R Constitutive Model

Given the extensive practical application of Equation 1.2, a lot of effort has been directed towards refining its predictive capabilities. In this context, considerable work has been invested in harnessing soil properties to develop models that yield more representative regression coefficients (namely, k_1, k_2, and k_3), thereby subsequently improving the prediction capability of Equation 1.2. The reported models for fine-grained soils are summarized in Table 2.1.

However, since the models shown in Table 2.1 are data-driven, they have generalizability limitations because of three reasons: (1) spatial variability of the soil; (2) sample size of the data used for model; and (3) the models may rely on features that might not be available in new datasets.

2.3.2 Predictive M_R Models Using Machine Learning Algorithms

As a result of the progress made in Machine Learning (ML) methods over the past decades, numerous research endeavors have concentrated on creating ML-driven models for predicting M_R based on fundamental soil characteristics (Azam et al., 2022; Hanittinan, 2007; Kardani et al., 2022; Khasawneh & Al-jamal, 2019; Pal & Deswal, 2014; Sadrossadat, Heidaripanah, & Ghorbani, 2016, Sadrossadat, Heidaripanah, & Osoul, 2016; Hoang & Nguyen, 2021). The frequency of use of different features for

TABLE 2.1
Summary of MEPDG regression coefficients prediction equations

Model	Soil Type	Reference
$k_1 = 404.166 + 42.933\ PI + 52.26\gamma_d - 987.353\left(\frac{w}{w_{opt}}\right)$ $k_2 = 0.25113 - 0.0292\ PI + 0.5573\left(\frac{w}{w_{opt}}\right) \times \left(\frac{\gamma_d}{\gamma_{dmax}}\right)$ $k_3 = -0.20772 + 0.23088\ PI + 0.00367\gamma_d - 5.4238\left(\frac{w}{w_{opt}}\right)$	Fine-grained	Elias and Titi (2006)
$Log\ k_1 = 6.99969 - 0.11144\ OMC - 1.1532\ MCR - 0.00154\ \gamma_{dmax} + 0.01875\ PI - 0.02339\ S1 + 0.00445\ p200$ $k_2 = 0.55494 + 0.25904\ MCR - 0.00651\ PI - 0.00785\ p4 + 0.00712\ p40 - 0.00266\ p200 - 0.00318\ CLAY$ $k_3 = 2.08483 - 0.03626\ w - 0.00044337\ \gamma_{dmax} + 0.01104\ LL - 0.02024\ S1 + 0.00494\ SN80 + 0.01012\ CSAND + 0.00392\ FSAND + 0.00287\ SILT$	Fine-grained	Malla and Joshi (2008)
$Ln\ k_1 = 1.334 + 0.0127\ P200 + 0.016\ LL - 0.036\ \gamma_{dmax} - 0.011\ MCCL + 0.001\ MCD_{DmaxP}$ $k_2 = 0.722 + 0.0057\ LL - 0.00454(MCDDmaxPI)^{0.641} + 0.00324\ MCDDP^{1.28} - 0.875(P200)$ $k_3 = -7.48 + 0.235\left(\frac{\gamma_d}{w}\right) + 0.038LL - 0.0008MCPI + 0.033\gamma_{dmax} - 0.016MCDDP$	A-4 A-6 A-7-5 A-7-6	Nazzal and Mohammad (2010)
$k_1 = 15.2755\ ADDO + 102.276\ AMCO + 15.5439\ Clay + 11.5423\ Gravel - 2.68004\ LL - 30.39\ \gamma_{dmax} - 127.734\ OMC - 14.5558\ PI + 12.5928\ Sand + 10.2289\ Silt + 562.592\ SG$ $k_2 = -0.0537\ ADDO + 0.38034\ AMCO - 0.00158\ Clay - 0.00577\ Gravel + 0.05675\ LL + 0.07329\ \gamma_{dmax} - 0.41051\ OMC + 0.03855\ PI + 0.01153\ Sand + 0.00322\ Silt - 1.38553\ SG$ $k_3 = 0.05003\ ADDO + 0.53396\ AMCO - 0.00486\ Clay - 0.01169\ Gravel + 0.04337\ LL - 0.11358\ \gamma_{dmax} - 0.68651\ OMC + 0.025\ PI - 0.00194\ Sand - 0.00836\ Silt + 3.25546\ SG$	A-4	Ji et al. (2014)
$k_1 = 10.373\ ADDO - 160.526\ AMCO - 1.23003\ Clay - 1.41065\ Gravel + 14.3789\ LL + 0.66964\ \gamma_{dmax} + 123.671\ OMC + 6.63765\ PI + 4.59136\ Sand + 2.00541\ Silt + 469.806\ SG$ $k_2 = -0.01986\ ADDO + 0.01833\ AMCO - 0.00484\ Clay - 0.01011\ Gravel - 0.00358\ LL + 0.02925\ \gamma_{dmax} - 0.01324\ OMC - 0.00985\ PI - 0.00646\ Sand - 0.00263\ Silt - 0.06798\ SG$ $k_3 = -0.01523\ ADDO - 0.00001\ AMCO + 0.00908\ Clay + 0.01576\ Gravel - 0.00555\ LL + 0.00598\ \gamma_{dmax} - 0.02212\ OMC + 0.01139\ PI + 0.0116\ Sand + 0.00892\ Silt + 0.06054\ SG$	A-6	Ji et al. (2014)
$k_1 = 40.4804\ ADDO - 71.9733\ AMCO + 13.5356\ Clay + 14.7768\ Gravel + 3.28479\ LL - 46.8996\ \gamma_{dmax} + 43.007\ OMC + 5.8166\ PI + 15.0582\ Sand + 19.0182\ Silt - 115.307\ SG$ $k_2 = -0.09401\ ADDO + 0.41507\ AMCO + 0.06418\ Clay + 0.0352\ Gravel - 0.02432\ LL + 0.0268\ \gamma_{dmax} - 0.50741\ OMC + 0.02229\ PI + 0.09066\ Sand + 0.07255 + 0.81695\ SG$ $k_3 = 0.13677\ ADDO - 0.16641\ AMCO - 0.03009\ Clay - 0.02735\ Gravel + 0.00768\ LL - 0.11805\ \gamma_{dmax} + 0.18766\ OMC - 0.00922\ PI - 0.03127\ Sand - 0.02671\ Silt + 0.15138\ SG$	A-7-6	Ji et al. (2014)

Note: *MCR* = moisture content ratio = moisture content/optimum moisture content; *DDR* = dry density ratio = γ_d / γ_{dmax}; *SG* = specific gravity; *ADDO* = actual dry density at optimum moisture content.

γ_d = dry density; *w* = moisture content; *CSAND* = percent coarse sand (particles of size 2–0.42 mm); *FSAND* = percent fine sand (particles of size 0.42–0.074 mm); *MCCL* = (moisture content-optimum moisture content)/Clay; *AMCO* = actual moisture content at optimum moisture content; *SN80* = percent passing of sieve No. 80; *S1* = percent passing 1-in. sieve; *p4* = percent passing sieve No. 4, *p40* = percent passing sieve No. 40.

$$MCDD\ max\ P = p200 \times \frac{moisture\ content - optimum\ moisture\ content}{optimum\ moisture\ content} \times \frac{\gamma_d}{\gamma_{d,\ max}}$$

$$MCDD\ max\ PI = plasticity\ index \times \frac{moisture\ content - optimum\ moisture\ content}{optimum\ moisture\ content} \times \frac{\gamma_d}{\gamma_{d,\ max}}$$

$$MCPI = plasticity\ index \times \frac{moisture\ content - optimum\ moisture\ content}{optimum\ moisture\ content}$$

$$MCDDP = p200 \times \frac{\gamma_d}{moisture\ content}$$

M_R model development in the various studies reviewed in the literature is listed in Figure 2.1.

Most of the studies referenced conducted sensitivity analyses to rank the importance of features in their model development process. Since the range of sensitivity analysis results varies for each study, the values are normalized for each column by subtracting the minimum value and dividing by the range of that column, transforming the values to range from 0 to 1 (Table 2.2). The summation of the normalized sensitivity analysis results for each parameter from different studies is then used as a measure of parameter importance in M_R modeling.

Figure 2.2 shows the normalized importance ranking of each feature in M_R prediction based on previous research.

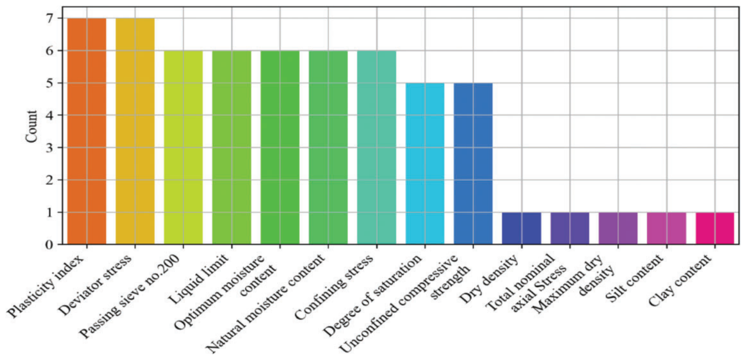


Figure 2.1 Adoption frequency of each feature for M_R model development in other studies.

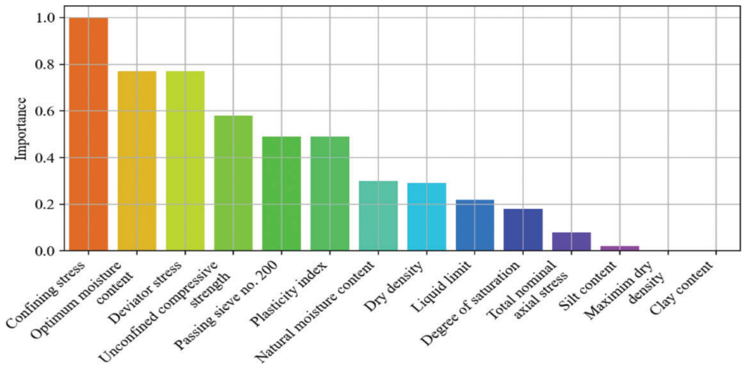


Figure 2.2 Average feature importance in M_R modeling based on literature review.

TABLE 2.2
Normalized sensitivity analysis results from previous research

	References						Total
Feature	**Pal & Deswal (2014)**	**Sadrossadat, Heidaripanah, & Ghorbani (2016)**	**Sadrossadat, Heidaripanah, & Osoul (2016)**	**Khasawneh & Al-jamal (2019)**	**Azam et al. (2022)**	**Kardani et al. (2022)**	**Sensitivity**
Confining Stress	1.00	0.11	0.86	–	0.54	0.95	3.46
Optimum Moisture Content	0.00	1.00	1.00	0.00	0.67	–	2.67
Deviator Stress	0.39	0.03	0.57	1.00	0.00	0.67	2.66
Unconfined Compression Strength	0.14	0.25	0.71	–	0.89	–	1.99
Passing Sieve200	0.032	0.11	0.57	0.00	1.00	–	1.71
Plasticity Index	0.05	0.83	0.00	0.00	0.44	0.38	1.70
Natural Moisture Content	0.09	0.23	0.43	–	0.30	0.00	1.05
Dry Density	–	–	–	–	–	1.00	1.00
Liquid Limit	0.016	0.06	0.43	0.00	0.27	–	0.77
Degree of Saturation	0.12	0.00	0.29		0.20	–	0.61
Total Nominal Axial Stress	–	–	–	0.26	–	–	0.26
Maximum Dry Density	–	–	–	0.00	–	–	0.00
Clay Content	–	–	–	0.00	–	–	0.00
Silt Content	–	–	–	0.08	–	–	0.08

In summary, the conducted literature review provided valuable insights into current practices and adopted ML approaches, as well as a clear understanding of the features affecting the M_R of soil. However, due to spatial variability in soil, the models and features identified cannot be directly applied for M_R model development in other locations without caution. Thus, a new model needed to be developed specifically for Indiana. Also, it should be noted that while the feature importance analysis aided in the decision-making for feature selection in model development, not all those features are available within the INDOT's dataset.

3. DATA COMPILATION AND DATA PREPROCESSING

An Excel file containing a dataset of 2,008 M_R records, each from a separate repeated load triaxial test conducted by INDOT over seven years from 2016 to 2022, was provided to the research team. This dataset contained information about the test date, the soil sample location (county and road), boring number, sample type (Shelby tube or bagged sample), AASHTO classification, textural classification, Atterberg limits (liquid limit and plastic limit), natural moisture content, optimum moisture content (only available for bagged samples)), maximum dry density (only available for bagged samples), natural wet density, dry density, and the measured M_R corresponding to the confining stress of 2 psi and the deviator stress of 6 psi. However, this Excel file includes several attributes with missing values. Figure 3.1 provides a detailed breakdown of each feature with missing values, highlighting the percentage of missing data for each attribute.

To verify the accuracy of the Excel file dataset and identify missing values, the original raw data —individual triaxial test reports for each recording provided by INDOT—was needed. Given the large number of files and the impracticality of manual review, PDF parsing was employed to automatically extract information from the reports. This enabled the identification of missing data, ensured the integrity of the entire dataset, and augmented the data with additional features such as stress levels, as discussed in the following sections. For example, moisture content values were erroneously entered in the optimum moisture content column and have been relocated to the correct moisture content column.

Moreover, the sieve analysis results of each recording were added as new feature to the database, and thus, the AASHTO classification was checked for correctness, and the miss-classified data points were corrected. Figure 3.2 shows the AASHTO classification zones, and the miss-classified points.

Similarly, the textural classification for each data point was re-evaluated according to the United States Department of Agriculture (USDA) textural classification chart and corrected where necessary.

Most research conducted on developing M_R prediction models has incorporated sieve analysis results as part of the predictor features (e.g., Hanittinan, 2007; Pal & Deswal, 2014; Sadrossadat, Heidaripanah, & Ghorbani, 2016). Therefore, the sieve analysis and hydrometer testing reports for each record were reviewed to determine the percentages of gravel, sand, silt, clay, and the material passing sieve No. 200 (p200). This information was then added to the database for records where it was available.

Also, the M_R values were augmented with corresponding confining and deviator stress combinations for each data point, as M_R is highly dependent on soil stress state (Kim & Siddiki, 2005). According to AASHTO T-307 specifications (AASHTO, 2003), each soil sample has 15 M_R values, corresponding to 15 unique combinations of confining and deviator stress. By incorporating these stress combinations, a more comprehensive understanding of M_R values can be achieved.

As noted earlier, the database comprises two types of soil samples: undisturbed samples collected from

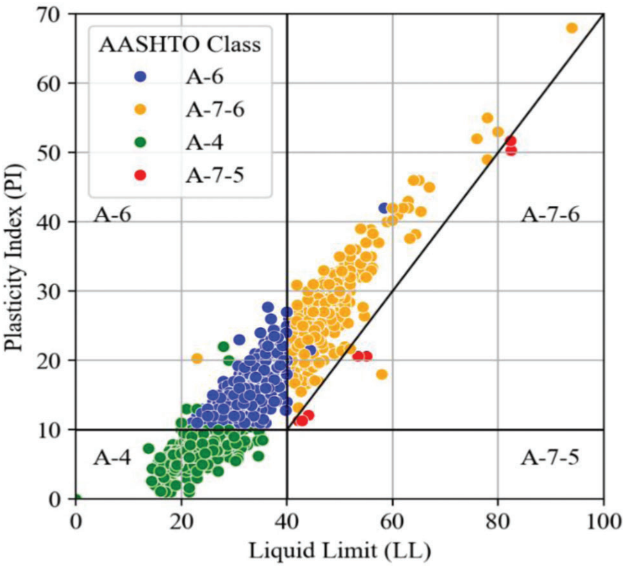


Figure 3.2 Data points' clusters within AASHTO classification zones (before correction).

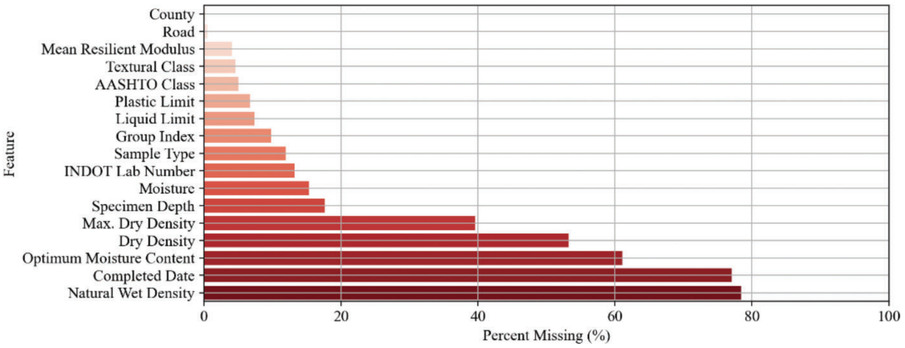


Figure 3.1 Incomplete data entries in the Excel dataset supplied by INDOT.

Shelby tubes and remolded samples gathered from sites in bags, referred to as "bag samples." This distinction is crucial, as the sampling methods yield different M_R values. Also, from the entire dataset, 77% of the recordings are for Shelby tube samples, and the remaining are for bag samples. Thus, the focus of the project is on the Shelby tube samples.

This study prioritizes the dominant soil types in Indiana, which are predominantly fine-grained (A-4, A-6, and A-7-6 soils). Due to their scarcity and distinct properties, the small number of coarse-grained soils (A-1, A-2, and A-5 soils) were excluded from the analysis to maintain a focused and representative dataset.

4. DATA CLEANING

To guarantee the accuracy and consistency of the compiled dataset, a thorough data cleaning process was performed to verify that each feature, as well as their combinations, fall within logical ranges and exhibit reasonable relationships.

4.1 Exclusion of Incomplete Test Reports

Data points featuring an individual test comprising less than 15 stress levels were removed from the dataset. It was observed during the data compilation phase that certain tests contained less than the required 15 stress levels, which does not adhere to the standard guidelines set forth by AASHTO T307. There were 1,294 records that had all 15 M_R stress levels bringing the total data count to 19,410.

Additionally, the tests that exhibit missing values for important features such as plastic limit, liquid limit, moisture content, and dry density, were eliminated from the dataset, bringing the total data count to 19,365.

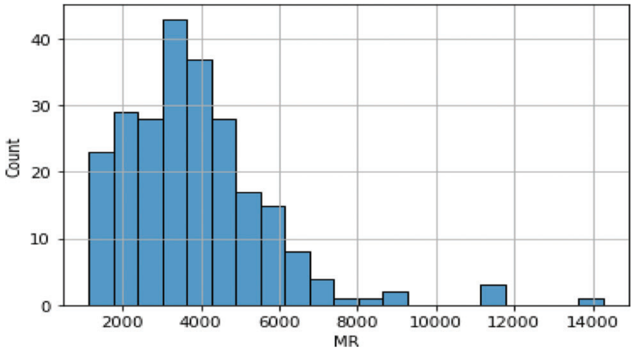


Figure 4.1 M_R distribution of samples with moisture content exceeding their liquid limit.

Furthermore, as sieve analysis results are established as reliable predictors of M_R in the existing literature, the analysis concentrated on the subset of the dataset where complete sieve analysis data is available, bringing the data count to 15,556.

4.2 Selection of Soil Samples with Stress-Softening Behavior

Since the soil classification under investigation includes A-4, A-6, and A-7 soils, it is anticipated that M_R will either decrease or remain constant as the deviator stress increases, based on previous research (Ji et al., 2014; Kim & Kim, 2007; Kim & Siddiki, 2005). To focus on the relevant data, a specialized algorithm is developed to identify soil samples that exhibit a decline in M_R with increasing deviator stress. This filter brought the dataset to a total of 11,851 datapoints.

4.3 Removing Inconsistent Liquid Limit and Moisture Content Data

Upon examining the compiled dataset, it was observed that certain data points exhibited an illogical relationship between natural moisture content and liquid limit, as depicted in Figure 4.1. Specifically, these points showed moisture content exceeding the liquid limit, yet still displayed M_R values like other data points. Since soils in a liquid state cannot be molded or withstand repeated load triaxial testing, these data points were deemed erroneous and removed from the dataset. After this correction, the dataset consisted of 11,311 records.

4.4 Removing Unreasonable Extreme M_R Values

A significant number of data points displayed out-of-range M_R values. To rectify this, M_R values are capped at 10,800 psi based on expertise of INDOT Study Advisory Committee and the box plot visualization (Figure 4.2), as any value exceeding this threshold is deemed an outlier. This cap is endorsed by INDOT, as values beyond 10,800 psi are not typical for the stress level of interest. The dataset now contains 10,877 records.

Figure 4.3 illustrates the data cleaning and preprocessing workflow employed for this dataset, showcasing

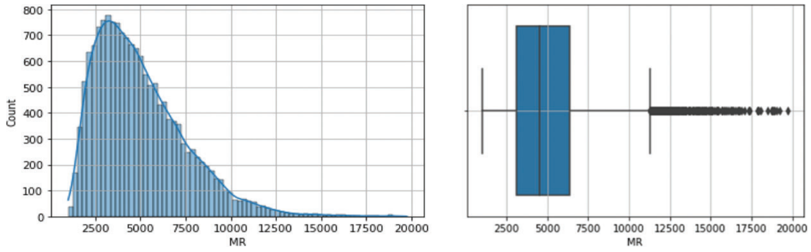


Figure 4.2 M_R distribution.

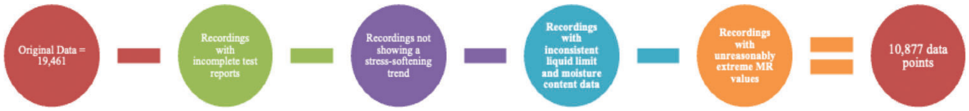

Original Data = 19,461
Recordings with incomplete test reports
Recordings not showing a stress-softening trend
Recordings with inconsistent liquid limit and moisture content data
Recordings with unreasonably extreme MR values
10,877 data points

Figure 4.3 Sanity check summary.

the steps taken to refine and prepare the data for analysis.

After compiling the dataset, filling missing values, and performing data cleaning, a database was formed using the Pandas Library in the Python programing language using a cloud service to store the data.

5. ANOMALY DETECTION

Anomaly detection is a crucial aspect of data analysis, aimed at identifying data points that deviate significantly from the norm within a dataset. These anomalies, often indicative of errors or unexpected patterns, can distort the performance of machine learning models if left unaddressed. In this study, two approaches were implemented to identify possible anomalies: (1) The similarity algorithm; and (2) the unsupervised learning approach, which will be discussed in the following subsections.

5.1 Similarity Algorithm

A specialized algorithm was developed utilizing the Python programming language to address anomalies within the dataset. This algorithm focuses on grouping data points with similar features but drastically different M_R values, which may confuse machine learning models. Operating on the principles of measuring Euclidean distance and cosine similarity concurrently, the algorithm effectively groups data exhibiting such similarities. Table 5.1 provides an illustrative example of two groups derived through this approach.

While the similarity algorithm proves effective in grouping data points that share similar soil properties but exhibit varying M_R values, the challenge arises in determining if a point within a group is an anomaly, due to the inherent variability of soil. Consequently, no points were excluded based solely on the algorithm's groupings. Nonetheless, this algorithm may prove useful in other datasets where anomalies are more distinct. Additionally, this method can help identify similar points within a dataset when applying developed models to new datasets, thereby enhancing decision-making, and validating the model's predictions.

5.2 Unsupervised Learning Algorithms

Unsupervised learning is a branch of machine learning where algorithms are tasked with extracting patterns and insights from unlabeled data without explicit guidance or supervision. In the context of anomaly detection, unsupervised learning algorithms excel in identifying anomalies by learning the inherent structure of the data and flagging instances that deviate significantly from the norm.

Isolation Forest is an unsupervised anomaly detection algorithm that isolates anomalies from normal data points by randomly partitioning the data. The core concept is that anomalies are few and different, making them more susceptible to isolation than normal instances.

The algorithm constructs an ensemble of isolation trees, which are binary trees that recursively partition the data by randomly selecting a feature and a random split value within the feature's range. Anomalies are isolated closer to the root of the tree, requiring fewer partitions, while normal instances require more partitions to be isolated.

The key mathematical idea is the path length required to isolate a data point in an isolation tree. Anomalies tend to have shorter path lengths since they are easier to isolate due to their distinct nature.

1. *Random Partitioning:* The algorithm begins by randomly selecting a feature and then randomly selecting a split value within that feature's range. This process partitions the data recursively, creating a tree structure. Each node in the tree represents a split point, and this process continues until the data points are isolated or until a predefined tree depth is reached.
2. *Isolation Trees (iTrees):* Multiple iTrees are constructed from random subsets of the data. The structure of these trees is crucial because anomalies are isolated closer to the tree's root, having shorter path lengths due to their rarity and distinctness. Figure 5.1 illustrates the Isolation Trees framework.
3. *Path Lengths and Anomaly Score:* The path length is the number of edges traversed from the root node to the terminal node for each data point. The path length reflects how quickly a data point can be isolated. Anomalies, being different and fewer, tend to have shorter path lengths. The anomaly score is then calculated based on the average path length across all iTrees. Mathematically, the anomaly score $s(x,n)$ $s(x,n)$ for a data point x in a forest of n trees is defined as:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}} \quad \text{(Eq. 5.1)}$$

where, $E(h(x))$ is the average path length of xx over all trees, and $c(n)$ is the average path length in an iTrees forest with nn external nodes (normalizing factor).
4. *Decision Function:* The anomaly scores are used to classify data points. The data point is likely an anomaly if the score is close to 1. If it is much smaller than 0.5, it is considered normal. The threshold can be adjusted based on the contamination parameter, which is the expected proportion of outliers in the dataset.

TABLE 5.1
Results of similarity algorithm

Liquid Limit	Plastic Limit	Moisture Content	Dry Density	Gravel Content	Sand Content	Silt Content	Clay Content	Passing Sieve 200	AASHTO Class	M_R
Group 1										
33.5	17.5	17.5	114.3	4.1	17.5	53.8	24.6	78.5	A-6	7,670
35.8	18.7	17.9	113.9	3.9	18	55.3	22.8	78.1	A-6	4,679
Group 2										
Liquid Limit	**Plastic Limit**	**Moisture Content**	**Dry Density**	**Gravel Content**	**Sand Content**	**Silt Content**	**Clay Content**	**Passing Sieve 200**	**AASHTO Class**	M_R
37	18	20.8	106.3	1	16.8	58.3	23.9	82.2	A-6	4,161
37	16	21	106	1.5	17.1	58.2	23.3	81.5	A-6	2,567

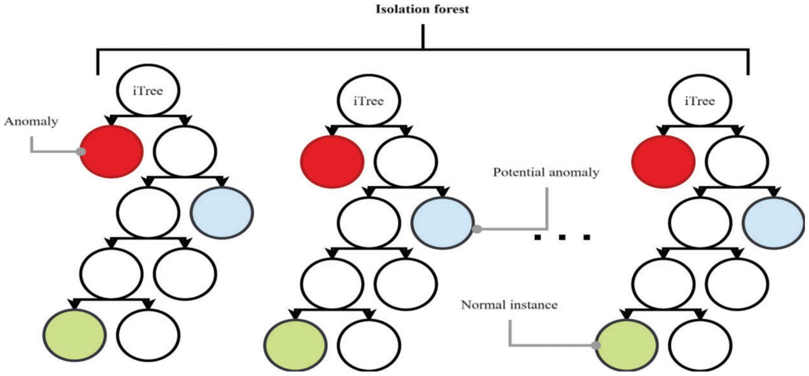


Figure 5.1 Isolation Trees framework (adopted from Regaya et al., 2021).

Although the Isolation Forest algorithm identified a subset of the data as potential anomalies, the performance of machine learning models did not improve when using the data excluding this subset.

In conclusion to the anomaly detection section, while various algorithms may detect different anomalies based on their respective frameworks, it is ultimately the geotechnical professional's expertise that should determine whether these points are anomalies. Exclusion from analysis is better based on geotechnical reasoning rather than solely relying on algorithmic-based detection methods.

6. EXPLORATORY DATA ANALYSIS

Given the inherent uncertainties often associated with geotechnical data, merely implementing data-cleaning procedures and training machine learning (ML) models may not be sufficient. ML models operate on the premise of being data-driven, yet specific data points may be considered unreasonable from a geotechnical standpoint. Thus, a more comprehensive approach was undertaken in this study, which delved into the intricacies of the data from a geotechnical perspective. In this regard, our study undertook a rigorous exploration of the data, focusing on several key aspects crucial to understanding the underlying geological and environmental dynamics.

Firstly, the potential geological grouping or zoning within the dataset was investigated. By discerning spatial patterns and identifying geological clusters, this approach aimed to uncover underlying trends and correlations that could inform model development and improve predictive accuracy.

Additionally, the textural classification differences for soil samples were examined. Understanding soil texture and composition variations is paramount in geotechnical analysis, as they can profoundly influence material properties and behavior under cyclic loading conditions.

Furthermore, the effects of seasonal changes on M_R variation were explored. Seasonal variations in environmental conditions can significantly impact soil behavior and engineering properties.

6.1 Seasonal Variation

To explore potential seasonal fluctuations in M_R, a line plot analysis was conducted to compare the mean M_R values across different seasons. However, merely examining mean values might not capture the complete picture of M_R variation. Thus, it was imperative to also assess the dispersion of M_R values for a more comprehensive representation. Consequently, the line plots were augmented to include the range of ± 2 standard deviations of M_R for each season, as depicted in Figure 6.1. Upon thorough investigation, it was

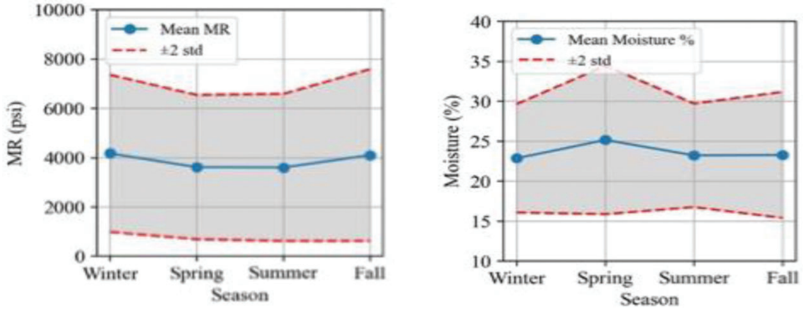


Figure 6.1 Seasonal variation of M_R and moisture content.

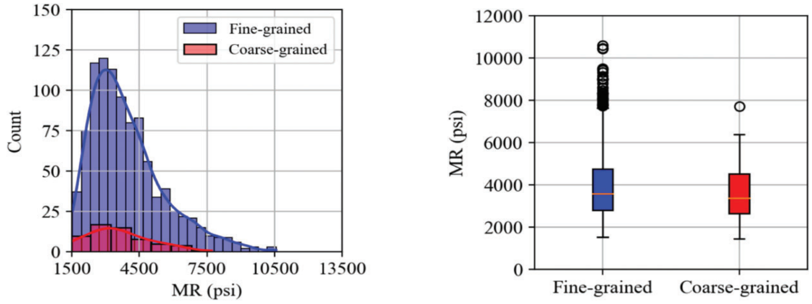


Figure 6.2 Effect of fines content on $M_{R.}$

deduced that the dataset exhibited considerable variability in M_R values across seasons, with no discernible pattern indicative of soil behavior. This underscores the complex and dynamic nature of soil properties, challenging any straightforward interpretation based solely on seasonal trends.

6.2 Fines Content

A further investigation delved into the influence of fines content on M_R values. This inquiry entailed a comparative analysis between two subsets of the dataset: one comprising soil samples with a percentage of passing sieve #200 greater than 50%, indicative of fine soil according to the Unified Soil Classification System (USCS), and another consisting of coarse-grained soils with passing #200 percentages below 50%. The distribution and boxplot of M_R values were plotted for each subset (Figure 6.2). Upon analysis, it was determined that there is no noticeable difference in the distribution or the boxplot representation between the two subsets. This observation implies that, within the scope of this dataset, fines content does not directly affect the M_R values.

6.3 Geographical Grouping

To explore potential geographical variations within the dataset, a thorough investigation was undertaken employing two distinct methodologies: (1) grouping counties by soil texture classification; and (2) grouping counties by surficial geology map. These two approaches are elaborated upon in detail in the subsequent subsection.

6.3.1 Grouping Counties by Soil Textural Classification

Leveraging the USDA soil textural classification chart illustrated in Figure 6.3, each data point was categorized into one of the four primary sections of the classification triangle: sandy, loamy, clayey, and silty. By assessing the predominant soil texture within each county, geographical groupings were established accordingly, as depicted in Figure 6.4.

Subsequently, to scrutinize potential differences between these groupings, box plots were generated to compare M_R and other relevant soil properties across the designated groups, as demonstrated in Figure 6.5. However, upon thorough analysis, no discernible differences based on this geographical grouping were observed. This finding suggests that, within the scope of this investigation, soil texture classification alone may not significantly influence the variability of M_R and associated soil properties across geographical regions.

6.3.2 Grouping Counties by Surficial Geology Map

Another investigation was conducted utilizing the surficial geology map of Indiana State, sourced from Indiana University. This map delineates the state into four distinct geological regions. Subsequently, counties within the dataset were grouped according to these geological formations, as illustrated in Figure 6.6.

By employing boxplots for M_R and other relevant soil properties, the values of these properties were juxtaposed across the designated geological regions, as depicted in Figure 6.7.

However, like previous analyses, no discernable differences were evident between the soil parameters

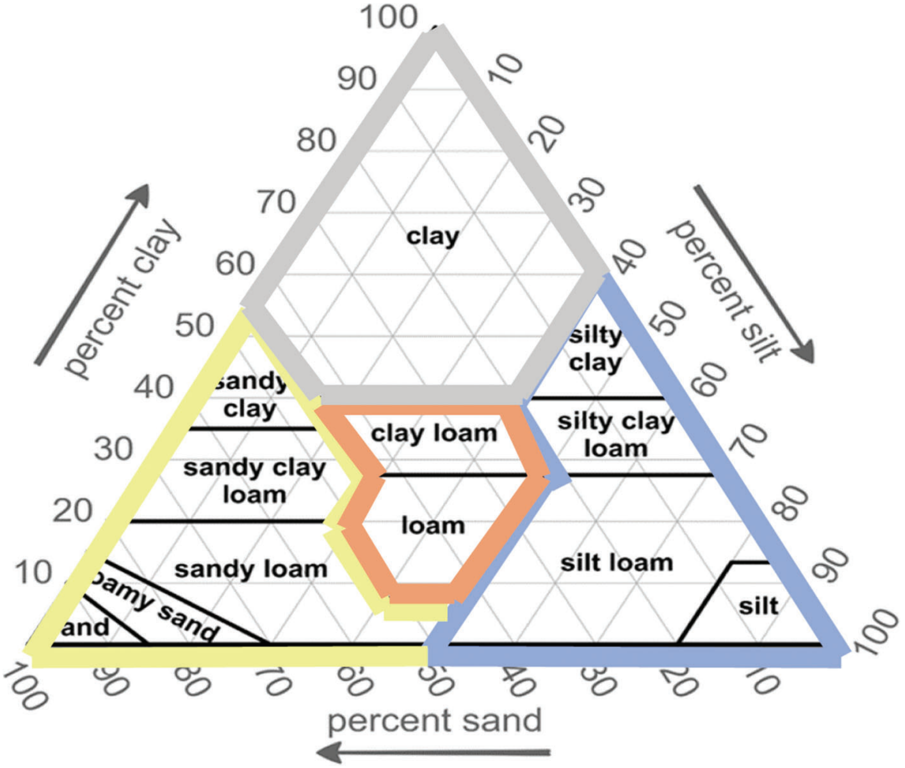


Figure 6.3 Data grouping based on USDA textural classification.

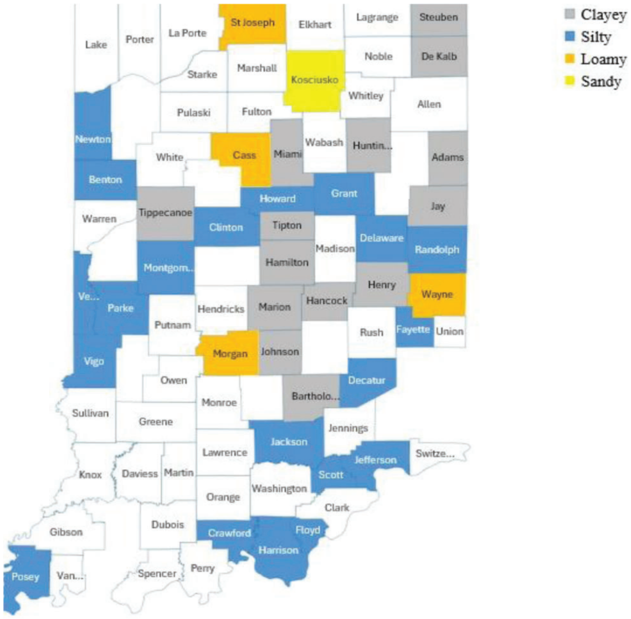


Figure 6.4 Visual presentation of county grouping based on USDA textural classification.

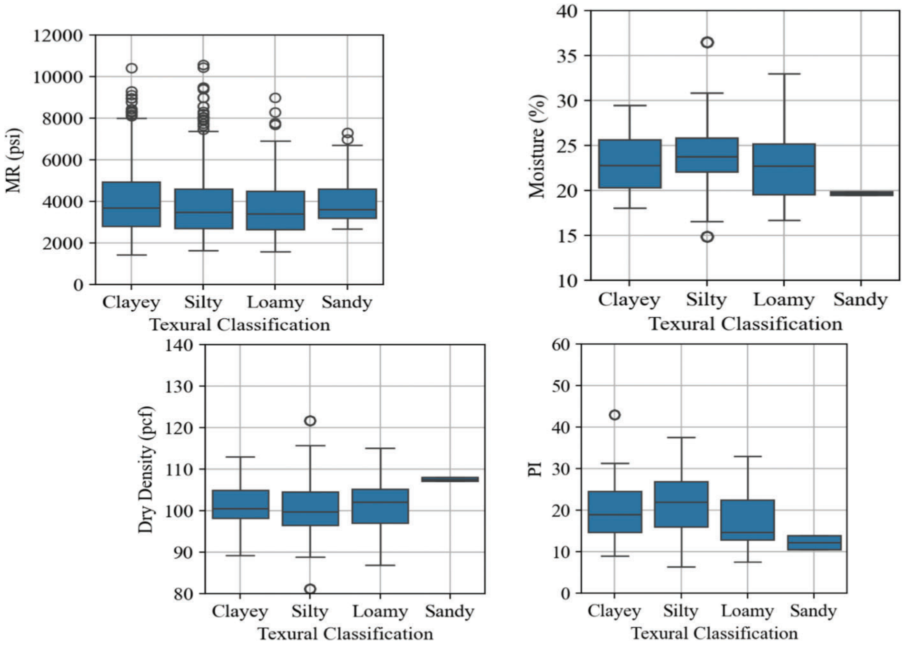


Figure 6.5 Regional variation of key soil properties across different USDA textural classifications.

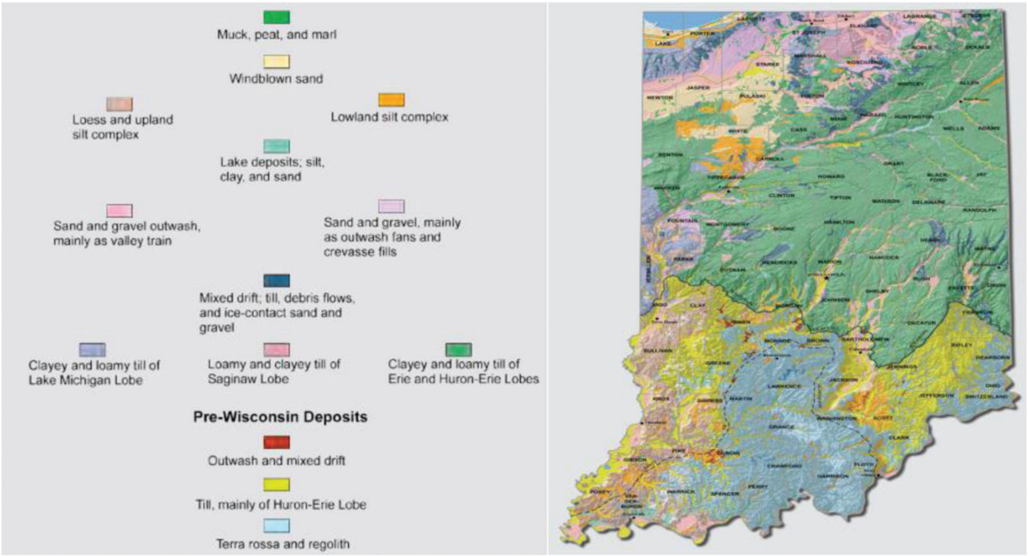

Muck, peat, and marl
Windblown sand
Loess and upland silt complex
Lowland silt complex
Lake deposits; silt, clay, and sand
Sand and gravel outwash, mainly as valley train
Sand and gravel, mainly as outwash fans and crevasse fills
Mixed drift; till, debris flows, and ice-contact sand and gravel
Clayey and loamy till of Lake Michigan Lobe
Loamy and clayey till of Saginaw Lobe
Clayey and loamy till of Erie and Huron-Erie Lobes
Pre-Wisconsin Deposits
Outwash and mixed drift
Till, mainly of Huron-Erie Lobe
Terra rossa and regolith

Figure 6.6 Geographical zoning of Indiana adopted for clustering of counties based on surficial geology.

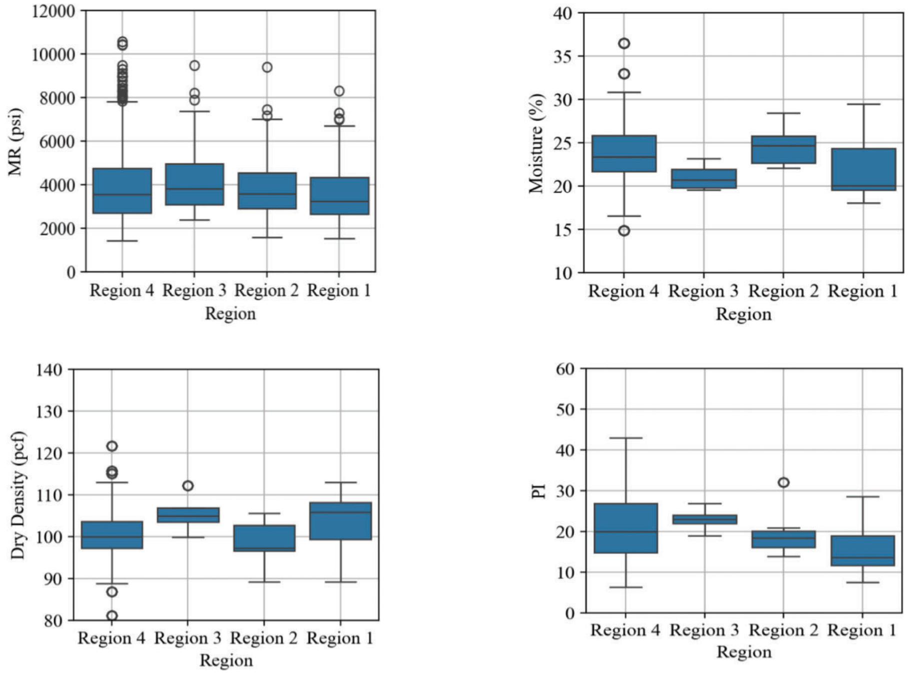


Figure 6.7 Soil variation across geological regions.

within these regions. This finding suggests that, despite geological distinctions, the variability of M_R and associated soil properties remains consistent across the surveyed areas within Indiana State.

7. MODEL DEVELOPMENT

7.1 Feature Selection

The dataset provided by INDOT consisted of 11 numerical features: liquid limit, plastic limit (PL), moisture content, dry density, gravel content, sand content, silt content, clay content, fines content (p200), confining stress, and deviator stress. Pearson's correlation coefficient (R) was calculated among features to avoid having redundant features. These relationships were visualized in the heatmap shown in Figure 7.1.

As expected, a strong linear correlation was found between the dry density and moisture content, and between the fines content and sand percentage (see Figure 7.1). With the intention of upholding the effectiveness of the baseline models, the features with high linear correlations were excluded. Excluding collinear features was shown to improve model stability, generalization, interpretability, and performance as it allowed the model to focus on capturing the most meaningful relationships within the data while avoiding the pitfalls associated with multicollinearity and overfitting.

7.2 Model Creation

The main ML models developed in this project were linear regression, polynomial regression, random forest (RF), decision tree (DT), artificial neural networks (ANNs), and extreme gradient boosting (XGBoost). One reason for selecting the XGBoost model is its strong performance when handling tabular data, as observed in comparative studies (e.g., Grinsztajn et al., 2022; Shwartz-Ziv & Armon, 2021).

7.2.1 Linear Regression Model

Linear regression is a fundamental statistical method widely employed across various fields to model the relationship between a dependent variable and one or more independent variables. Its objective is to establish a linear equation that effectively captures the connection between these variables, enabling predictions and comprehension of the dependent variable's behavior based on the independent ones. Despite its extensive use, traditional linear regression is susceptible to challenges like overfitting, multicollinearity, and model instability when handling high-dimensional data or correlated predictor variables. To address these concerns, regularization techniques such as lasso (L1 regularization), ridge (L2 regularization), and elastic net (L1 and L2 regularization) have emerged as pivotal approaches within the domain of linear regression. These techniques introduce penalty terms into the ordinary least squares cost function shown in Equation 7.1, facilitating the selection of pertinent predictor variables and enhancing model generalization performance.

The expressions of the lasso, ridge, and elastic net loss functions are detailed in the following equations (James et al., 2013).

Lasso

Suppose the linear equation is to be defined as:

$$\hat{y} = \beta_0 + \sum\nolimits_{j=1}^{p} \beta_j x_j \qquad \text{(Eq. 7.1)}$$

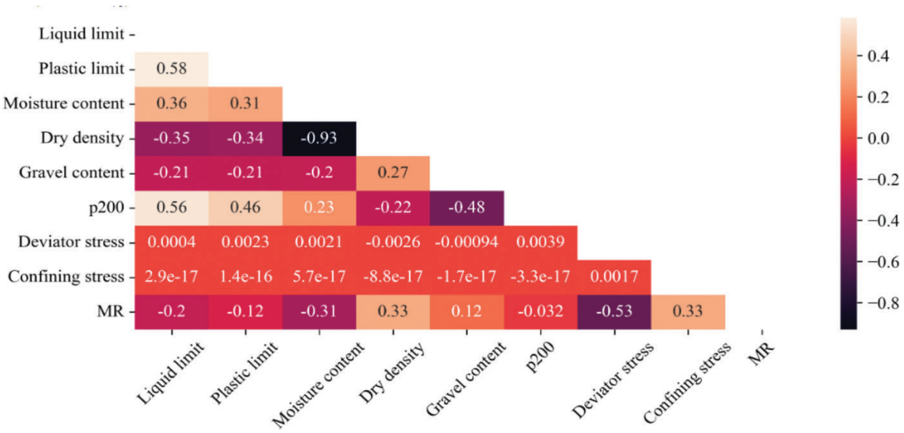


Figure 7.1 Feature correlation heatmap.

where $\hat{y}$ is the estimated value; β_0 is the linear model intercept; β_j are the model weights; and x_j represent the independent or explanatory variables.

The best fitting model is characterized by the lowest residual sum of squares (RSS) as shown in Equation 7.2.

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 \qquad \text{(Eq. 7.2)}$$

While lasso regression penalizes the absolute values of the residuals by incorporating a coefficient, as shown in Equation 7.2, Ridge regression adds a tuning parameter (λ ≥ 0) to the Residual Sum of Squares (RSS), as demonstrated in Equation 7.3.

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad \text{(Eq. 7.3)}$$

Lastly, elastic net combines both squared and absolute penalties on the RSS, as presented below.

$$RSS + \lambda \sum_{j=1}^{p} \left|\beta_j^2\right| \qquad \text{(Eq. 7.4)}$$

$$RSS + \alpha \Big[\sum_{j=1}^{p} \beta_j^2 + \sum_{j=1}^{p} |\beta_j| \qquad \text{(Eq. 7.5)}$$

The developed model using linear regression is shown below.

$$\begin{aligned} M_R = {} & -2{,}028.7126 + (-37.5318) * LL + (27.0172) \\ & * PL + (-9.4884) * Moisture + (60.7980) * dd \\ & + (28.8646) * Gravel + (18.2576) * p200 \\ & + (-298.7301) * \sigma_d + (318.0464) * \sigma_3 \end{aligned} \qquad \text{(Eq. 7.6)}$$

where LL: liquid limit; PL: plastic limit; moisture: moisture content; dd: dry density (pcf); gravel: gravel percent; p200: percent passing sieve No. 200; σ_d: deviator stress (psi); σ_3: confining stress (psi).

7.2.2 Polynomial Regression

Polynomial regression represents an extension of the traditional linear regression framework, tailored to address nonlinear relationships between variables. While linear regression models assume a linear association between the dependent and independent variables, real-world phenomena often exhibit more complex behaviors that cannot be effectively characterized by straight lines. Polynomial regression overcomes this limitation by incorporating polynomial functions of various degrees, enabling the model to fit intricate data patterns and yield more accurate predictions. The developed 2nd degree polynomial regression models are as follows.

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{m=1}^{p} \beta_m x_m^2 + \sum_{n=1}^{p} \beta_n x_n x_{n+1} \qquad \text{(Eq. 7.7)}$$

where $\hat{y}$ is the estimated value; β_0 is the model intercept; β_j are the original features' weights; x_j represent the features; β_m are the transformed features' weights; x_m^2 represent the features after transforming to a second degree; and β_n are the interaction terms' weights.

The polynomial $\mathrm{M_R}$ model is shown below.

$$\begin{aligned} M_R = {} & -12{,}208.4987 + (-99.4841) * LL + (-160.6817) \\ & * PL + (155.1812) * dd + (143.8614) * Gravel \\ & + (258.3643) * p200 + (197.8345) * \sigma_d + (-0.2500) \\ & * LL^2 + (6.0333) * LL * PL + (-0.1052) * LL \\ & * Moisture + (-0.3206) * LL * dd + (0.7471) * LL \\ & * Gravel + (-0.1625) * LL * p200 + (3.2016) * LL * \sigma_d \\ & + (-2.4054) * LL * \sigma_3 + (-10.8483) * PL^2 + (13.8025) \\ & * PL * Moisture + (-5.0184) * PL * dd \\ & + (43.3871) * PL * Gravel + (6.5280) * PL * p200 \\ & + (-6.5148) * PL * \sigma_d + (-6.5148) * PL * \sigma_d \\ & + (-1.8836) * PL * \sigma_3 + (-2.2455) * Moisture^2 \\ & + (-3.1656) * Moisture * dd + (9.6691) * Moisture \\ & * Gravel + (0.9306) * Moisture * p200 + (-0.6171) \end{aligned}$$

$$
\begin{aligned}
&\ast \mathit{Moisture} \ast \sigma_d + (-3.7327) \ast \mathit{Moisture} \\
&\ast \sigma_3 + (0.3736) \ast dd^2 + (-2.3778) \ast dd \ast \mathit{Gravel} \\
&+ (0.3897) \ast dd \ast p200 + (-7.4236) \ast dd \ast \sigma_d \\
&+ (3.5897) \ast dd \ast \sigma_3 + (4.2583) \ast \mathit{Gravel}^2 \\
&+ (-14.9271) \ast \mathit{Gravel} \ast p200 + (0.2191) \\
&\ast \mathit{Gravel} \ast \sigma_d + (4.1249) \ast \mathit{Gravel} \ast \sigma_3 + (-2.5449) \\
&\ast p200^2 + (-0.4704) \ast p200 \ast \sigma_d + (0.8532) \ast p200 \\
&\ast \sigma_3 + (37.1562) \ast {\sigma_d}^2 + (-35.5576) \ast \sigma_d \\
&\ast \sigma_3 + (38.7649) \ast {\sigma_3}^2 \qquad \text{(Eq. 7.8)}
\end{aligned}
$$

where LL: liquid limit; PL: plastic limit; moisture: moisture content; dd: dry density (pcf); gravel: gravel percent; p200: percent passing sieve No. 200; σ_d: deviator stress (psi); σ_3: confining stress (psi).

The polynomial model is assessed using the same RSS method as outlined in the linear regression section.

7.2.3 Decision Tree

Decision Trees (DT) are a popular machine learning algorithm for classification and regression tasks. They are a tree-based model that recursively partitions the input space into smaller regions, forming a tree-like structure. In the context of regression problems, DTs aim to predict a continuous target variable based on the input features.

The fundamental concept behind DTs is to recursively split the data into smaller subsets based on the input features, creating a tree-like structure. Each internal node of the tree represents a decision based on a specific feature, and the branches stemming from that node correspond to the possible values or ranges of that feature. The splitting process continues until a stopping criterion is met, such as reaching a maximum depth or a minimum number of instances in a node (Song & Lu, 2015).

The splitting criteria used in DTs for regression tasks are typically based on measures that evaluate the impurity or heterogeneity of the target variable within each node. Common measures include the mean squared error (MSE) or the variance of the target variable. The goal is to find the feature and split point that minimizes the impurity or heterogeneity in the resulting child nodes (Rokach & Maimon, 2005).

Once the tree is constructed, the prediction for a new instance is made by traversing the tree from the root node to a leaf node, following the decision path based on the instance's feature values. The predicted value at the leaf node is typically the mean or median of the target variable for the instances that reach that leaf.

7.2.4 Random Forest

Random Forest (RF) is a powerful ensemble learning method that combines multiple decision trees to improve predictive performance and robustness, especially in regression tasks. Proposed by Breiman (2001), RF represents a significant advancement in tree-based models, addressing the issue of overfitting and reducing the variance associated with individual decision trees.

The core idea behind RF is to construct a group of decision trees, each trained on a bootstrap sample of the original training data (Breiman, 1996). Additionally, a random subset of features is considered for each split during the tree construction process, introducing randomness and correlating the individual trees (Ho, 1998). This randomization process helps reduce the ensemble's variance and improve generalization.

When making predictions for a new instance, the RF algorithm aggregates the predictions from all the individual trees. For regression tasks, the final prediction is typically the average of the predictions from all trees (Breiman, 2001). This ensemble approach significantly reduces the risk of overfitting and enhances the model's overall predictive accuracy and stability.

One of the advantages of RF is its ability to handle high-dimensional data and automatically capture nonlinear relationships and interactions between features (Louppe, 2014). Additionally, RF provides an estimate of feature importance, which can be valuable for feature selection and interpretation (Breiman, 2001; Genuer et al., 2010).

7.2.5 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a class of machine learning models inspired by the biological neural networks found in the human brain. This approach has garnered widespread attention for its ability to discern complex data patterns and relationships.

ANNs consist of layers of interconnected nodes, or neurons, each contributing to the network's ability to process information. The fundamental principles underlying ANNs involve computing weighted sums of inputs, followed by applying nonlinear activation functions to produce outputs. This architecture enables ANNs to model intricate nonlinear relationships between input and output variables (Goodfellow et al., 2016; LeCun et al., 2015).

Within ANNs, various components play pivotal roles in shaping their functionality. Neurons serve as the basic processing units, computing weighted sums of inputs and applying activation functions to generate outputs. These connections between neurons are defined by weights, representing the strength of influence of one neuron on another. Additionally, biases enable neurons to introduce shifts in activation functions, further enhancing the network's expressiveness. ANNs typically comprise multiple layers, each performing distinct transformations on input data. The choice of activation function, such as sigmoid, tanh, or Rectified Linear Unit (ReLU), introduces crucial nonlinearities into the network, facilitating the learning of complex relationships (Bishop, 1995/2023;

Rumelhart et al., 1986). In this study, the ReLU activation function was utilized to develop the ANN model. This function outputs the input directly if it is positive; otherwise, it outputs zero. ReLU was chosen due to its efficiency in overcoming the vanishing gradient problem, which is often encountered with other activation functions such as sigmoid or tanh (Glorot et al., 2011).

7.2.6 XGBoost Algorithm

XGBoost stands as a prominent and powerful machine learning algorithm, recognized for its efficacy in predictive modeling and its capability to handle diverse data types and complexities. Originating from the gradient boosting framework, XGBoost refines and extends this methodology through the integration of regularization techniques, adaptive learning rates, and novel data handling mechanisms. By aggregating a multitude of weak learners, XGBoost constructs a robust and accurate predictive model, capable of addressing both regression and classification tasks.

The XGBoost algorithm builds an ensemble model by iteratively adding decision trees to the ensemble. Each decision tree is designed to correct the errors of the previous trees, with greater emphasis on the instances that were inaccurately predicted. The final prediction is a weighted sum of the predictions from all individual trees. XGBoost's objective function combines a loss term, representing the discrepancy between predicted and actual values, with regularization terms to prevent overfitting. The objective function, to be minimized, is defined as:

$$Obj(\theta) = RSS + \sum_{i=1}^{n} \Omega(f_n) \qquad \text{(Eq. 7.9)}$$

where θ represents the set of model parameters, RSS is the residual sum of squares, and $\Omega(f_n)$ is the regularization term for each tree (f_n).

The boosting process of the XGboost model could be summarized in the following six steps as presented in the algorithm's documentations (Chen & Guestrin, 2016):

Step 1: Initial Prediction

Start with initial prediction for each data point, this is often the mean value of the target variable as shown in Equation 7.10.

$$\hat{y}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{(Eq. 7.10)}$$

Step 2: Compute Residuals

The error $r_{i,t}$ represents the error in the predictions at each step t, which is calculated as shown in Equation 7.11.

$$r_{i,t} = y_i - \hat{y}_{t,t-1} \qquad \text{(Eq. 7.11)}$$

where y_i is the actual value for the i^{th} observation, and $\hat{y}_{t,t-1}$ is the predicted value for the i^{th} observation at the previous step (*t-1*).

Step 3: Fit a Tree to Residuals

A regression tree is fitted to the residuals to find the tree structure and leaf node values that minimize the function presented in Equation 7.12.

$$\sum_{i=1}^{n} (r_{i,t} - f(x_i))^2 + \Omega(f_t) \qquad \text{(Eq. 7.12)}$$

where $\Omega(f_t)$ is the regularization term for the tree (f_t)

Step 4: Compute Optimal Weights for Leaves

Once the tree structure is fixed, the optimal weight (w_j*) for each leaf (*j*) minimizes the following term:

$$w_j^* = \frac{\sum r_{i,t}}{\sum 1 + \lambda} \qquad \text{(Eq. 7.13)}$$

where λ is a regularization parameter that shrinks the leaf weights and helps prevent overfitting.

Step 5: Update the Model

The model is updated as shown in Equation 7.14.

$$\hat{y}_{i,t} = \hat{y}_{t,t-1} + \eta f_t(x_i) \qquad \text{(Eq. 7.14)}$$

where η represents the learning rate, which helps in controlling the contribution of each tree.

Step 6: Iterate

Steps 2–5 are repeated for a predefined iteration number (*T*), or until other stopping criteria is met.

The Python code for the XGBoost model is provided in Appendix A.

8. MODEL EVALUATION

To enable the training of the models, the dataset was divided into two sets: 80% of the data was allocated for training, while the remaining 20% was set aside for testing. To enhance robustness and minimize bias, five distinct random seeds were employed for conducting the data split. By comparing the model predictions across these five random divisions, the consistency and stability of the model's performance under varying data configurations were evaluated. This approach ensured that the model's performance wasn't disproportionately influenced by specific data splits or random initialization, providing a more dependable assessment of its ability to generalize. Model performance was evaluated using R^2 (R-squared), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error); explained as follows.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad \text{(Eq. 8.1)}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad \text{(Eq. 8.2)}$$

TABLE 8.1
Performance summary of the developed ML models (all stress levels)

	Training Set			Testing Set			
Model	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**Training Time (sec)**
Linear	0.51	1,123	865	0.55	1,117	866	0.004
2nd Degree Polynomial	0.68	900	670	0.60	1,049	760	
Decision Tree	0.66	938	694	0.45	1,234	934	0.1
Random Forest	0.83	661	505	0.65	987	725	8
ANN	0.61	1,000	770	0.58	1,077	838	12
XGBoost	*0.97*	*261*	*184*	*0.95*	*366*	*274*	*1.2*

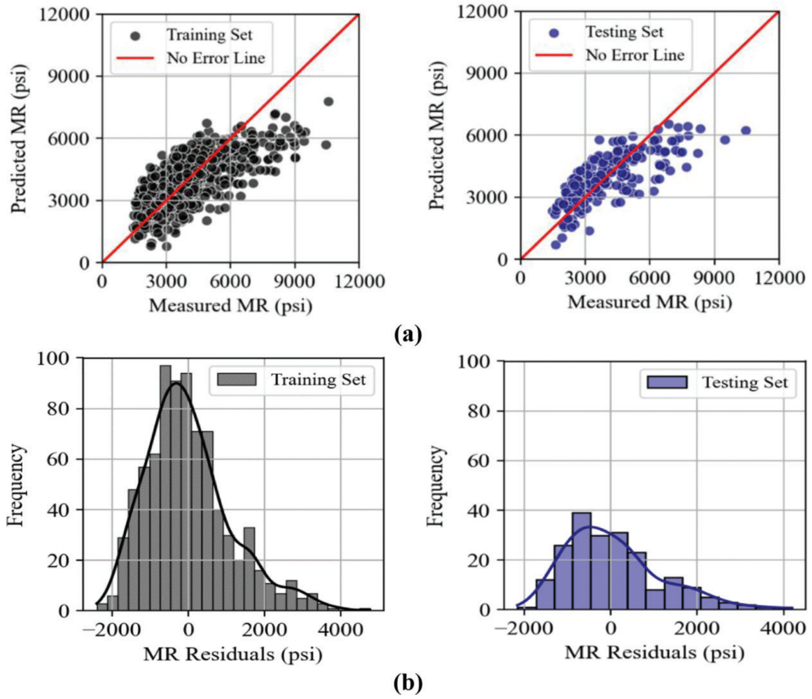


Figure 8.1 Performance of the linear regression model on the training and testing set: (a) scatter plots, and (b) histograms of residuals.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{(Eq. 8.3)}$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, $\bar{y}_i$ is the mean of actual values, and n is the number of observations.

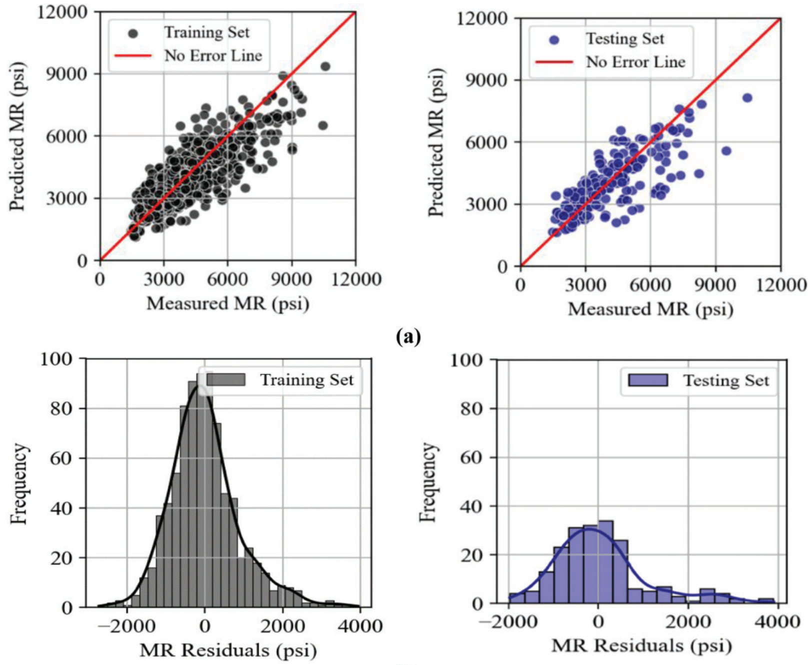
It should be noted that the performance of the models was evaluated under two stress combination scenarios: (1) all 15 stress level combinations in RLT testing; and (2) a common real-world pavement design scenario with a confining stress of 2 psi and deviator stress of 6 psi. Figures 8.2 through 8.6 present scatter plots and residual histograms showing the performance of the developed models on training and testing data.

A comprehensive representation of the training and testing performance for all developed models is outlined in Tables 8.1 and 8.2.

As shown in the model performance summary, XGBoost model show a robust and stable performance for all stress levels as well as a single stress level for both the training and the testing sets. The following figures shows the scatter plots and the residual plots for each model. The Python code for the developed XGBoost model is provided in Appendix A.

TABLE 8.2
Performance summary of the developed ML models (single stress levels)

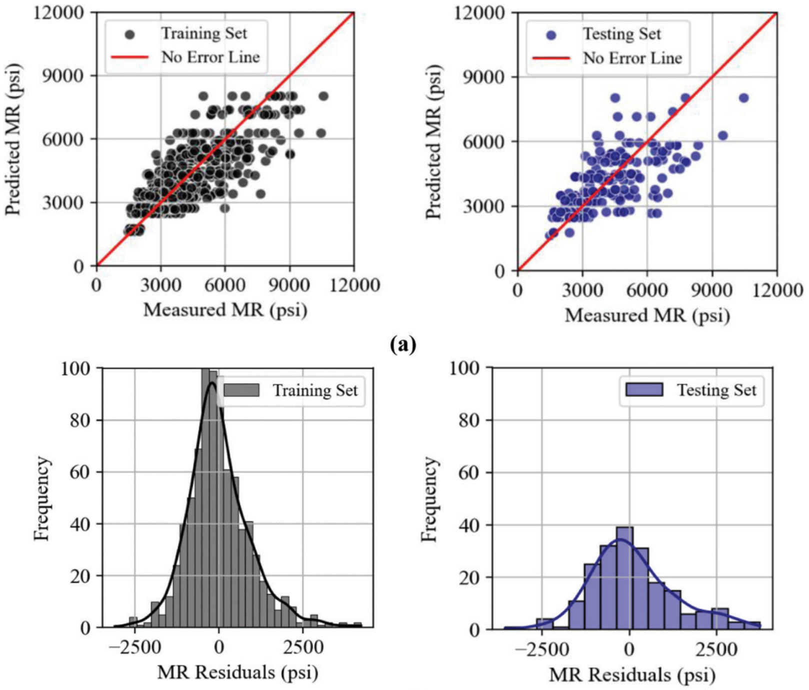
	Training Set			Testing Set		
Model	**R^2**	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**
Linear	0.51	1,123	865	0.10	823	718
2nd Degree Polynomial	0.68	900	670	0.36	693	584
Decision Tree	0.66	938	694	0.20	778	547
Random Forest	0.83	661	505	0.51	605	463
ANN	0.61	1,000	770	0.03	855	752
XGBoost	*0.97*	*261*	*184*	*0.92*	*235*	*176*



(a)

(b)

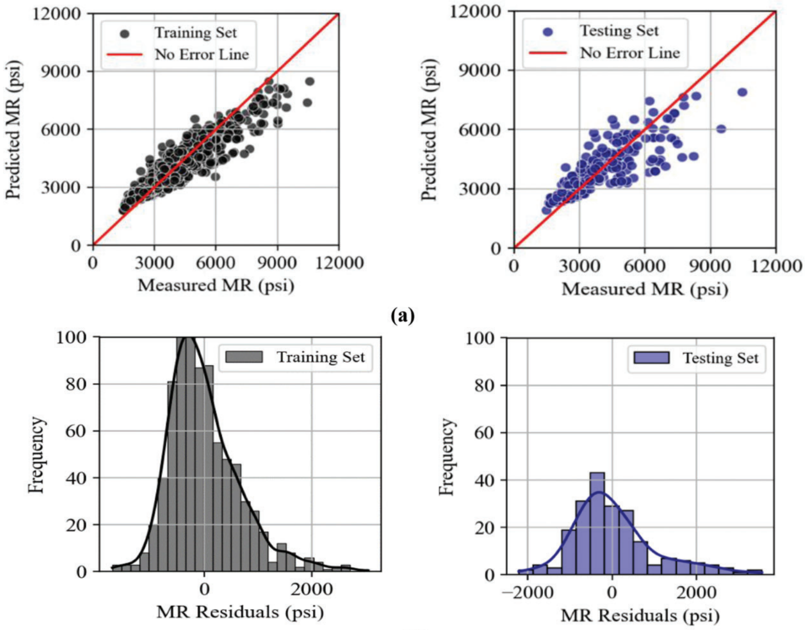
Figure 8.2 Performance of the 2nd degree polynomial regression model on the training and testing set: (a) scatter plots, and (b) histograms of residuals.



(a)

(b)

Figure 8.3 Performance of the decision trees model on the training and testing set: (a) scatter plots, and (b) histograms of residuals.



(a)

(b)

Figure 8.4 Performance of the random forest model on the training and testing set: (a) scatter plots, and (b) histograms of residuals.

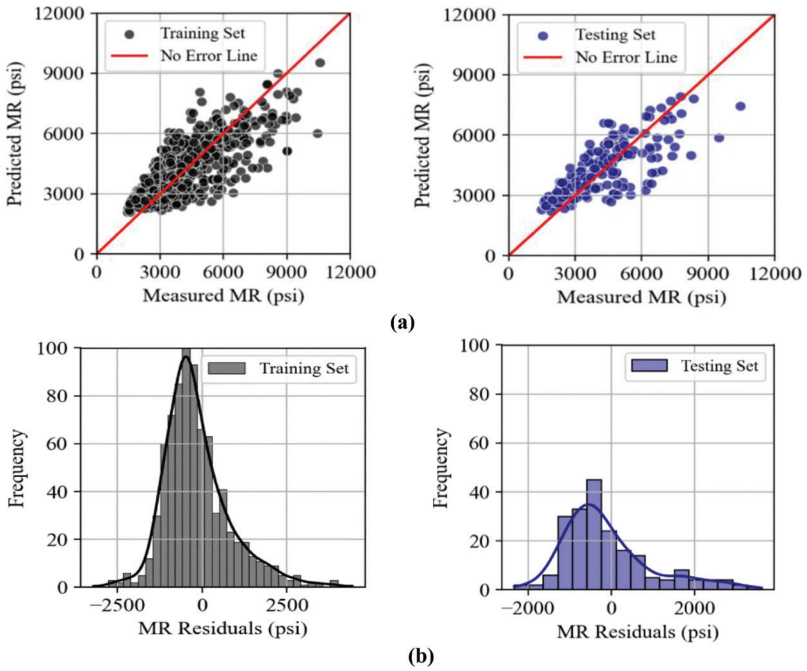


Figure 8.5 Performance of the artificial neural networks model on the training and testing set: (a) scatter plots, and (b) histograms of residuals.

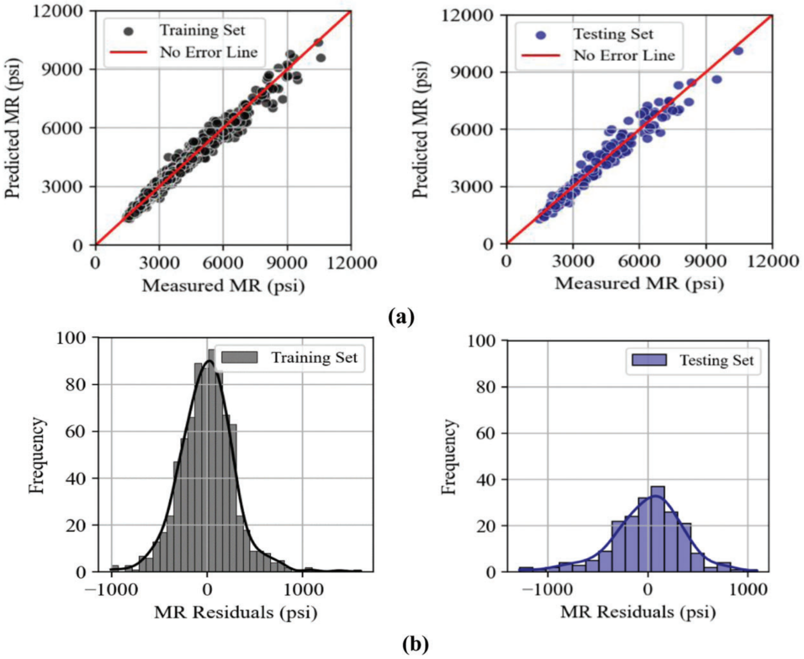


Figure 8.6 Performance of the XGBoost model on the training and testing set: (a) scatter plots, and (b) histograms of residuals.

9. BIAS-VARIANCE ANALYSIS

The bias-variance tradeoff is a fundamental principle in machine learning that balances a model's bias against its variance. Variance indicates how much the model's predictions vary across different training data subsets, with high variance suggesting sensitivity to noise. Bias measures deviation from the true values, with high bias leading to oversimplification and underfitting as shown in Figure 9.1.

Figure 9.2 shows the bias and variance calculated for all developed models using the RMSE metric for the training and testing sets. A larger RMSE gap between these error metrics with minimal training error indicates the overfitting of the model.

The bias-variance analysis showed that the decision tree (DT) model exhibits noticeable overfitting, as indicated by a significant gap between training and testing errors, unlike other models. The linear model, while showing no signs of overfitting, achieves lower accuracy. The Artificial Neural Network (ANN) also displayed no overfitting but had similarly lower accuracy. In contrast, the XGBoost model demonstrated superior accuracy with minimal disparity between training and testing errors, indicating robust performance without overfitting.

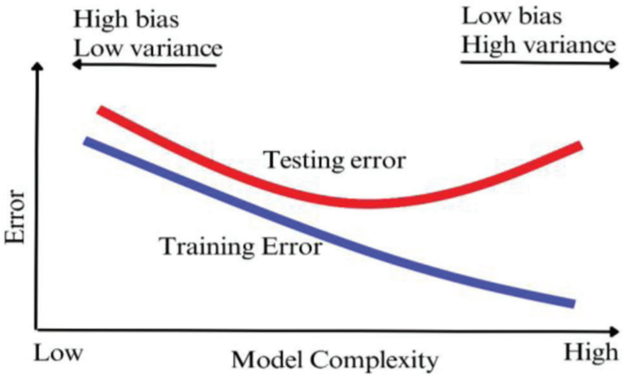


Figure 9.1 Bias-variance tradeoff illustration.

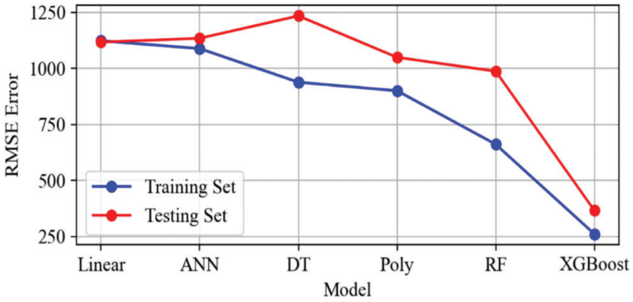


Figure 9.2 Bias-variance analysis summary.

10. DISCUSSIONS

10.1 Assessing Data Reduction Impact on Model Efficiency

To explore if the reduced dataset of 1,050 data points after data cleaning is sufficient to maintain the model's performance, a series of experiments were conducted. This approach involved training a linear model on five different subsets of the dataset, each concluding of 130 data points, and comparing their performance to the linear model trained on other subsets and the entire dataset. The results showed that the model's performance on each subset was approximately the same as its performance on the entire dataset. Additionally, all five subsets had approximately the same performance (see Table 10.1). This consistency indicates that the smaller data count did not negatively impact the model's effectiveness, demonstrating that the reduced dataset is sufficient for reliable modeling.

10.2 Error Analysis

In the context of error analysis from a geotechnical perspective, points with high residuals, defined as those exhibiting an absolute error of more than 1,000 psi between the actual and predicted M_R, were clustered and examined (Table 10.1). Each point was then matched with similar training set points based on Euclidean distance between input features. Next, the resilient modulus was calculated for each cluster, and the predicted M_R is compared to the M_R of similar points to evaluate the geotechnical reasonableness of the predictions.

Table 10.3 selects a sample point with a high residual from Table 10.2 and identifies similar points from the dataset—based on input feature similarity using Euclidean distance—to examine the corresponding M_R values of these comparable data points. As seen in the table, although these points are similar in input features, they have noticeably different M_R values, which may explain the high residual for the selected point. This underscores the importance of integrating domain expertise into the machine learning framework, rather than solely relying on error metrics without

TABLE 10.1
Subsets performance comparison

Subset	Data Points	Training R^2
Subset 1	130	0.40
Subset 2	130	0.35
Subset 3	130	0.34
Subset 4	130	0.39
Subset 5	130	0.36
Entire Cleaned Dataset	1,050	0.51

TABLE 10.2
High residuals points

Liquid Limit	Plastic Limit	Moisture Content (%)	Dry Density (*pcf*)	Gravel Content (%)	Passing Sieve200 (%)	Deviator Stress (*psi*)	Confining Stress (*psi*)	Predicted M_R (*psi*)	Measured M_R (*psi*)	Absolute Residual (*psi*)
31.2	16.3	22.4	103.6	5.6	60.5	2	4	5,618	4,611	1,007
42.3	20.9	21.7	106.3	2.4	83.04	2	6	7,752	8,781	1,029
30	20	25.8	97.3	1.9	83	2	4	5,787	6,878	1,091
36.7	16.8	22.1	102.1	0.7	80.6	2	6	8,294	9,422	1,128
32	18	19.5	108.1	0.7	67.8	2	4	5,743	4,615	1,128
31.2	16.3	22.4	103.6	5.6	60.5	4	4	4,940	3,812	1,128
30	20	25.8	97.3	1.9	83	4	6	5,195	6,354	1,159
44	21	23.2	99.9	6.3	87	2	4	7,049	8,221	1,172
27	14	17.5	115.1	4.8	55.5	2	2	6,692	5,498	1,194
38.4	18.8	33	87	12.1	46.3	2	6	6,412	7,705	1,293
44	21	23.2	99.9	6.3	87	2	6	8,144	9,486	1,342
28	15	22.4	104.6	21.8	57.4	2	6	8,119	9,480	1,361
27	15	19.7	105.1	3.1	74	2	4	6,917	8,317	1,400
47	20	19.9	104.8	0.2	94.9	2	2	5,226	3,743	1,483
47	19	18.2	113.1	0.9	83.2	2	6	8,949	10,434	1,485
52.1	20.8	27.7	95.2	2	79.11	2	4	5,426	6,931	1,505
49	20.4	23.8	100.4	0	95.75	2	6	6,712	8,320	1,608

TABLE 10.3
Sample point for error analysis

Liquid Limit	Plastic Limit	Moisture Content (%)	Dry Density (*pcf*)	Gravel Content (%)	Passing Sieve200 (%)	Deviator Stress (*psi*)	Confining Stress (*psi*)	Measured M_R (*psi*)	Euclidean Distance
42.3	**20.9**	**21.7**	**106.3**	**2.4**	**83.1**	**2**	**6**	**8,781**	**0**
51	22	23.9	102.8	0.1	98.9	2	6	5,044	2.004
46.5	20.2	23.5	101.4	0.1	81.2	2	6	3,766	2.160
47	19	24.7	99.4	8.4	78.2	2	6	4,777	2.188
45	18	22	99.8	3.1	73.7	2	6	4,354	2.375
46	21.6	19.6	104.7	1.6	60.9	2	6	7,044	2.452

Note: Boldface numbers are from Table 10.2. They are included as a point of comparison with the dataset.

further analysis. By incorporating geotechnical expertise, we can increase confidence and reliability in the developed model, ultimately leading to more informed decision-making in geotechnical applications.

11. ADDITIONAL MODEL DEVELOPMENT

To assess the efficacy of the constitutive model employed by INDOT for predicting M_R in A-4, A-6, and A-7 subgrade soils, a comprehensive evaluation of the existing model was conducted. The model, defined by Equation 11.1, utilizes regression coefficients (k_1, k_2, and k_3) specific to each test and determined by the testing device.

$$M_R = k_1 p_a \left(\frac{\theta}{p_a}\right)^{k_2} \left(\frac{\sigma_d}{p_a}\right)^{k_3} \qquad \text{(Eq. 11.1)}$$

where p_a is the atmospheric pressure, θ is the bulk stress, and σ_d is the deviator stress.

Upon closer examination, it was revealed that the coefficients derived from the testing device for each test showed considerable variability. This variability can result in errors in M_R predictions, with the model potentially overestimating or underestimating values, as clearly depicted in the scatter plots shown in Figure 11.1. This finding highlights the need for a more robust approach to ensure accurate M_R predictions.

To address this issue and explore the potential for improvements, a curve-fitting approach from the SciPy Python library was implemented. This method involves fitting a mathematical curve to the data points of the RLT test, allowing to determine the best-fitting coefficients for the constitutive model. By optimizing these coefficients through this procedure, the predicted M_R values would align more closely with the actual M_R, thus enhancing the model's predictive accuracy as shown in Figure 11.2.

Table 11.1 shows a comparison of the values of the original and the optimized k values for two random tests.

The M_R values predicted using the optimized coefficients demonstrated a stronger alignment with the

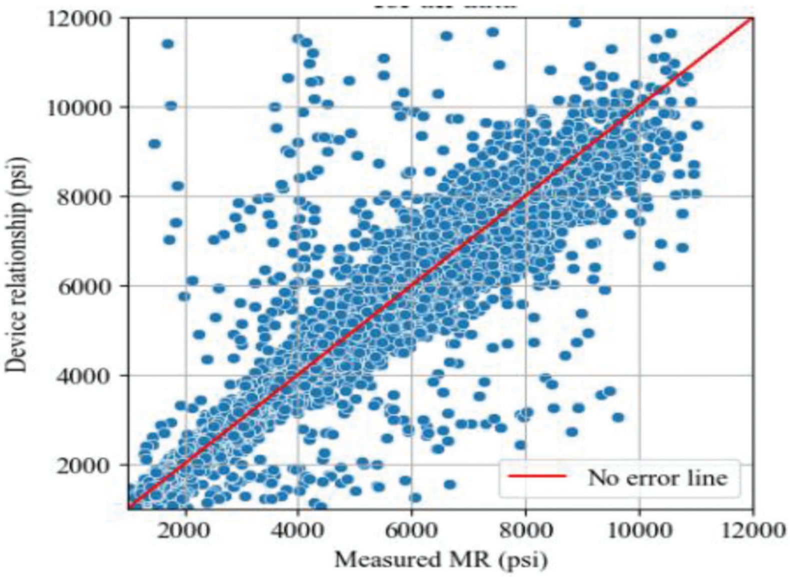


Figure 11.1 Measured M_R vs. model-predicted M_R using the RLT constitutive model.

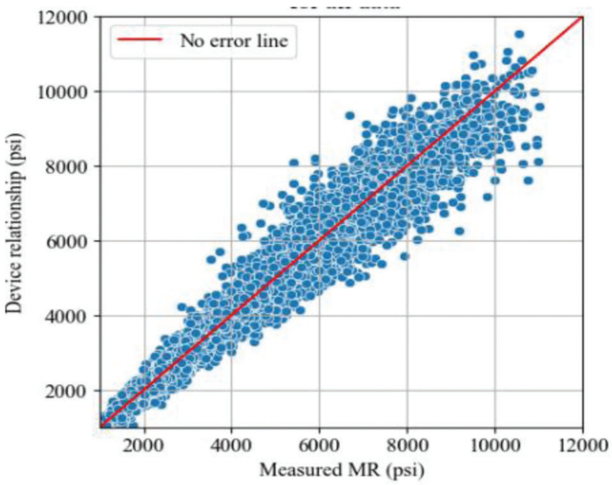
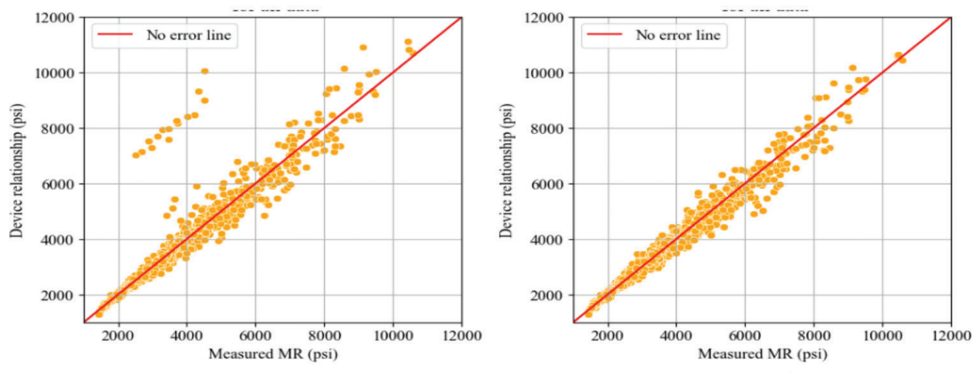


Figure 11.2 Measured M_R vs. model-predicted M_R using the optimized coefficients.

TABLE 11.1
One example of optimized k values

Coefficients of the Original Model			**Coefficients of the Optimized**		
k_1	k_2	k_3	k_1	k_2	k_3
156.96	0.826	-0.373	145.843	0.912	-0.4254
298.59	0.156	-0.325	308.318	0.125	-0.302

actual measured M_R values, indicating a more accurate representation of soil behavior. Notably, when focusing on the stress-softening portion of the data, the *k* values from the original device equation exhibited a remarkable consistency, with minimal scatter observed for most data points, as illustrated in Figure 11.3.



(a) **(b)**

Figure 11.3 Comparison of the performance of constitutive models in the RLT device: (a) the existing model, and (b) the updated model.

12. ADDITIONAL NOTES

It should be noted that the importance of soil features and stresses in M_R model development could differ based on the adopted model. Figures 12.1 and 12.2 show the heat map and bar plot of the normalized feature importance for each model.

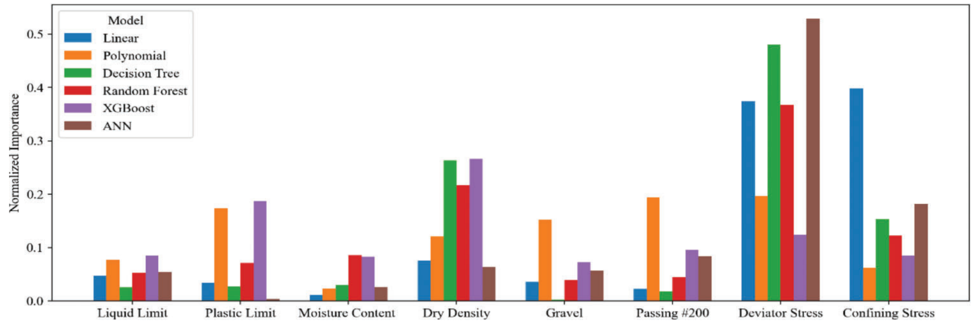


Figure 12.1 Comparison of feature importance in different models.

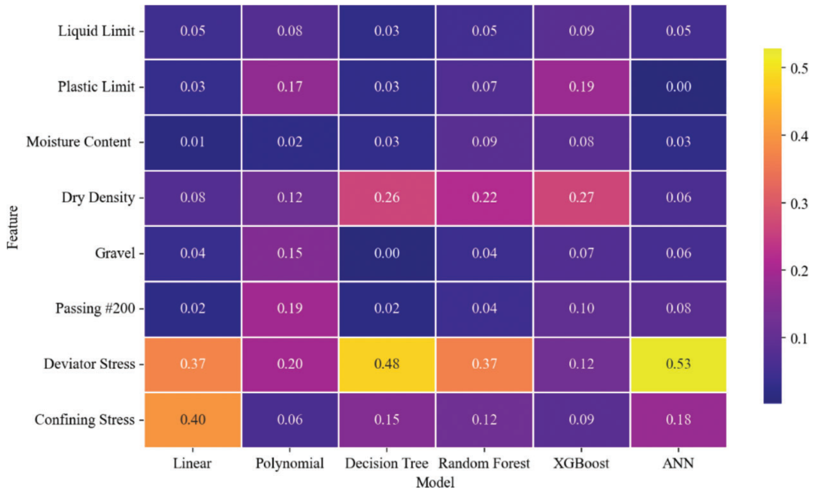


Figure 12.2 Feature importance heatmap for different ML models.

13. RECOMMENDATIONS

Based on the exploratory data analysis, data cleaning, and the error analysis conducted in this study, a list of recommendations is provided for future sampling, laboratory testing, and data collection.

13.1 Recommendations for Future Laboratory Testing

1. Ensure RLT results comply with AASHTO-T307 standards by including all 15 required stress levels.
2. Verify that the stress behavior plot generated from Repeated Load Testing (RLT) exhibits the characteristic stress response expected for the specific soil type, ensuring consistency with anticipated soil behavior.
3. Verify the range of measured M_R values for reasonability and consistency.
4. Conduct a thorough review of laboratory test results to confirm.
   - All necessary features are present and accounted for.
   - Liquid limit values are not less than moisture content values.
   - Sieve results add up to 100% to ensure accurate soil composition.
   - Soil classifications are accurate and align with test results.
   - Perform additional testing to determine unconfined compression strength, which has been shown to have a strong correlation with M_R in numerous studies (Azam et al., 2022; Pal & Deswal, 2014; Sadrossadat, Heidaripanah, & Ghorbani, 2016, Sadrossadat, Heidaripanah, & Osoul, 2016). Obtaining this value could lead to the development of more accurate and reliable models.

13.2 Recommendations for Future Sampling

1. Add latitude and longitude coordinates to the dataset to enable spatial analysis, which could be useful for the following:
   - examining the spatial distribution of soil properties and M_R values;
   - identifying patterns and correlations between soil characteristics and geographic location;

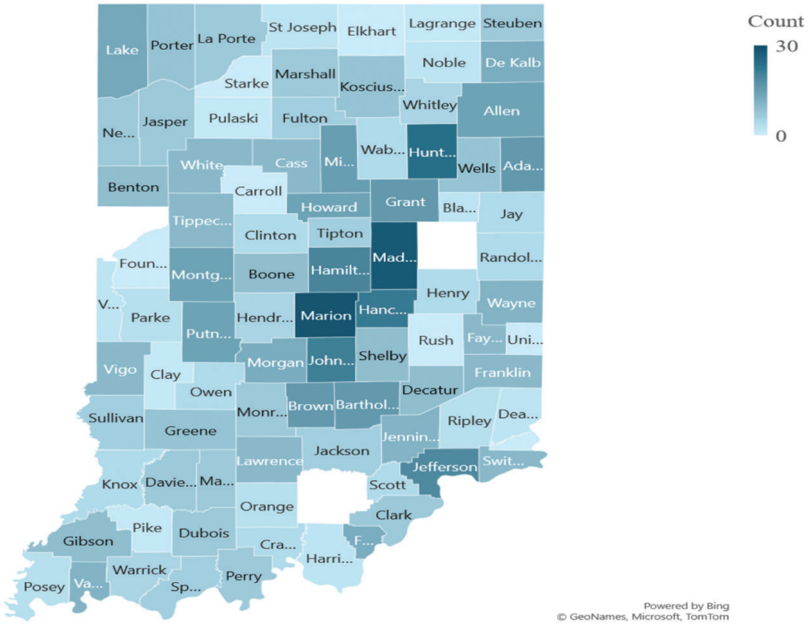


Figure 13.1 Current count of sampling locations.

TABLE 13.1
Counties with the highest population growth (Kinghorn, 2024)

County	Population Growth (%)	No. of Current M_R Locations	Notes
Hancock	3.7	22	–
Boone	2.4	9	More sampling recommended
Hendricks	1.8	5	More sampling recommended
Hamilton	1.7	20	–
Morgan	1.3	12	–
Johnson	1.3	21	–
Rush	1.0	0	More sampling recommended
White	1.0	10	–
Clark	1.0	7	More sampling recommended
Warrick	1.0	5	More sampling recommended

- investigating how geological changes, such as soil formation processes and tectonic activity, influence M_R values; and
- future development of predictive models that account for spatial variability in soil properties and M_R.

2. Strategically select future sampling sites.
   - Counties that are under sampled but now exhibit higher population growth, focusing on areas likely to require increased infrastructure development. Table 13.1 highlights the counties with the highest growth rates in 2023 and indicates whether each county has below-average soil sample locations. These counties are likely to need more soil investigation to cope with infrastructure development in the future.
3. After cleaning the data for the bagged sample test results, only five tests passed the sanity check and contained all necessary information. This quantity was insufficient for model development. To facilitate the development of M_R prediction model for bagged samples, it is recommended that more bagged samples be collected in the future.

13.3 Recommendations for the Use of Developed Models

The distribution of existing sampling locations across counties is illustrated in Figure 13.1. When using the models in unrepresented regions, caution should be exercised, and potential limitations acknowledged. If feasible, additional data should be collected through sampling and RLT testing to improve model performance and confidence.

REFERENCES

AASHTO. (2003). *Standard method of test for determining the resilient modulus of soils and aggregate materials*. American Association of State Highway and Transportation Officials.

AASHTO. (2015). *Mechanistic-empirical pavement design guide: A manual of practice* (1st ed.). American Association of State Highway and Transportation Officials.

AASHTO. (2020). *Mechanistic-empirical pavement design guide: A manual of practice* (3rd ed.). American Association of State Highway and Transportation Officials.

Allen, J. J. (1973). *The effects of non-constant lateral pressures on the resilient response of granular materials* [Doctoral dissertation, University of Illinois at Urbana-Champaign].

Alnedawi, A., Ullah, S., Azam, A., Mousa, E., Obaid, I., & Yosri, A. (2022). Integrated and holistic knowledge map of resilient pavement materials: A scientometric analysis and bibliometric review of research frontiers and prospects. *Transportation Geotechnics*, *33*, 100711.

Andrei, D., Witczak, M. W., & Houston, W. N. (2009). Resilient modulus predictive model for unbound pavement materials. *Contemporary Topics in Ground Modification, Problem Soils, and Geo-Support* (pp. 401–408). American Society of Civil Engineers. https://doi.org/10.1061/41023(337)51

ARA. (2004). *Guide for mechanistic-empirical pavement design of new and rehabilitated pavement structures* (NCHRP Project 1-37A). Applied Research Associates.

Azam, A., Bardhan, A., Kaloop, M. R., Samui, P., Alanazi, F., Alzara, M., & Yosri, A. M. (2022). Modeling resilient modulus of subgrade soils using LSSVM optimized with swarm intelligence algorithms. *Scientific Reports*, *12*, 14454. https://doi.org/10.1038/s41598-022-17429-z

Ba, M., Nokkaew, K., Fall, M., & Tinjum, J. (2013). Effect of matric suction on resilient modulus of compacted aggregate base courses. *Geotechnical and Geological Engineering*, *31*, 1497–1510.

Bishop, C. M. (1995/2023). *Neural networks for pattern recognition*. Oxford Academic. https://doi.org/10.1093/oso/9780198538493.001.0001

Boyce, J. R. (1976). *The behaviour of a granular material under repeated loading* [Doctoral dissertation, University of Nottingham]. Nottingham eTheses. https://eprints.nottingham.ac.uk/id/eprint/11853

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Brown, S., & Selig, E. T. (1994). The design of pavement and rail track foundations. *XIII ICSMFE*. ISSMGE.

Carmichael, R. F., III., & Stuart, E. (1978). Predicting resilient modulus: A study to determine the mechanical properties of subgrade soils. *Transportation Research Record*, *1043*, 145–148.

Cary, C. E., & Zapata, C. E. (2011). Resilient modulus for unsaturated unbound materials. *Road Materials and Pavement Design*, *12*(3), 615–638.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. arXiv. https://doi.org/10.48550/arXiv.1603.02754

Çoleri, E. (2007). *Relationship between resilient modulus and soil index properties of unbound materials* [Master's thesis, Middle East Technical University].

Dawson, A., Thom, N., & Paute, J. (1996). Mechanical characteristics of unbound granular materials as a function of condition. *Gomes Correia*. Balkema, Rotterdam.

Dehlen, G. L., & Monismith, C. L. (1970). Effect of nonlinear material response on the behavior of pavements under traffic. *Highway Research Board*.

Domitrović, J., Rukavina, T., & Lenart, S. (2019). Effect of freeze-thaw cycles on the resilient moduli and permanent deformation of RAP/natural aggregate unbound base mixtures. *Transportation Geotechnics*, *18*, 83–91.

Duong, T. V., Cui, Y.-J., Tang, A.-M., Dupla, J.-C., Canou, J., Calon, N., & Robinet, A. (2015). Effects of water and fines contents on the resilient modulus of the interlayer soil of railway substructure. *Acta Geotechnica*, 11(1).

Ekblad, J. (2008). Statistical evaluation of resilient models characterizing coarse granular materials. *Materials and Structures*, *41*, 509–525.

Elias, M. B., & Titi, H. H. (2006). Evaluation of resilient modulus model parameters for mechanistic–empirical pavement design. *Transportation Research Record*, *1967*(1), 89–100.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, *15*, 315–323.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). *Why do tree-based models still outperform deep learning on tabular data?* arxiv. https://doi.org/10.48550/arXiv.2207.08815

Hajj, E. Y., Thavathurairaja, J., Stolte, S., Sebaaly, P. E., & Motamed, R. (2018). *Resilient modulus prediction models of unbound materials for Nevada*. University of Nevada.

Han, Z., & Vanapalli, S. K. (2015). Model for predicting resilient modulus of unsaturated subgrade soil using soil-water characteristic curve. *Canadian Geotechnical Journal*, *52*(10), 1605–1619. https://doi.org/10.1139/cgj-2014-0339

Hanittinan, W. (2007). *Resilient modulus prediction using neural network algorithms*. The Ohio State University.

Hardcastle, J. H. (1992). *Subgrade resilient modulus for Idaho pavements* (FHWA Report No. RP110-D). Idaho Department of Transportation.

Hicks, R. G., & Monismith, C. L. (1971). Factors influencing the resilient properties of granular materials. *Highway Research Record*, *345*, 15–31.

Ho, T. K. (1998, August). The random subspace method for constructing decision forests. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Hoang, H.-G. T., Nguyen, T. A. (2022). An artificial intelligence approach to predict the resilient modulus of subgrade pavement or unbound material. In C. Ha-Minh, A. M. Tang, T. Q. Bui, X. H. Vu, & D. V. K. Huynh (Eds.) *CIGOS 2021, Emerging Technologies and Applications for Green Infrastructure. Lecture Notes in Civil Engineering, vol 203*. Springer. https://doi.org/10.1007/978-981-16-7160-9_177

Houston, W. N., Houston, S. L., & Anderson, T. W. (1993). Stress state considerations for resilient modulus testing of pavement subgrade. *Transportation Research Record*, *1406*, 124–132.

Jackson, K. D. (2015). *Laboratory resilient modulus measurements of aggregate base materials in Utah* [Master's thesis, Brigham Young University]. BYU Scholars Archive. https://scholarsarchive.byu.edu/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Ji, R., Siddiki, N., Nantung, T., & Kim, D. (2014). Evaluation of resilient modulus of subgrade and base materials in Indiana and its implementation in MEPDG. *The Scientific World Journal*, *Vol 2014*(1).

Kardani, N., Aminpour, M., Raja, M. N., Kumar, G., Bardhan, A., & Nazem, M. (2022). Prediction of the resilient modulus of compacted subgrade soils using ensemble machine learning methods. *Transportation Geotechnics*, *36*, 100827.

Khasawneh, M. A. (2019). Investigation of factors affecting the behaviour of subgrade soils resilient modulus using robust statistical methods. *International Journal of Pavement Engineering*, *20*(10), 1193–1206. https://doi.org/10.1080/10298436.2017.1394101

Khasawneh, M. A., & Al-jamal, N. F. (2019). Modeling resilient modulus of fine-grained materials using different statistical techniques. *Transportation Geotechnics*, *21*, 100263.

Kim, D., & Kim, J. R. (2007). Resilient behavior of compacted subgrade soils under the repeated triaxial test. *Construction and Building Materials*, *21*(7), 1470–1479.

Kim, D., & Siddiki, N. Z. (2005). *Simplification of resilient modulus testing for subgrades* (Joint Transportation Research Program FHWA/IN/JTRP-2005/23). West Lafayette: IN: Purdue University. https://doi.org/10.5703/1288284313388

Kinghorn, M. (2024). *Indiana population projections to 2060* Indiana Business Research Center, Indiana University Kelley School of Business. https://doi.org/www.ibrc.indiana.edu/ibr/2024/summer/article1.html

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Liang, R. Y., Rabab'ah, S., & Khasawneh, M. (2008). Predicting moisture-dependent resilient modulus of cohesive soils using soil suction concept. *Journal of Transportation Engineering*, *134*(1), 34–40.

Louppe, G. (2014). *Understanding random forests: From theory to practice* [Doctoral dissertation, University of Liège]. https://orbi.uliege.be/handle/2268/170309

Malla, R. B., & Joshi, S. (2008). Subgrade resilient modulus prediction models for coarse and fine-grained soils based on long-term pavement performance data. *International Journal of Pavement Engineering*, *9*(6), 431–444.

Mitry, F. (1965). *Determination of the modulus of resilient deformation of untreated base course materials* [Doctoral dissertation, University of California Berkeley].

Mohammad, L. N., Puppala, A. J., & Alavilli, P. (1995). Resilient properties of laboratory compacted subgrade soils. *Transportation Research Record*, *1504*, 87–102.

Nazzal, M. D., & Mohammad, L. N. (2010). Estimation of resilient modulus of subgrade soils for design of pavement structures. *Journal of Materials in Civil Engineering*, *22*(7).

Pal, M., & Deswal, S. (2014). Extreme learning machine based modeling of resilient modulus of subgrade soils. *Geotechnical and Geological Engineering*, *32*, 287–296.

Pezo, R., & Hudson, W. (1994). Prediction models of resilient modulus for nongranular materials. *Geotechnical Testing Journal*, *17*(3), 349–355. https://doi.org/10.1520/GTJ10109J

Qu, Y.-l., Chen, G.-l., Niu, F.-j., Ni, W.-k., Mu, Y.-h., & Luo, J. (2019). Effect of freeze-thaw cycles on uniaxial mechanical properties of cohesive coarse-grained soils. *Journal of Mountain Science*, *16*, 2159–2170.

Rahim, A. M. (2005). Subgrade soil index properties to estimate resilient modulus for pavement design. *Internation Journal of Pavement Engineering*, *6*(3), 163–169.

Regaya, Y., Fadli, F., & Amira, A. (2021). Point-Denoise: Unsupervised outlier detection for 3D point clouds enhancement. *Multimedia Tools and Applications*, *80*, 28161–28177. https://doi.org/10.1007/s11042-021-10924-x

Rokach, L., & Maimon, O. (2005). Decision Trees. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 165–192). Springer. https://doi.org/10.1007/0-387-25465-X_9

Rumelhart, D. E., Hinton, G., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536. https://doi.org/10.1038/323533a0

Saberian, M., & Li, J. (2021). Effect of freeze–thaw cycles on the resilient moduli and unconfined compressive strength of rubberized recycled concrete aggregate as pavement base/subbase. *Transportation Geotechnics*, *27*, 100477.

Sadrossadat, E., Heidaripanah, A., & Ghorbani, B. (2016). Towards application of linear genetic programming for indirect estimation of the resilient modulus of pavements subgrade soils. *Road Materials and Pavement Design*, *19*(1), 139–153.

Sadrossadat, E., Heidaripanah, A., & Osouli, S. (2016b). Prediction of the resilient modulus of flexible pavement subgrade soils using adaptive neuro-fuzzy inference systems. *Construction and Building Materials*, *123*, 235–247.

Seed, H. B., Chan, C. K., & Lee, C. E. (1962, August 20–24). *Resilience characteristics of subgrade soils and their relation to fatigue failures in asphalt pavements* [Conference session].

International Conference on the Structural Design of Asphalt Pavements, Ann Arbor, MI.

Shwartz-Ziv, R., & Armon, A. (2021). *Tabular data: Deep learning is not all you need*. arxiv. https://doi.org/10.48550/arXiv.2106.03253

Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135.

Sweere, G. T. H. (1990). *Unbound granular bases for roads* [Doctoral dissertation, Delft University of Technology]. TU Delft Repositories. http://resolver.tudelft.nl/uuid:1cc1c86a-7a2d-4bdc-8903-c665594f11eb

Thach Nguyen, B., & Mohajerani, A. (2014). Determination of CBR for fine-grained soils using a dynamic lightweight cone penetrometer. *International Journal of Pavement Engineering*, *16*(2), 180–189. https://doi.org/10.1080/10298436.2014.937807

Thom, N. (1988). *Design of road foundations* [Doctoral dissertation, University of Nottingham]. Nottingham eTheses. https://eprints.nottingham.ac.uk/id/eprint/10281

Uzan, J. (1985). Characterization of granular material. *Transportation Research Record*, *1022*, 52–59.

Witczak, M. W., & Uzan, J. (1988). *The universal airport pavement design system, Report I of V: Granular material characterization*. University of Maryland, College Park.

Yang, S.-R., Huang, W.-H., & Tai, Y.-T. (2005). Variation of resilient modulus with soil suction for compacted subgrade soils. *Transportation Research Record*, *1913*(1), 99–106. https://doi.org/10.1177/0361198105191300110

Zvonarić, M., Barišić, I., Galić, M., & Minažek, K. (2021). Influence of laboratory compaction method on compaction and strength characteristics of unbound and cement-bound mixtures. *Applied Sciences*, *11*(11), 4750.

APPENDICES

Appendix A. Python Code for the XGBoost Model

APPENDIX A. PYTHON CODE FOR THE XGBOOST MODEL

Python Code for the XGBoost Model

```
#Import necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler, StandardScaler, PowerTransformer, RobustScaler, QuantileTransformer
from sklearn.metrics.pairwise import euclidean_distances
from sklearn.metrics.pairwise import cosine_similarity
import xgboost as xgb
from sklearn.metrics import mean_squared_error,r2_score,mean_absolute_error
from xgboost import XGBRegressor
#Import dataset
df_r=pd.read_csv(‘\project\training_data.csv’)
#take only the part with LL>Moisture%
df_r=df_r[df_r['LL']>df_r['Moisture % ']]
#Exclude tests with MR >= 11,000 psi
# Group by 'INDOT Lab Number' and check if any 'MR' value is less than 11000
valid_lab_numbers = result_df.groupby('INDOT Lab Number')['MR'].max() <= 11000
# Filter the original DataFrame based on valid lab numbers
result_df = result_df[result_df['INDOT Lab Number'].isin(valid_lab_numbers[valid_lab_numbers].index)]
#Select only stress softening RLT tests
eliminated = pd.DataFrame(columns=df_r.columns)
good = pd.DataFrame(columns=df_r.columns)
# Get unique INDOT Lab Numbers
unique_tests = df_r['INDOT Lab Number'].unique()
# Loop through each unique test
for test in unique_tests:
    # Get the subset of the dataframe for the current test
    test_df = df_r[df_r['INDOT Lab Number'] == test]
    # Group by conf_stress and check if MR increases with dev_stress for the same conf_stress
    for conf_stress in test_df['rounded_conf_pandas'].unique():
        conf_df = test_df[test_df['rounded_conf_pandas'] == conf_stress]
        # Check if there is any increase in MR with an increase in dev_stress
        if (conf_df['MR'].diff() > 0).any():
            eliminated = pd.concat([eliminated, test_df])
            break  # Break the loop if the trend is not followed for this conf_stress
    else:
        good = pd.concat([good, test_df])
```

```
# Reset index for the resulting dataframes
eliminated.reset_index(drop=True, inplace=True)
good.reset_index(drop=True, inplace=True)

eliminated_k = pd.DataFrame(columns=good.columns)
good_k = pd.DataFrame(columns=good.columns)

# Get unique INDOT Lab Numbers
unique_tests = good['INDOT Lab Number'].unique()

# Loop through each unique test
for test in unique_tests:
    # Get the subset of the dataframe for the current test
    test_df = good[good['INDOT Lab Number'] == test]

    # Group by conf_stress and check if MR increases with dev_stress for the same conf_stress
    for conf_stress in test_df['rounded_dev_pandas'].unique():
        conf_df = test_df[test_df['rounded_dev_pandas'] == conf_stress]

        # Check if there is any increase in MR with an increase in dev_stress
        if (conf_df['MR'].diff() >= 0).any():
            eliminated_k = pd.concat([eliminated_k, test_df])
            break  # Break the loop if the trend is not followed for this conf_stress
    else:
        good_k = pd.concat([good_k, test_df])
# Reset index for the resulting dataframes
eliminated_k.reset_index(drop=True, inplace=True)
good_k.reset_index(drop=True, inplace=True)
#check the shape of the developed dataframes
eliminated.shape, good.shape, eliminated_k.shape,good_k.shape

#Model development
df= good_k
def split_data_with_similar_statistics(data, test_size=0.2, max_attempts=1000):
    # Step 1: Calculate statistics for each feature in the entire dataset
    feature_statistics = data.describe().loc[['mean', 'std']]
    # Step 2: Randomly shuffle the dataset
    data = data.sample(frac=1, random_state=42).reset_index(drop=True)
    # Step 3: Split the dataset into training and testing sets
    num_samples = len(data)
    split_index = int((1 - test_size) * num_samples)
    train_data, test_data = data[:split_index], data[split_index:]
    # Step 4: Check if the statistics of each feature in the training and testing sets are similar
    attempts = 1
    while attempts <= max_attempts:
        train_statistics = train_data.describe().loc[['mean', 'std']]
        test_statistics = test_data.describe().loc[['mean', 'std']]

        # Check if the absolute difference between train and test statistics is within a tolerance level
        tolerance = 0.1
        is_similar = np.allclose(train_statistics, test_statistics, rtol=tolerance, atol=tolerance)
        if is_similar:
            break
        else:
            # Reshuffle the data and try again
            data = data.sample(frac=1, random_state=42).reset_index(drop=True)
            train_data, test_data = data[:split_index], data[split_index:]
            attempts += 1
    if attempts > max_attempts:
        print("Warning: Maximum attempts reached. Statistics may not be perfectly matched.")
    return train_data, test_data
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
import numpy as np

def regression_metrics(model, X_train, y_train, X_test, y_test):
    """
    Compute and print regression metrics for both training and testing sets.
    Parameters:
    - model: Fitted regression model.
    - X_train, X_test: Training and testing feature matrices.
    - y_train, y_test: Training and testing target vectors.
    """
    # Training set predictions
    y_train_pred = model.predict(X_train)
    # Testing set predictions
    y_test_pred = model.predict(X_test)
    # Regression metrics
    metrics_names = [
        "R^2", "MSE", "RMSE", "MAE", "MAPE", "Performance Index (PI)", "Scatter Index (SI)"
    ]
    metrics_values_train = [
        r2_score(y_train, y_train_pred),
        mean_squared_error(y_train, y_train_pred),
        np.sqrt(mean_squared_error(y_train, y_train_pred)),
        mean_absolute_error(y_train, y_train_pred),
        np.mean(np.abs((y_train - y_train_pred) / y_train)) * 100,
        np.sum(np.abs(y_train - y_train_pred)) / np.sum(np.abs(y_train)) * 100,
        np.sqrt(np.sum((y_train - y_train_pred)**2)) / np.sqrt(np.sum((y_train - np.mean(y_train))**2)) * 100
    ]
    metrics_values_test = [
        r2_score(y_test, y_test_pred),
        mean_squared_error(y_test, y_test_pred),
        np.sqrt(mean_squared_error(y_test, y_test_pred)),
        mean_absolute_error(y_test, y_test_pred),
        np.mean(np.abs((y_test - y_test_pred) / y_test)) * 100,
        np.sum(np.abs(y_test - y_test_pred)) / np.sum(np.abs(y_test)) * 100,
        np.sqrt(np.sum((y_test - y_test_pred)**2)) / np.sqrt(np.sum((y_test - np.mean(y_test))**2)) * 100
    ]
    # Print metrics
    print("Metrics for Training Set:")
    for name, value in zip(metrics_names, metrics_values_train):
        print(f"{name}: {value:.4f}")
    print("\nMetrics for Testing Set:")
    for name, value in zip(metrics_names, metrics_values_test):
        print(f"{name}: {value:.4f}")
#Select input features
good_feat=df[['INDOT Lab Number','LL', 'PL', 'Moisture % ', 'dd (pcf)', 'Gravel', 'p200', 'rounded_dev_pandas',
'rounded_conf_pandas', 'MR']]
good_feat['rounded_conf_pandas']=good_feat['rounded_conf_pandas'].astype(int)
good_feat['rounded_dev_pandas']=good_feat['rounded_dev_pandas'].astype(int)
train,test=split_data_with_similar_statistics(good_feat)
X_train=train.drop(['MR'],axis=1)
y_train=train['MR']
X_test=test.drop(['MR'],axis=1)
y_test=test['MR']

# Create the optimized XGBoost Regressor
ga_xgboost = XGBRegressor(
    n_estimators=int(184),
    max_depth=int(2),
    learning_rate=0.9680342,
    random_state=101
)

# Fit the model
```

```
xgboost_model = ga_xgboost.fit(X_train.drop(['INDOT Lab Number'], axis=1), y_train)
xgboost_model.feature_importances_
regression_metrics(xgboost_model, X_train.drop(['INDOT Lab Number'],axis=1), y_train, X_test_single.drop(['INDOT Lab Number'],axis=1), y_test_single)
```

About the Joint Transportation Research Program (JTRP)

On March 11, 1937, the Indiana Legislature passed an act which authorized the Indiana State Highway Commission to cooperate with and assist Purdue University in developing the best methods of improving and maintaining the highways of the state and the respective counties thereof. That collaborative effort was called the Joint Highway Research Project (JHRP). In 1997 the collaborative venture was renamed as the Joint Transportation Research Program (JTRP) to reflect the state and national efforts to integrate the management and operation of various transportation modes.

The first studies of JHRP were concerned with Test Road No. 1—evaluation of the weathering characteristics of stabilized materials. After World War II, the JHRP program grew substantially and was regularly producing technical reports. Over 1,600 technical reports are now available, published as part of the JHRP and subsequently JTRP collaborative venture between Purdue University and what is now the Indiana Department of Transportation.

Free online access to all reports is provided through a unique collaboration between JTRP and Purdue Libraries. These are available at http://docs.lib.purdue.edu/jtrp.

Further information about JTRP and its current research program is available at http://www.purdue.edu/jtrp.

About This Report

An open access version of this publication is available online. See the URL in the citation below.

Khoshnevisan, S., Norouzi, M., & Sadik, L. (2024). *Use of machine learning methods to obtain a reliable predictive model for resilient modulus of subgrade soil* (Joint Transportation Research Program Publication No. FHWA/IN/JTRP-2024/27). West Lafayette, IN: Purdue University. https://doi.org/10.5703/1288284317768