Machine Learning strategy in comparison to Physics-based models to predict the Resilient Modulus response of Soil-polymer composites under triaxial and cyclic loads

Ruan A. Carvalho[1], Francisco E. Junior[2]

[1]*Department of Civil and Environmental engineering, University of Brasilia*
Darcy Ribeiro,70910-900, Brasilia - DF, Brazil
runcarvalho@gmail.com, fejr.unb@gmail.com

Abstract. The objective of the work is to propose a Machine Learning (ML) strategy to predict the Resilient Modulus (RM) for soil-polymer composite materials. It is developed regression models capable of accurately predicting the material stiffness under cyclical tests based not only on traditional predictor variables defining the stress state but also incorporating information such as curing time and polymer dosage in the composite. This strategy aims to answer the question of whether ML data-based models, considering a larger range of independent variables, can perform better than various physics-based constitutive models specialized to specific ranges of the independent variables. Gaussian Process Regression (GPR) models are trained from data from triaxial cyclic load tests performed on specimens of different polymer dosages and curing times. The predictor variables are confining stress, deviator stress, curing time, and percentage of polymer incorporated into the composite, whereas the Resilient Modulus (RM) is the response variable. The optimization of hyperparameters and model performance measurements were employed using cross-validation methods. The results show that the accuracy of the ML models is competitive and, in some cases, better than the ones provided by the physics-based constitutive models traditionally used to model the RM response.

Keywords: Machine Learning, Data-based constitutive models, Resilient modulus, Soil-polymer composites.

1 Introduction

In Brazil, there are special areas called Environmental Protection Areas (APA's) that aim to reconcile conservation of nature with sustainable use of natural resources. Some of these areas, *e.g.*, Jalapão in Tocantins, show great potential to sustainable tourism development [1], but the region can only be accessed and explored through dirt roads. These types of roads are highly affected by erosion and rains, thus requiring constant maintenance, what causes environmental damage to some degree [2]. In this context, the implementation of alternative materials can lower maintenance frequency and environmental damage and improve the general performance of dirt roads. Soil-polymer composites can be used for this purpose, although requiring a better understanding of the mechanical behavior of these materials.

Constitutive models based on traditional stress state variables, such as the ones developed by Silva [3], can perform good predictions of the Resilient Modulus (RM) of soil-polymer composites, but cannot incorporate important characteristics of the material, such as curing time and dosage of polymer. Sharifzadeh *et al.* [4] successfully developed GPR models to forecast wind power and solar power, the optimization of these models was done by establishing the best kernel function that could be used to predict the response variable. Also, Zhao *et al.* [5] developed GPR models to predict sulfate content of lakes in China, using 5-fold cross-validation for accuracy testing.

In this paper, GPR models are trained with the same experimental data used by Silva [3], with k-fold cross-validation and tuning optimally three hyperparameters. The objective is to create a single model that can predict the Resilient Modulus of soil-polymer composites while inputting also their polymer dosage and curing time,

aiming to be as competitive as standard physics-based constitutive models. The background of GPR technique is explained based on the literature of Rasmussen and Williams [6] and Bishop [7], and the study of MacKay [8,9] and Williams [10]. A comparison between GPR models and Silva's [3] physics-based models is done by their prediction accuracy, measuring values of Mean Squared Error (MSE) and Mean Squared Error of normalized response variable (MSEn).

2 Background

In this section, it is presented how ML-based and physics-based constitutive models are formulated to predict the response variable, since each technique presents its own set of equations and predictor variables.

2.1 Physics-based constitutive models for Resilient Modulus

Silva [3] built nine separate models so that every combination of SP and curing time was accomplished, with exception of SP 0%, that had a single model for all curing times. These models were created with the software LABFit Curve Fitting – V.7.2.50 and are based on the traditional stress state of the material, defining only confining stress and deviator stress, as described in the equation

$$RM = k_1 \sigma_3^{k_2} \sigma_d^{k_3}, \tag{1}$$

where k_1, k_2, k_3 are constants of the regression model, σ_3 is the confining stress and σ_d is the deviator stress.

2.2 Gaussian Process Regression

GPR is considered a probabilistic nonparametric technique that relies on kernel functions to train its models. In the Gaussian Process (GP) viewpoint, it defines a prior probability distribution over functions directly, what it difficult to work with over the infinite space of functions. But, since the training set is finite, only the values of the functions at the discrete set of input variables x_n, corresponding to training set and test set data points, needs to be considered, so, in practice, it can be considered a finite space. The technique addresses the question of predicting the value of a response variable y_{new}, given the new input vector x_{new}, and the set of training data, which includes predictor variables and observed response variables.

In order to apply GP models to the regression problem, it needs to be taken into account the noise on the observed target values, which is given by

$$t_n = y_n + \epsilon_n, \tag{2}$$

where $y_n = y(x_n)$, and ϵ_n is a random noise variable whose value is chosen independently for each observation n. Because the noise is independent for each data point, the joint distribution of the target value $t = (t_1, \dots, t_N)^T$ conditioned on the values of $y = (y_1, \dots, y_N)^T$ is given by an isotropic Gaussian of the form

$$p(t|y) = \mathcal{N}(t|y, \beta^{-1} I_N), \tag{3}$$

where I_N denotes the N x N unit matrix. The marginal distribution of t, $p(t)$, is given by

$$p(t) = \int p(t|y)p(y)dy = \mathcal{N}(t|0, C) \tag{4}$$

where the covariance matrix C has elements

$$C(x_n, x_m) = k(x_n, x_m) + \beta^{-1}\delta_{nm}. \tag{5}$$

One of the kernel functions commonly used for GPR models is given by the exponential of a quadratic form, with the addition of constant and linear terms to give

$$k(x_n, x_m) = \theta_0 \, exp\left\{-\frac{\theta_1}{2} \, ||x_n - x_m||^2\right\} + \theta_2 + \theta_3 x_n^T x_m. \tag{6}$$

Considering that the goal is to predict the target variable t_{N+1} for a new input vector x_{N+1}, the predictive distribution $p(t_{N+1}|t_N)$ needs to be evaluated. The joint distribution over $t_1, \dots, t_{N+1}$will be given by $p(t_{N+1}) = \mathcal{N}(t_{N+1}|0, C_{N+1})$, where C_{N+1} is an $(N+1)$ x $(N+1)$ covariance matrix with elements given by Eq. (6). To find

the conditional Gaussian distribution, the covariance matrix is partitioned as

$$C_{N+1} = \begin{pmatrix} C_N & k \\ k^T & c \end{pmatrix}, \tag{7}$$

where C_N is an N x N covariance matrix given by Eq. (6) for $n, m = 1, \dots, N$, the vector k has elements $k(x_n, x_{N+1})$ for $n = 1, \dots, N$, and the scalar $c = k(x_{N+1}, x_{N+1}) + \beta^{-1}$.

3 Methodology

The data set was obtained experimentally in laboratory and consists of a 180x5 matrix, which columns correspond to confining stress (σ_3), deviator stress (σ_D), polymer dosage in the composite (SP), curing time and Resilient Modulus (RM). Confining and deviator stresses form 15 different pairs of values that are repeated twelve times, four for each curing time and three for each SP, totalizing 180 rows. While σ_3 and σ_D range, respectively, from 0.0207 to 0.1379 MPa and 0.0207 to 0.2758 MPa, polymer dosage assumes values of 0%, 2.5% and 5%, and curing time 7, 15, 30 or 45 days.

3.1 Experimental data

The experimental data was obtained through tests that followed AASHTO T 307-99 [11] with a Triaxial Cyclic equipment (Fig. 1), developed by the company ELE International, located at University of Brasilia in the Infrastructure Laboratory. The equipment uses compressed air for load application and consists of a removable triaxial chamber with a pneumatic loading device.

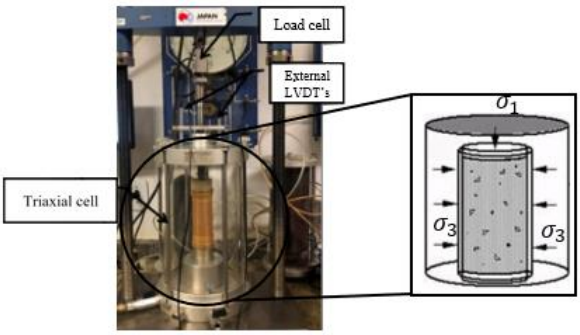

Figure 1. Triaxial Cyclic equipment

3.2 Machine Learning

The cross-validation method known as k-fold consists of first dividing the data set into a number k of groups, after that, one of the groups is selected as the test set and the rest is as the training set, then this process is repeated k times always selecting a different group as the test set (Fig. 2).

MSEn is calculated in each repetition of the described process, so that mean and Standard Deviation (SD) of the k values of MSEn can be determined. This whole process helps avoiding overfitting and is key to hyperparameter optimization, since it can measure how well the parameters do at predicting data.

Every time the data set is partitioned, its proportion is maintained to avoid biased partition, for example, if 20% of the set is to be isolated and used as validation set, it is selected 3 of each 15 specimens' group, instead of picking up randomly 36 out of the 180 total.

The predictors of the models are σ_3, σ_D, SP and curing time, whereas the response variable is the Resilient Modulus. Different models were created just to compare them with each other and select only the one that had the most accurate predictions.

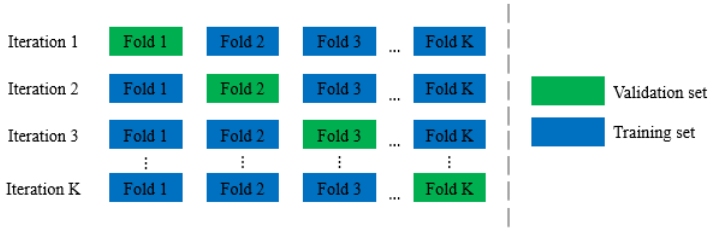

Figure 2. K-fold cross-validation in Machine Learning.

The tool utilized to train, validate and test the models was the software called MATLAB. Hyperparameters were optimized with assist of built-in functions of the software, using Bayesian optimization with *expected-improvement-plus* family of acquisition functions. The parameters optimized were *KernelFunction*, *BasisFunction* and *Standardize*. The best value of each parameter is selected for each one of the k = 2, 3, 4, 5 and 6 folds, then the one that appeared the most is selected as the overall best value of that parameter and is used in the final model.

4 Results

Predictions of the experimental data were plotted to show the varying results depending on SP, curing time and training or validation sets. The graphics are displayed in Fig. 3, where the horizontal axis presents the expected value and the vertical one presents the predicted value of that same point. So, if a point is above the line, it predicted a higher value than what it should be, if it is under the line, the prediction was of lower-than-expected value.

MSE and MSEn were calculated for each SP and curing time value (Tab.2) in order to quantify the prediction accuracy of the models and to compare them with each other. Comparing the results of both techniques, its noticeable that MSE and MSEn are almost always lower on the ML models, a fact that indicates a high goodness-of-fit.

Table 2. Comparison between MSE and MSEn of the models

Accuracy indicator	SP	Curing time	Physics-based	GPR
MSE	0% (Soil)	All	243.72	120.96
	2.5%	7 days	987.08	805.91
		15 days	374.08	230.85
		30 days	2,662.66	2,267.55
		45 days	161.62	118.06
	5%	7 days	3,035.66	1,742.86
		15 days	3,105.54	2,770.83
		30 days	3,714.19	2,419.63
		45 days	185.24	635.49
MSEn	0% (Soil)	All	0.0008	0.0004
	2.5%	7 days	0.0045	0.0025
		15 days	0.0013	0.0007
		30 days	0.0082	0.0069
		45 days	0.0006	0.0004
	5%	7 days	0.0135	0.0053
		15 days	0.0119	0.0085
		30 days	0.0118	0.0074
		45 days	0.0010	0.0019

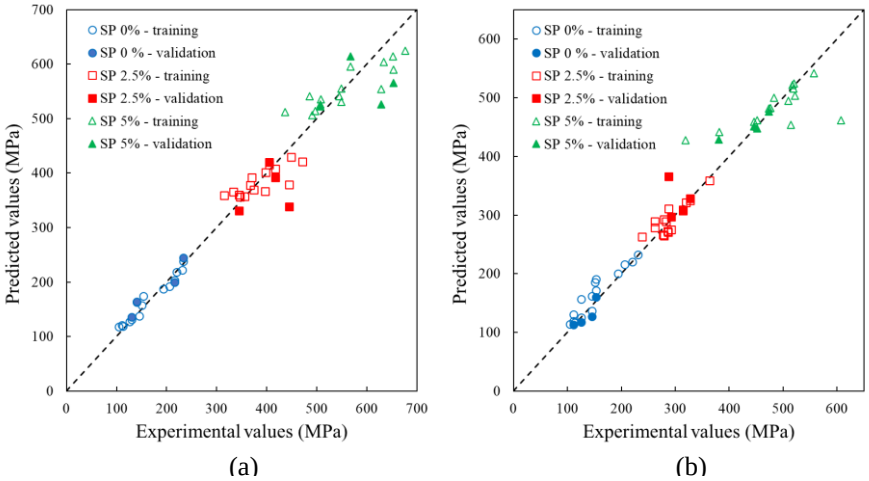

Figure 3. GPR model predictions by curing times: a) 7 days; b) 15 days

5 Conclusion

In the present work, comparisons were done between proposed ML models and physics-based models respective predictions. In the prediction accuracy comparison, it is showed that, in almost every case, ML models present better MSE and MSEn. In conclusion, proposed ML-based models are as competitive as stress-state-based ones, presenting similar predictions, and show better predictor accuracy indicators such as MSE and MSEn.

Acknowledgements. The authors gratefully acknowledge the financial support granted by DEG/DAIA (in Portuguese, *Decanato de Ensino de Graduação/Diretoria de Acompanhamento e Integração Acadêmica)* from University of Brasilia and general support of the brazilian research agency CNPq (in Portuguese, *Conselho Nacional de Desenvolvimento Científico e Tecnológico*).

Authorship statement. The authors hereby confirm that they are the sole liable persons responsible for the authorship of this work, and that all material that has been herein included as part of the present paper is either the property (and authorship) of the authors, or has the permission of the owners to be included here.

References

[1] R. A. F. Benvindo. Análise comparativa dos instrumentos de regulamentação das políticas de proteção ambiental e de promoção do ecoturismo: caso do Parque Estadual do Jalapão – TO. Master Thesis, School of Architecture and Urbanism, University of Brasilia, 2009.
[2] B. M. Carneiro and L. F. M. Adorno. A implantação de vias e o manejo de Áreas Protegidas na região do Jalapão: a rodovia TO-110 e seu entorno. *Caminhos de Geografia*, vol. 11, n. 33, pp. 120-137, 2010.
[3] N. A. B. S. Silva. Desempenho de um compósito solo-polímero para aplicabilidade em obras geotécnicas e de pavimentação. Master Thesis, Department of Civil and Environmental Engineering, University of Brasília, 2020.
[4] M. Sharifzadeh, A. Sikinioti-Lock and N. Shah. Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and Gaussian Process Regression. *Renewable and Sustainable Energy Reviews*, vol. 108, pp. 513–538, 2019.
[5] J. Zhao, H. Guo, M. Han, H. Tang, X. Li. Gaussian Process Regression for Prediction of Sulfate Content in Lakes of China. *Journal of Engineering and Technological Sciences*, vol. 51, n. 2, pp. 198-215, 2019.
[6] C. E. Rasmussen e C. K. I. Williams. *Gaussian Processes for Machine Learning.* Massachusetts Institute of Technology. MIT Press, 2006.
[7] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
[8] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.
[9] D. J. C. MacKay. Introduction to Gaussian Processes. In: C. M. Bishop (ed.), *Neural Networks and Machine Learning* (NATO ASI Series, Kluwer), pp. 133–166, 1998.
[10] C. K. I. Williams. Regression with Gaussian Processes, 1997.
[11] AASHTO (American Association of State and Highway Transportation Officials, Standard Method of Test for Determining the Resilient Modulus of Soils and Aggregate Materials, 2007.