REVISTAMATĒRIA V.28 N.4
ISSN 1517-7076 Articles e20230188, 2023

A statistical methodology for reliable evaluation of calibrated dynamic modulus regression models

Isadora Guimarães dos Santos[1], Rogério Pinto Espíndola[1], Francisco Thiago Sacramento Aragão[1],
Luis Alberto Herrmann Nascimento[2]

[1]Universidade Federal do Rio de Janeiro, Departamento de Engenharia Civil, Centro de Tecnologia. Av. Athos da Silveira Ramos, 149, 21941-909, Cidade Universitária, Rio de Janeiro RJ, Brasil.

[2]PETROBRAS, Centro de Pesquisas, Desenvolvimento e Inovação Leopoldo Américo Miguez de Mello. Av. Horácio Macedo, 950, 21941-915, Cidade Universitária, Rio de Janeiro, RJ, Brasil.

e-mail: isadora@coc.ufrj.br, rogerio@coppe.ufrj.br, fthiago@coc.ufrj.br, luisnascimento@petrobras.com.br

ABSTRACT

Predictive models are useful tools for reliable estimations of key mechanical properties when properly calibrated. Several research efforts have compared calibrated models in the literature, but the sampling techniques adopted in the model calibration, selection, and evaluation were not the focus of these studies. This work reviews different sampling techniques and employs hold-out and repeated k-fold cross-validation (CV) to evaluate three empirical dynamic modulus equations calibrated using a database containing 1,806 records from 65 asphalt mixtures. The results indicate that hold-out can induce unrealistic conclusions about the estimations, while repeated k-fold CV is a reliable methodology.

Keywords: Asphalt mixtures; dynamic modulus; empirical predictive models; resampling methods; repeated cross-validation.

1. INTRODUCTION

The use of predictive models to estimate asphalt mixture properties is an interesting alternative to laboratory tests, especially in the initial stages of pavement design, for example, the material selection process. Such models allow the preliminary and expedited evaluation of different combinations of mixture constituents. Furthermore, empirical models can be used to estimate design parameters, such as the dynamic modulus of asphalt mixtures ($|E^*|$), an important input in the Mechanistic-Empirical Pavement Design Guide (MEPDG) [1].

Several authors proposed empirical models to predict $|E^*|$, expressed as equations that relate characteristics of the component and mix design parameters to the stiffness properties of the mixtures, including WITCZAK and FONSECA [2], CHRISTENSEN JUNIOR *et al.* [3], BARI and WITCZAK [4], MATEOS and SOARES [5], YANG and YOU [6], and SAKHAEIFAR *et al.* [7].

Recently, with the accumulation of experimental data and the use of more powerful computers, predictive models developed with different machine learning techniques have emerged. These include artificial neural network (ANN) models proposed by FAR *et al.* [8], CEYLAN *et al.* [9], RAHMAN and TAREFDER [10], and BARUGAHARE *et al.* [11], random forests by DANESHVAR and BEHNOOD [12], and a gene expression programming (GEP) model by LIU *et al.* [13].

Empirical models are generally developed based on a statistical analysis of experimental data using regression techniques. The goal is to obtain a model as simple as possible, ensuring very good predictions regardless of the dataset considered (representation and generalization skills). However, since a model is an approximate representation of a real system, its predictions are uncertain and the adjustment of its parameters can affect the bias and variance of the results. Therefore, additional adjustments to the model parameters are often necessary, especially when the material characteristics differ from those used in the original dataset.

Although several $|E^*|$ predictive models have been published in the literature, the sampling techniques adopted in their evaluations were not the focus of these studies. A common approach is to randomly divide a dataset into two non-overlapping subsets, i.e., training and testing datasets. During training, the model parameters are fitted to the available data through regressions. Later, in the testing step, the model is evaluated based

Corresponding Author: Isadora Guimarães Dos Santos Received on 27/06/2023 Accepted on 15/08/2023

DOI: https://doi.org/10.1590/1517-7076-RMAT-2023-0188

on its learned parameters. The downside of this procedure, called hold-out [14], is that the model performance is dependent on the dataset splitting criteria and the sampled subsets may not statistically represent the population.

In search of more reliable evaluations of the performance of predictive models, this work presents sampling techniques usually adopted in machine learning studies for the calibration of $|E^*|$ predictive equations. In the analysis, the repeated cross-validation (CV) is adopted for the calibration of CHRISTENSEN JUNIOR *et al.* [3], BARI and WITCZAK [4], and SAKHAEIFAR *et al.* [7] equations, considering a database of 65 asphalt mixtures used in Brazil. The analyses indicate that even robust models present variations in the performance when evaluated with a hold-out scheme. Therefore, the analysis of calibrated models through repeated cross-validation is a more reliable methodology for the evaluation of calibrated regression models, given that it is based on the arithmetic mean of values instead of on a single value as the performance estimator. In addition, it allows the determination of a confidence interval (CI), which allows a better estimate of the performance of a model according to different sampled data. The following section presents basic concepts about sampling and some of the techniques available in the literature for evaluating predictive models.

2. RESAMPLING TECHNIQUES

A sample is a subset of elements from a population. If population data is widely available or it is feasible to acquire multiple datasets, both training and test subsets can be large and diverse enough to be representative. However, such situations are rare in science and one way to mitigate this problem is to perform sampling from the known dataset, treat it as a proxy for the population data, and resample it repeatedly [15]. Thus, resampling is the process of sampling available data several times to make more reliable inferences about the behavior of a statistical estimator.

Regarding predictive models, resampling techniques can be applied for different purposes, such as selection and evaluation tasks. The latter consists of estimating the predictive performance of a model on unseen data, which can be used as a criterion for model selection when some are considered [16]. Predictive performance estimates can also be used in the adjustments of model parameters, as in the calibration of asphalt mixture empirical models to unseen data from other sources, and in the search for the best modeling parameters (hyperparameters) to reduce bias and variance of the results.

Several resampling techniques have been proposed in the literature to assess the performance of predictive models. Some of the main ones available in the literature are described below.

2.1. Hold-out

Hold-out is a simple sampling technique that relies on randomly dividing the available data into two mutually exclusive subsets: training data and testing data, only once. Often, the training set contains about 70% to 90% of the available data, while the testing set contains the remaining 30% to 10% [17, 18], respectively. The testing data are held out for evaluation purposes and they are not used for training [14].

The model performance is evaluated based on the testing sample predictions. However, considering an asphalt mixture dataset, these testing data can represent three different scenarios: a) The best-case condition, wherein the testing set comprises mixtures that closely adhere to the calibrated equation derived during training; b) The worst-case condition, where the test mixtures significantly differ from those employed to train the model; and c) an intermediate scenario, which does not fall into either the best or worst case. Thus, hold-out is sensible to the splitting criteria adopted and when a single sampling is performed for the training and testing datasets, there is no guarantee that the value obtained for the determination coefficient (R^2) of a model is a strong performance estimate.

Since part of the data is not used in the calibration, testing samples may have a different distribution than the full dataset, an undesirable situation for model performance reliability. Furthermore, data on the testing set may be valuable for training and the prediction performance may be affected if they are held out, again leading to skewed results [19]. In other words, model evaluations can differ significantly depending on the selection of elements in each subset.

2.2. Bootstrap

Bootstrap is a family of techniques that perform successive random sampling with replacement from observations. The statistical models are fitted to the sampled data, while the observations that were left out compose a testing set. Details of bootstrap variations are described in EFRON and TIBSHIRANI [20].

The advantage of sampling with replacement is that the training subset can be as large as the original dataset. However, it is important to highlight that samples must be representative, independently, and identically distributed. As the sampled data with bootstrap may have different distributions than those of the available data, this technique is not recommended in cases of small sample sizes and when there is an interest in estimating extreme values of the distributions. Relying on asymptotic results in small samples or treating dependent data as if they were independent can underestimate sampling variation and make the results appear better than they are [21].

2.3. Cross-validation (CV) - general concept

It is a sampling method without replacement that allows the assessment of the generalization ability of predictive models, being an interesting strategy for actions that try to avoid an overfitted modeling to a particular set of training data [16, 22, 23]. The CV method assumes that data are identically distributed and that the training and testing subsets are independent. There are different variations of CV sampling, as detailed below.

2.3.1. Leave-one-out cross-validation (LOOCV)

The method was introduced by STONE [24] and consists of splitting the data into training and testing subsets so that in each iteration nearly all the data, except for a single observation, are used for training and the model is tested on that single observation. This process is repeated until all dataset elements have been used in the test subset.

An advantage of LOOCV is that it has a small bias and does not tend to overestimate the testing error rate, as learning is repeated using training sets with almost all data ($n - 1$). Due to the absence of randomness in particular training/testing sets, different LOOCV runnings produce the same results. On the other hand, a disadvantage of this approach is that it can be computationally expensive for large datasets. This technique requires n models to be learned to evaluate a calibration process [16].

2.3.2. k-fold cross-validation (k-fold CV)

It was introduced by GEISSER [25] as an alternative to the computationally expensive LOOCV for non-small datasets [26]. Figure 1 illustrates the k-fold CV method, which involves randomly splitting the sample set into k nearly equal folds. Subsequently, k iterations of training and testing are performed such that within each iteration a different fold is reserved for testing and the ($k - 1$) remaining folds are used as the training subset. This allows the use of all records in the training and testing subsets, but not simultaneously. As a result, each iteration provides a scenario for learning and evaluating the model, where some may be more pessimistic and others more optimistic. As with the other resampling methods, the final model is built using the entire dataset, but its expected performance is calculated as the average of the evaluations obtained in each iteration, which provides a more robust evaluation than hold-out, since k-fold considers several evaluation scenarios. Selecting an appropriate value for k is important to obtain a reliable estimation of the model's performance. Usually, k values are between five and ten [27], and $k = 10$ was here adopted based on a study by KOHAVI [17], where different values were evaluated and the results indicated that most estimates were reasonably good at ten folds. Smaller values of k may result in a more biased estimation, while larger values of k may require longer processing time. Therefore, $k = 10$ is a common choice for k-fold cross-validation due to its balance between bias and variance, as well as its computational feasibility.

According to BREIMAN [28], as each training subset is used ($n - 1$) times during the learning, its iterations are not independent of each other, implying a variance of performances that may be large, but not

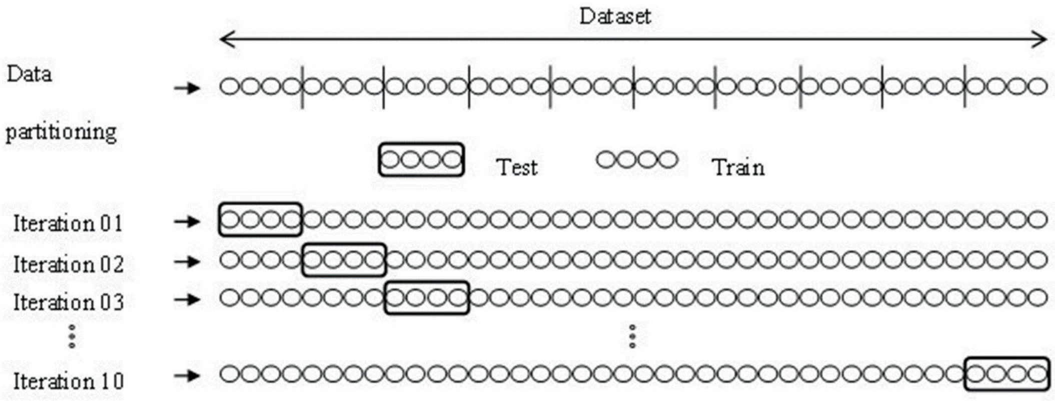

Figure 1: k-fold CV (k = 10).

as large as those observed in applying the hold-out sampling multiple times. KRSTAJIC *et al*. [29] performed 10-fold cross-validations 50 times for several datasets and compared the distributions of optimal cross-validatory parameters for each dataset, proving that the model selected by a single cross-validation may have high variance, which points out the need for repeated cross-validation.

2.3.3. Repeated k-fold cross-validation

To deal with the variability of cross-validation results, some authors recommend the repeated CV [16, 17, 29, 30], an improved method that runs k-fold CV multiple times and shuffles the data before each repetition. The expected performance is calculated as the average of the evaluations obtained in each CV. Although the computational cost of the calibration process is increased, as the mean is a better estimate of a variable value than a single value, this method is an efficient tool based on a robust sampling strategy that can be adopted for the evaluation of asphalt mixture predictive models.

3. MODEL PERFORMANCE ASSESSMENT INDEXES

The performance evaluation of the models is accomplished by observing the determination coefficient, R^2, calculated as shown in Equation 1.

$$R^2(SS_{res}, SS_{tot}) = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{1}$$

where: $\bar{y}$ is the sample mean; SS_{res} is the quadratic sum of the residuals, $y_i - \hat{y}_i$; and SS_{tot} is the sum of the squared deviations, $y_i - \bar{y}$.

The ratio between the standard error of the estimated modulus values and the deviation of the measured values, S_e/S_y, is calculated from the results of Equations 2 and 3.

$$S_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(n-1)}} \tag{2}$$

$$S_y = \sqrt{\frac{\sum(y_i - \bar{y}_i)^2}{(n-1)}} \tag{3}$$

For qualitative assessments of model calibrations, the criteria proposed by WITCZAK *et al.* [32], as shown in Table 1, were adopted in this paper.

4. DYNAMIC MODULUS EMPIRICAL MODELS

In this paper, repeated k-fold cross-validation was used in the calibration of three $|E^*|$ predictive equations, as described below.

4.1. Christensen Junior *et al.* (2003)

The model proposed by CHRISTENSEN JUNIOR *et al.* [3] was developed from 206 data points of 18 asphalt mixtures used in the United States. It is an equation based on the law of mixtures that considers three independent variables: mixture voids in the mineral aggregate (*VMA*), %, and voids filled with asphalt (*VFA*), %, and

Table 1: Criteria for qualitative model evaluation.

PREDICTIVE POTENTIAL	R^2	S_e/S_y
Excellent	> 0.90	< 0.35
Good	0.70-0.89	0.36-0.55
Fair	0.40-0.69	0.56-0.75
Poor	0.20-0.39	0.76-0.90
Very poor	< 0.19	> 0.90

binder dynamic shear modulus ($|G_b^*|$), psi. From this information, the contact volume between aggregate particles (P_c) and the $|E^*|$ are calculated, as shown in Equations 4 and 5.

$$\left|E^*\right| = P_c\left[4{,}200{,}000\left(1-\frac{VMA}{100}\right)+3\left|G_b^*\right|\left(\frac{VFA\cdot VMA}{10{,}000}\right)\right]+(1-P_c)\left[\frac{1-\dfrac{VMA}{100}}{4{,}200{,}000}+\frac{VMA}{3\cdot VFA\cdot\left|G_b^*\right|}\right]^{-1} \tag{4}$$

$$P_c = \frac{\left(20+\dfrac{VFA\cdot 3\left|G_b^*\right|}{VMA}\right)^{0.58}}{650+\left(\dfrac{VFA\cdot 3\left|G_b^*\right|}{VMA}\right)^{0.58}} \tag{5}$$

When the model was evaluated with the original dataset, it obtained a R^2 of 0.98, in logarithmic scale, using the 90–10 hold-out scheme. However, when BARI and WITCZAK [4] applied the equations to the expanded dataset with 7,400 $|E^*|$ values, this model presented R^2 of 0.61 in logarithmic scale and R^2 of 0.23 in arithmetic scale.

4.2. Bari and Witczak (2006)

The BARI and WITCZAK [4] model was developed considering a dataset containing 7,400 $|E^*|$ values from 346 mixtures used in the United States. The goodness of fit was evaluated without a resampling scheme, i.e., all available data was used to assess the model. The results are shown in two ways: in logarithmic scale, the model presented R^2 of 0.90 and S_e/S_v of 0.32, and in arithmetic scale, R^2 was 0.80 and S_e/S_v was 0.45. The model, presented in Equation 6 in its logarithmic version, includes as variables the binder $|G_b^*|$, mixture volumetric properties, and aggregate gradation information.

$$\begin{aligned}\log_{10}\left|E^*\right| &= -0.349+0.754\left|G_b^*\right|^{-0.0052}\\ &\times\left(6.65-0.032\rho_{200}+0.0027\rho_{200}^2+0.011\rho_4^2+0.006\rho_{38}-0.00014\rho_{38}^2-0.08V_v-1.06\left(\frac{V_{be}}{V_v+V_{be}}\right)\right)\\ &+\frac{2.56+0.03V_v+0.71\left(\dfrac{V_{be}}{V_v+V_{be}}\right)+0.012\rho_{38}-0.0001\rho_{38}^2-0.01\rho_{34}}{1+e^{\left(-0.7814-0.5785\log\left|G_b^*\right|+0.8834\log\delta_b\right)}}\end{aligned} \tag{6}$$

where: δ_b is the binder phase angle, degree; ρ_{200} are the particles (by weight of the total particles) passing through sieve No. 200, %; ρ_4 are the cumulative particles (by weight of the total particles) retained on sieve No. 4, %; ρ_{38} are the cumulative particles (by weight of the total particles) retained on sieve 3/8”, %; ρ_{34} are the cumulative particles (by weight of the total particles) retained on sieve 3/4”, %; V_v are the air voids (by volume of the mix), %; and V_{be} is the effective binder content (by volume of the mix), %.

4.3. Sakhaeifar *et al.* (2015)

The SAKHAEIFAR *et al.* [7] model was developed on a database of 20,209 data points from 1008 mixtures used in the United States gathered from several sources. Equation 7 considers the viscoelastic material behavior, as well as different physical and mechanical properties of the mixtures. The final model was calibrated using hold-out with 90% of the data for training and the remaining for testing. Considering the logarithmic scale, the model presented R^2 of 0.98 and of 0.14 for the training dataset, and R^2 of 0.99 and S_e/S_v of 0.13 for the testing dataset. In arithmetic scale, the authors obtained R^2 of 0.95 and S_e/S_v of 0.22 for the training dataset, and R^2 of 0.93 and S_e/S_y of 0.27 for the testing dataset.

$$\log_{10}\left|E^*\right| = 6.4197 - 0.00014\rho_{34}^2 - 0.00547\rho_{38} - 0.11786\rho_{200} - 0.05528V_v - 0.16266V_{be} + 0.00487V_{be}^2 + \frac{0.57677 + 0.00713\rho_{38} + 0.16167\rho_{200} - 0.0052\rho_{200}^2 + 0.01889V_v + 0.16031V_{be} - 0.00592V_{be}^2}{1 + e^{1.8645 - 0.95991\log\left|G^*\right|}} \tag{7}$$

5. BRAZILIAN DATASET

The dataset of this research includes information from 65 asphalt mixtures used in Brazil. The same dataset was utilized in the study conducted by SANTOS *et al.* [33], with the objective of developing an artificial neural network (ANN) model for predicting the dynamic modulus of asphalt mixtures. It is a comprehensive database that considers aggregates from five Brazilian states, mixtures with polymer-modified and conventional (unmodified) asphalts, in a total of 36 different binders. The dataset consisted of 1,806 dynamic modulus master curve values, information about binders, gradation, and volumetric properties of those mixtures. Table 2 presents a statistical description of the parameters.

As the coefficient of variation is a statistical measure of the relative dispersion of values in a data series around their mean, the order of magnitude of a variable does not change its interpretation. The coefficient of variation for gradation parameters varied more significantly for fractions retained on sieves 3/4" and 3/8", while the fractions retained on sieve No. 4 presented the smallest variation. In general, small variations were observed for the volumetric parameters, except for V_{be}, which presented a coefficient of variation of 20.2%. Mixtures with air voids ranging between 3.9% and 7.1% and ρ_{200} up to 8.3% were analyzed in this research.

Table 2: Data description.

PARAMETER	SYMBOL	MEAN	STD. DEV.	COEF. OF VARIATION	MIN.	MAX.
Binder dynamic shear modulus at 20°C (MPa)	$\|G_b^*\|$	90.184	153.843	170.6%	1.39E-4	1.12E+3
Binder phase angle (°)	δ_b	52.660	20.541	39.0%	8.158	89.202
Aggregates passing sieve No. 200	ρ_{200}	0.047	0.010	21.3%	0.030	0.083
Cumulative retained material on sieve No. 4	ρ_4	0.488	0.070	14.3%	0.300	0.601
Cumulative retained material on sieve 3/8"	ρ_{38}	0.246	0.084	34.1%	0.050	0.342
Cumulative retained material on sieve 3/4"	ρ_{34}	0.020	0.028	140.0%	0.000	0.072
Air voids	V_v	0.054	0.008	14.8%	0.039	0.071
Effective binder volume	V_{be}	0.094	0.019	20.2%	0.038	0.130
Voids in the mineral aggregate	*VMA*	0.149	0.015	10.1%	0.107	0.175
Voids filled with asphalt	*VFA*	0.628	0.083	13.2%	0.352	0.764
Dynamic modulus (MPa)	$\|E^*\|$	13,328	13,331	100.0%	62	70,873

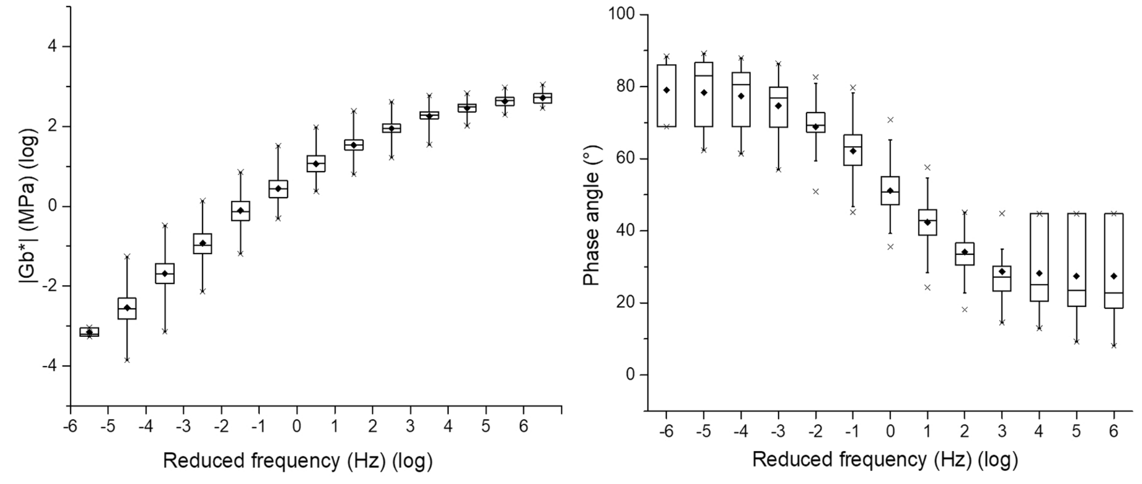

Figure 2: Variation of stiffness and phase angle for the binders evaluated in the study.

The plots in Figure 2 illustrate the variation in stiffness and phase angle values of the binder master curves at 20°C. The amplitudes of $|G^*|$ values are reduced for higher frequencies, with some tendency of stabilization near 2 GPa at the upper limit of $|G^*|$, regardless of the binder. The phase angle values, however, presented more dispersed values for extreme frequencies, which was expected, given the diversity of binders in the database.

6. MODEL CALIBRATION PROCESS USING REPEATED k-FOLD CV

In this work, CHRISTENSEN JUNIOR *et al.* [3], BARI and WITCZAK [4], and SAKHAEIFAR *et al.* [7] predictive equations were calibrated using a code written in Python language and run in a conventional computer. The purpose of the calibration was to adjust the coefficients of each equation to the Brazilian Dataset by the minimization of an objective function. In this work, the root mean-square error (*RMSE*) was adopted as objective function.

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \tag{8}$$

where: *RMSE* is the root mean-square error; y_i is the experimental value of $|E^*|$; $\hat{y}_i$ is the predicted $|E^*|$ value; and n is the sample size.

The L-BFGS-B (Limited-memory, Broyden-Fletcher-Goldfarb-Shanno, Bound-constrained) optimization algorithm [31] was adopted to perform the calibration, which is a family of quasi-Newton methods that implement the BFGS optimization algorithm in limited computational memory environments for bound constrained minimization. The L-BFGS-B is a variation of the BFGS algorithm that utilizes a limited approximation of the inverse Hessian matrix to guide the search for the optimal solution. This approximation is constructed based on gradient information from previous iterations, enabling the algorithm to efficiently navigate towards the best solution through the variable (coefficients) space. It provides an efficient approach to optimization by avoiding the need to calculate and store the complete Hessian matrix.

To obtain statistically robust assessments for R^2, a 10-fold CV was repeated 30 times with random splits and confidence intervals for R^2 and RMSE were calculated. In addition, S_e/S_v was also calculated based on Equations 2 and 3 to perform the qualitative model evaluation.

At the end of this process, a final model equation was produced using the entire dataset and the expected performance of this model was estimated by the mean R^2 through the repeated CV. A pseudo-code adopted in the model calibration process with repeated CV is presented below.

6.1. Pseudo code for the k-fold cross-validation

```
State the model equation and the value of $k$
Initialize coefficients with values from the original model
Load dataset
For repetition from 1 to 30
    Shuffle dataset
    Randomly split dataset into $k$ folds
    For fold from 1 to $k$
        Test_set(i) = ith fold
        Train_set(i) = dataset without ith fold
        Calibration = minimize MSE over Train_set(i) with L-BFGS-B optimizer
        Calculate $|E^*|$ for Test_set(i)
        Evaluate model performance on Test_set(i)
    Calculate average performance over $k$-folds
Calculate average performance after repetitions and obtain confidence intervals
```

Finally, to verify how often the results produced with the hold-out sampling may represent unlikely scenarios (outside the CI of the results obtained with repeated cross-validation), the three models were calibrated with this same dataset, using the 90–10 splitting ratio adopted in each iteration of the 10-fold CV.

7. RESULTS AND DISCUSSION

7.1. Performance of original models

CHRISTENSEN JUNIOR *et al.* [3], BARI and WITCZAK [4], and SAKHAEIFAR *et al.* [7] original equations were used to predict $|E^*|$ experimental values of all 65 asphalt mixtures evaluated in this work. Table 3 presents R^2, *RMSE*, and the predictive potential of each model in arithmetic scale, which was between fair and good for the analyzed dataset.

These performances are probably related to the fact that all models were designed for mixtures used in the United States, which were produced with materials different than those adopted in the composition of Brazilian mixtures. Among the models evaluated, the BARI and WITCZAK [4] equation presented the worst performance, while SAKHAEIFAR *et al.* [7] was the original model that best fitted the Brazilian mixtures.

7.2. Performance of calibrated models

The calibrated equations obtained using the L-BFGS-B optimization algorithm with the entire dataset are presented below.

Calibrated CHRISTENSEN JUNIOR *et al.* [3]:

$$\left|E^*\right| = P_c\left[4,200,000\left(1-\frac{VMA}{100}\right)+3\left|G_b^*\right|\left(\frac{VFA\cdot VMA}{10,000}\right)\right]+\left(1-P_c\right)\left[\frac{1-\dfrac{VMA}{100}}{4,200,000}+\frac{VMA}{3\cdot VFA\cdot\left|G_b^*\right|}\right]^{-1} \tag{9}$$

$$P_c = \frac{\left(20+\dfrac{VFA\cdot 3\left|G_b^*\right|}{VMA}\right)^{(-0.0801)}}{650+\left(\dfrac{VFA\cdot 3\left|G_b^*\right|}{VMA}\right)^{(-0.0801)}} \tag{10}$$

Calibrated BARI and WITCZAK [4]:

$$\begin{aligned}\left|E^*\right| = EXP\Big(&-0.045+3.062\left|G_b^*\right|^{-0.0510}\\ &\times\left(7.449+2.588\rho_{200}-0.133\rho_{200}^2+4.656\rho_4-0.0496\rho_4^2-1.123\rho_{38}-0.186\rho_{38}^2+3.010V_v-0.352\left(\frac{v_{be}}{V_v+V_{be}}\right)\right)\\ &+\frac{+\ 2.874+0.045V_v+0.945\left(\dfrac{v_{be}}{V_v+V_{be}}\right)+0.022\rho_{38}+1.275\rho_{38}^2-0.0035\rho_{34}}{1+e^{\left(-0.988-0.306\log\left|G_b^*\right|+0.493\log\delta_b\right)}}\Bigg)\end{aligned} \tag{11}$$

Calibrated SAKHAEIFAR *et al.* [7]:

$$\begin{aligned}\left|E^*\right| = EXP\Big(&6.5257-0.0009\rho_{34}^2+0.00094\rho_{38}+0.04366\rho_{200}-0.3126V_v-0.0978V_{be}+0.00096V_{be}^2\\ &+\frac{0.7195+0.00188\rho_{38}-0.2253\rho_{200}+0.0164\rho_{200}^2+0.3872V_v+0.0439V_{be}+0.00049V_{be}^2}{1+e^{1.8805-0.91748\log\left|G^*\right|}}\Big)\end{aligned} \tag{12}$$

Table 4 shows the results obtained from the calibration process, including each model with their respective R^2 and *RMSE* in arithmetic scale, according to the sampling techniques used to evaluate the models during calibration, and the predictive potential of the calibrated models. 10-fold CV values were presented with four decimals to show the CI with 95% of confidence level. The CI for a population sampled mean was calculated. The statistical assumptions for this method include that the observations in the data should be independent of

Table 3: Performance of the original predictive models.

MODEL	R^2	*RMSE*	PREDICTIVE POTENTIAL
BARI and WITCZAK (2006)	0.532	9084	Fair
CHRISTENSEN JUNIOR *et al.* (2003)	0.605	8347	Fair
SAKHAEIFAR *et al.* (2015)	0.861	4942	Good

Table 4: Performance of the models after the calibrations.

MODEL	HOLD-OUT (90–10 SCHEME)		REPEATED 10-FOLD CROSS-VALIDATION		
	R^2 MIN	R^2 MAX	R^2	*RMSE*	PREDICTIVE POTENTIAL
BARI and WITCZAK (2006)	0.90	0.96	0.9344 ± 0.0010	0.1829 ± 0.0013	Excellent
CHRISTENSEN JUNIOR *et al.* (2003)	0.60	0.85	0.7256 ± 0.0011	6954.3688 ± 13.7838	Good
SAKHAEIFAR *et al.* (2015)	0.94	0.97	0.9603 ± 0.0001	0.1430 ± 0.0001	Excellent

each other, the population should have a sufficiently large sample size or the population itself should follow a normal distribution, and the sampling process should be random.

All models were calibrated in less than one minute and provided at least good predictions. Considering the use of a regular computer and numerous iterations involving the repetition of the 10-fold calibration process 30 times, achieving the calibration of all models in less than one minute can be considered an impressive feat in terms of execution time. The calibration experiments based on 90–10 hold-out scheme, the minimum and maximum values for the R^2 of 30 runnings are presented.

The results showed a remarkable improvement in the predictive potential for all models after calibration, which was expected. The model predictive potentials were considered good for CHRISTENSEN JUNIOR *et al.* [3] and excellent for the other two. The model by SAKHAEIFAR *et al.* [7] remained the highest R^2 value after calibration, considering the Student's *t*-test with 95% of confidence level for two independent samples with unequal variances (*p*-value = 5.7E-31). The underperformance of the CHRISTENSEN JUNIOR *et al.* [3] model may be related to the fact that it only uses three variables in the equation, which may potentially affect the ability of the model to explain the behavior of mixtures from this dataset.

The *CI* is an indicator of the uncertainty margin in relation to a sampled mean value, informing the region where the population mean is at a certain confidence level. For instance, if the analysis is repeated many times at 95% confidence level, in 95% of the times the *CI* would contain the population mean. A narrower *CI* for high confidence levels indicates a smaller uncertainty about the sample mean value, which may be considered a good estimate of the population mean. The small *CIs* for 95% confidence levels presented in Table 4 indicate the robustness of the obtained R^2 values. It should be noted that *CI* calculation requires multiple samples, which does not make sense when using resampling techniques such as hold-out. Furthermore, the use of individual assessments as in hold-out sampling often produced R^2 values outside the *CIs* obtained from repeated cross-validation, as suggested by the maximum and minimum values shown in Table 4 and observed from the results, since only 2% of them were inside the calculated *CIs*. Variations of less than 5% in R^2 may not be relevant for estimates in pre-design phases, but rigorous care is needed when concerns the process of sampling, development, and evaluation of prediction models for pavement design parameters in order to reduce the uncertainty about the expected performance of these models.

When analyzing the results of the 90–10 hold-out scheme, it is observed that the predictive potential of the CHRISTENSEN JUNIOR *et al.* [3] calibrated model was classified as fair in 30% of the runs or as good in the others, depending on the sampling. Furthermore, the performance of the BARI and WITCZAK [4] calibrated model might be erroneously considered superior to the SAKHAEIFAR *et al.* [7] one since its best result was superior than the worst of the latter. Thus, hold-out can induce unrealistic conclusions about the performance estimated for the models, even for robust and widely tested ones, such as those evaluated in this study. Therefore, hold-out is strongly not recommended as sampling scheme for model assessment.

The variability of R^2 values can be observed, particularly, in the analysis of the histograms shown in Figures 3 and 4. Figure 3 shows *RMSE* and R^2 histograms for the 90–10 hold-out scheme, while Figure 4 shows

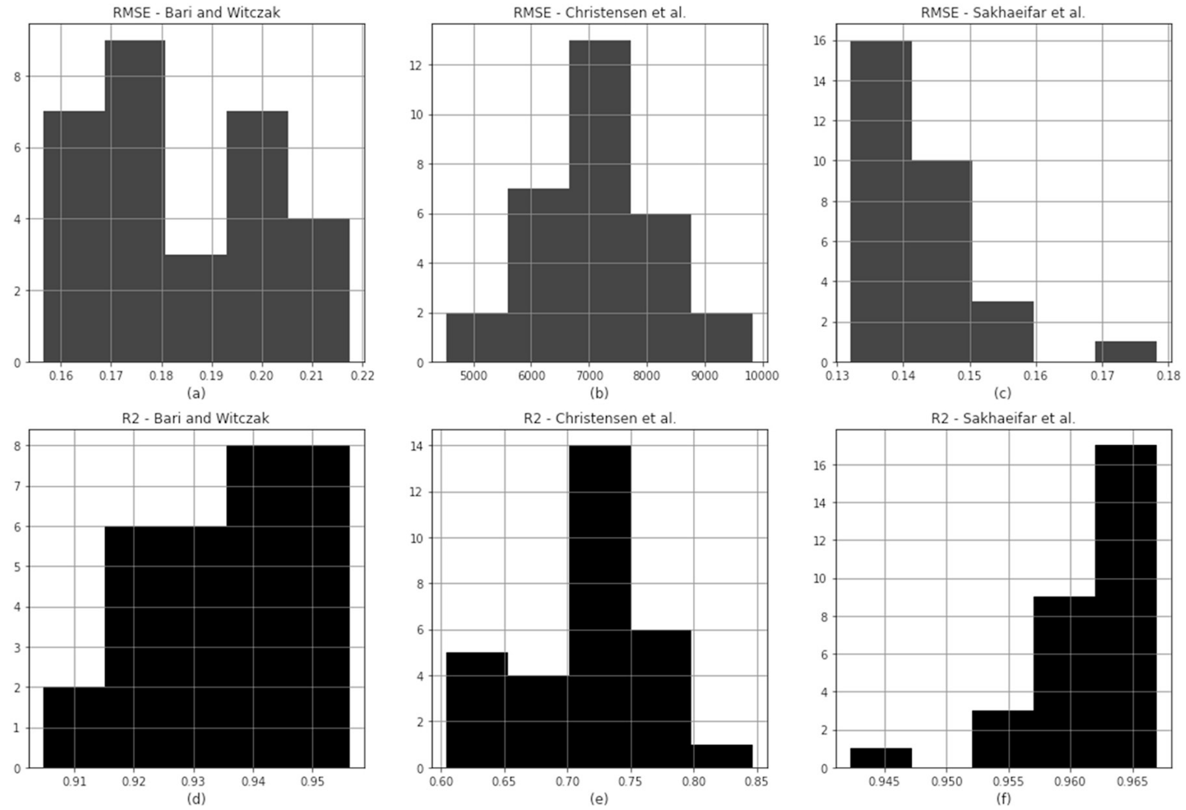

Figure 3: *RMSE* and R^2 histograms for 90–10 hold-out scheme.

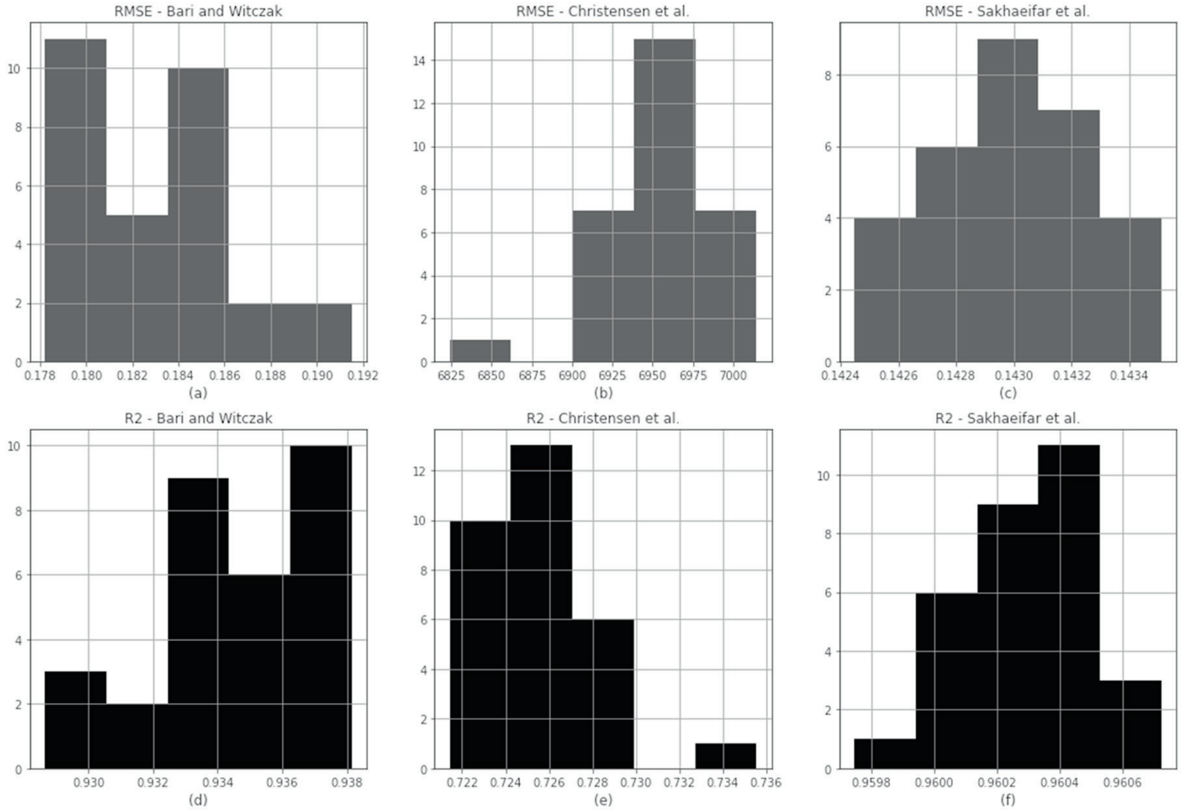

Figure 4: *RMSE* and R^2 histograms for repeated 10-fold CV**.**

them for the repeated 10-fold CV. Note in the histograms that the ranges of R^2 and *RMSE* values obtained for CV were considerably smaller compared to those obtained with hold-out. Regarding the histograms obtained with hold-out, it is noted that most of the cases evaluated are outside the values calculated with confidence intervals, which reinforces the idea that the results obtained with hold-out present a higher degree of uncertainty.

In general, it may be observed that repeated CV is a resampling technique that allows an evaluation of predictive models that is less sensitive to the subset of data used, providing estimates with smaller variances. Thus, as in machine learning community, this technique is strongly recommended to be adopted in calibrations of models that attempt to predict the mechanical characteristics of asphalt mixtures, such as their dynamic modulus.

8. SUMMARY AND CONCLUSIONS

This work reviewed several resampling techniques used in the calibration, selection, and evaluation of predictive models. Among them, although repeated k-fold CV is recognized as a good technique for the evaluation of the generalization ability and robustness of a model, it is not widely adopted on some other engineering tasks, such as the calibration of empirical models.

Hold-out and repeated k-fold CV were employed to evaluate the performance of the calibration of three dynamic modulus predictive models available in the literature. Those were proposed by CHRISTENSEN JUNIOR *et al.* [3], BARI and WITCZAK [4], and SAKHAEIFAR *et al.* [7]. For that, a dataset containing 1,806 dynamic modulus experimental values from 65 asphalt mixtures was applied.

The results using repeated k-fold CV demonstrated a considerable improvement in the predictive potential of the models after calibration, as expected. All calibrated models presented good or excellent predictive potentials. The results using the 90–10 hold-out scheme revealed that hold-out may induce unrealistic conclusions about the estimated performance of models, even for robust and widely tested ones, such as those evaluated in this study. Therefore, hold-out is strongly not recommended as sampling scheme for model assessment.

As each calibration took less than a minute to be performed using a regular computer, the recommendation to adopt more robust evaluation procedures becomes stronger, even if they have higher computational costs. The impact of multiple iterations in repeated k-fold CV on the calibration running time was irrelevant in view of the gain in performance reliability, since the method allowed finding better estimates and confidence intervals. The calibrated models presented small intervals, which gives less uncertainty to their R^2 values and expected performances when they are applied to new predictions. So, repeated k-fold CV is a highly recommended resampling technique to be adopted in model calibration processes, as it allows better assessment of the model's predictive capabilities regardless the subset of data considered.

9. ACKNOWLEDGMENTS

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de Financiamento 001, by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and by CENPES/PETROBRAS. The authors thank CENPES/PETROBRAS, Universidade Federal do Ceará, and Universidade Federal de Santa Maria for providing part of the experimental data used in this research.

10. BIBLIOGRAPHY

[1] NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM, *NRC guide for mechanistic-empirical design of new and rehabilitated pavement structures – final report*, Washington, 2004. http://onlinepubs.trb.org/onlinepubs/archive/mepdg/guide.htm, accessed in October, 2023.

[2] WITCZAK, M., FONSECA, O., "Revised predictive model for dynamic (complex) modulus of asphalt mixtures", *Transportation Research Record: Journal of the Transportation Research Board*, v. 1540, n. 1, pp. 15–23, 1996. doi: http://dx.doi.org/10.1177/0361198196154000103

[3] CHRISTENSEN JUNIOR, D.W., PELLINEN, T., BONAQUIST, R.F., "Hirsch model for estimating the modulus of asphalt concrete", *Electronic Journal of the Association of Asphalt Paving Technologists*, v. 72, pp. 97–121, 2003.

[4] BARI, J., WITCZAK, M.W., "Development of a new revised version of the Witczak *E** predictive model for hot mix asphalt mixtures", *Electronic Journal of the Association of Asphalt Paving Technologists*, v. 75, pp. 381–423, 2006.

[5] MATEOS, A., SOARES, J.B., "Validation of a dynamic modulus predictive equation on the basis of Spanish asphalt concrete mixtures", *Materiales de Construcción*, v. 65, n. 317, e047, 2015. doi: http://dx.doi.org/10.3989/mc.2015.01114

[6] YANG, X., YOU, Z., "New predictive equations for dynamic modulus and phase angle using a nonlinear least-squares regression model", *Journal of Materials in Civil Engineering*, v. 27, n. 3, pp. 1, 2014.

[7] SAKHAEIFAR, M.S., KIM, Y.R., KABIR, P., "New predictive models for the dynamic modulus of hot mix asphalt", *Construction & Building Materials*, v. 76, pp. 221–231, 2015. doi: http://dx.doi.org/10.1016/j.conbuildmat.2014.11.011

[8] FAR, M.S.S., UNDERWOOD, B.S., RANJITHAN, S.R., *et al.*, "Application of Artificial Neural Networks for estimating dynamic modulus of asphalt concrete", *Transportation Research Record: Journal of the Transportation Research Board*, v. 2127, n. 1, pp. 173–186, 2009. doi: http://dx.doi.org/10.3141/2127-20

[9] CEYLAN, H., GOPALAKRISHNAN, K., KIM, S., "Looking to the future: the next-generation hot mix asphalt dynamic modulus prediction models", *The International Journal of Pavement Engineering*, v. 10, n. 5, pp. 341–352, 2009. doi: http://dx.doi.org/10.1080/10298430802342690

[10] RAHMAN, A.S.M.A., TAREFDER, R.A., "Dynamic modulus predictive model based on artificial neural network for the superpave asphalt mixtures of New Mexico", In: *Proceedings of the ASME 2017 International Mechanical Engineering Congress and Exposition – IMECE2017*, Florida, EUA, 2017. doi: http://dx.doi.org/10.1115/IMECE2017-71800.

[11] BARUGAHARE, J., AMIRKHANIAN, A.N., XIAO, F., *et al.*, "ANN-based dynamic modulus models of asphalt mixtures with similar input variables as Hirsch and Witczak models", *The International Journal of Pavement Engineering*, v. 23, n. 5, pp. 1328–1338, 2022. doi: http://dx.doi.org/10.1080/10298436.2020.1799209

[12] DANESHVAR, D., BEHNOOD, A., "Estimation of the dynamic modulus of asphalt concretes using random forests algorithm", *The International Journal of Pavement Engineering*, v. 23, n. 2, pp. 250–260, 2022. doi: http://dx.doi.org/10.1080/10298436.2020.1741587

[13] LIU, J., YAN, K., YOU, L., *et al.*, "Prediction models of mixtures' dynamic modulus using gene expression programming", *The International Journal of Pavement Engineering*, v. 18, n. 11, pp. 971–980, 2017. doi: http://dx.doi.org/10.1080/10298436.2016.1138113

[14] DEVROYE, L., WAGNER, T.J., "Distribution-free performance bounds for potential function rules", *IEEE Transactions on Information Theory*, v. 25, n. 5, pp. 601–604, 1979. doi: http://dx.doi.org/10.1109/TIT.1979.1056087

[15] GOOD, P.I., *Resampling methods: a practical guide to data analysis*, 3 ed., Boston, Birkhäuser, 2006.

[16] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J.H., *The elements of statistical learning: data mining, inference, and prediction*, New York, Springer-Verlag, 2009. doi: http://dx.doi.org/10.1007/978-0-387-84858-7

[17] KOHAVI, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection", In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 2*, pp. 1137–1143, Montreal, Canada, 1995.

[18] BERRAR, D., "Cross-validation", In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (eds), Encyclopedia of bioinformatics and computational biology, vol. 1, Amsterdam, Elsevier, pp. 542–545, 2019. doi: http://dx.doi.org/10.1016/B978-0-12-809633-8.20349-X)

[19] REFAEILZADEH, P., TANG, L., LIU, H., "Cross-validation", In: Refaeilzadeh, P., Tang, L., Liu, H. (eds), Encyclopedia of database systems, Boston, Springer, 2009. doi: http://dx.doi.org/10.1007/978-0-387-39940-9_565

[20] EFRON, B., TIBSHIRANI, R., A*n introduction to the bootstrap*, New York, Chapman & Hall, 1993. doi: http://dx.doi.org/10.1007/978-1-4899-4541-9

[21] CHERNICK, M.R., *Bootstrap methods: a guide for practitioners and researchers*, 2 ed., Hoboken, Wiley, 2007. doi: http://dx.doi.org/10.1002/9780470192573

[22] BOUCKAERT, R.R., "Choosing between two learning algorithms based on calibrated tests", In: *Proceedings of 20th International Conference on Machine Learning*, pp. 51–58, Washington, DC, USA, 2003.

[23] DIETTERICH, T.G., "Approximate statistical tests for comparing supervised classification learning algorithms", *Neural Computation*, v. 10, n. 7, pp. 1895–1923, 1998. doi: http://dx.doi.org/10.1162/089976698300017197. PubMed PMID: 9744903.

[24] STONE, M., "Cross-validatory choice and assessment of statistical predictions", *Journal of the Royal Statistical Society. Series B. Methodological*, v. 36, n. 2, pp. 111–133, 1974. doi: http://dx.doi.org/10.1111/j.2517-6161.1974.tb00994.x

[25] GEISSER, S., "The predictive sample reuse method with applications", *Journal of the American Statistical Association*, v. 70, n. 350, pp. 320–328, 1975. doi: http://dx.doi.org/10.1080/01621459.1975.10479865

[26] BREIMAN, L., SPECTOR, P., "Submodel selection and evaluation in regression. the x-random case", *International Statistical Review*, v. 60, n. 3, pp. 291–319, 1992. doi: http://dx.doi.org/10.2307/1403680

[27] BURMAN, P., "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods", *Biometrika*, v. 76, n. 3, pp. 503–514, 1989. doi: http://dx.doi.org/10.1093/biomet/76.3.503

[28] BREIMAN, L., "Heuristics of instability and stabilization in model selection", *Annals of Statistics*, v. 24, n. 6, Dec. 1996. doi: http://dx.doi.org/10.1214/aos/1032181158

[29] KRSTAJIC, D., BUTUROVIC, L.J., LEAHY, D.E., *et al*., "Cross-validation pitfalls when selecting and assessing regression and classification models", *Journal of Cheminformatics*, v. 6, n. 1, pp. 10, 2014. doi: http://dx.doi.org/10.1186/1758-2946-6-10. PubMed PMID: 24678909.

[30] HARRELL, F., *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis*, New York, Springer, 2015. doi: http://dx.doi.org/10.1007/978-3-319-19425-7

[31] BYRD, R.H., LU, P., NOCEDAL, J., *et al*., "A limited memory algorithm for bound constrained optimization", *SIAM Journal on Scientific Computing*, v. 16, n. 5, pp. 1190–1208, 1995. doi: http://dx.doi.org/10.1137/0916069

[32] WITZCAK, M.W., KALOUSH, K., PELLINEN, T., *et al*., *NCHRP report 465: simple performance test for superpave mix design*, Washington, Transportation Research Board, 2002.

[33] SANTOS, I.G., ESPÍNDOLA, R.P., ARAGÃO, F.T.S., *et al*., "Prediction of the dynamic modulus of Brazilian asphalt mixtures using artificial neural networks", In: *Proceedings of the XLI Ibero-Latin American Congress on Computational Methods in Engineering – XLI CILAMCE*, Foz do Iguaçu, PR, Brasil, 2020.