



Material Consumption Estimation in the Construction of Concrete Road Bridges Using Machine Learning

Miljan Kovačević¹  and Borko Bulajić² 

¹ Faculty of Technical Sciences, University of Pristina, Knjaza Milosa 7,
38220 Kosovska Mitrovica, Serbia

`miljan.kovacevic@pr.ac.rs`

² Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6,
21000 Novi Sad, Serbia

`borkobulajic@uns.ac.rs`

Abstract. The state-of-the-art techniques of machine learning for assessing material consumption in the construction of prestressed concrete (PC) road bridges are described and analyzed in this paper. For model training and evaluation, a database of material consumptions and design features for 75 PC bridges is compiled. The achieved accuracy of the model for estimating prestressed steel consumption per m² was 6.55%, calculated as the mean absolute percentage error. Based on the proposed model, a MATLAB-based software with a graphical user interface was developed to allow for the input of basic variables and the calculation of estimated prestressing steel quantities. The use of such software to determine the ideal amount of steel will, in most cases, result in the construction of higher-quality structures that require less maintenance.

Keywords: Material consumption · Bridges · Prestressed concrete bridges · Machine learning

1 Introduction

In general, using modern algorithms to estimate the ideal amount of steel in a bridge or other RC structure will result in the construction of a higher-quality structures that will require less long-term maintenance. Early resource planning and optimization gives a better foundation for later maintenance that is easier and less expensive.

There are many examples of such algorithm used for bridges. Marcous et al. [1] used a neural network model to model optimal concrete volume and the weight of prestressing steel for 22 prestressed reinforced concrete bridges over the Nile River in Egypt. The result was an error size of 7.5% in terms of predicting the amount of concrete and 11.5% in terms of predicting the amount of prestressing steel. Fragkakakis et al. [2] used regression analyses in modeling the consumption of steel and concrete for bridge foundations in order to consider the costs in the conceptual phase of the project. Mučenski et al. [3] used a neural network model to estimate the required amounts of reinforcement and

concrete in multi-storey buildings based on data from 115 major projects of residential buildings. They obtained the best results using the network trained by Broyden–Fletcher–Goldfarb–Shanno algorithm with an average error of 12.49%. Frangkakis et al. [4] defined regression equations for estimating the consumption of concrete and steel per meter of culvert, based on the database of 104 culverts on the Egnatia highway in Greece. They tested the models using tenfold cross-validation and evaluated against MAPE criteria. The result showed the accuracy of the estimate in terms of consumed concrete and steel, which is 13.78% and 19.79%, respectively. Marineli et al. [5] created a model to estimate the material consumption in bridge superstructure using a multilayer perceptron neural network model based on the data from 68 bridges in Greece. The models were evaluated via the correlation coefficient R . The construction costs and consumption of steel and concrete in underpasses were analyzed by Antoniou et al. [6]. Their database included data from 28 underpasses in Greece. The paper points out the average values per m^2 of material consumption and the actual costs of construction of underpass bridges. Dimitriou et al. [7] analyzed a model for estimating the consumption of steel and concrete in the construction of road bridges using linear regression models and neural network models, based on the data on 68 bridges in Greece. The value of MAPE and R^2 have been used to assess model accuracy. In the prediction of concrete consumption only for the bridge superstructure, the value of MAPE ranged from 11.48% to 16.12% depending on the bridge span, while the values for R^2 range from 0.979 to 0.995. In terms of consumption of concrete and steel per column, the obtained values of MAPE and R^2 were, respectively, 37% and 0.962 for concrete and 31% and 0.962 for steel.

In this paper, we will develop, examine, and verify a number of prediction models for early estimation of prestressed steel consumption per m^2 , based on artificial intelligence approaches. For the assessment model, a database of design attributes for 75 PC bridges in Serbia is constructed. All models will be trained and tested using cross-validation under the same conditions.

2 Methods

2.1 Multilayered Perceptron Artificial Neural Network (MLP-ANN)

A multilayer perceptron is a forward signal-propagating neural network having minimum of three neurons layers: the input, the hidden, and the output. As shown in Fig. 1, every neuron of one layer is connected to every neuron of the next layer.

The number of neurons in the hidden layer and the activation type have a major impact on the network's features. When MLP-ANN is applied as an approximator, activation functions are usually chosen to be continuous and differentiable. A linear activation function is most typically used in the output layer. If the case of sufficient number of neurons in the hidden layer [8], an arbitrary multidimensional function can be approximated for a given data set using the MLP-ANN model with a single hidden layer whose neurons' activation function is bipolar sigmoid while for the output layer the activation function is linear.

The number of neurons in the hidden layer is established through trial and error, with certain guidelines taken into account. The following formulas [8–11] can be used to estimate the maximum of the number of hidden layer neurons N_H :

$$N_H \leq 2 \times N_i + 1 \tag{1}$$

$$N_H \leq \frac{N_s}{N_i + 1}. \tag{2}$$

where N_i is the number of inputs, and N_s is the number of instances used for training. Less than two acquired values for the number of neurons are accepted utilizing these formulas based on the obtained values for the number of neurons.

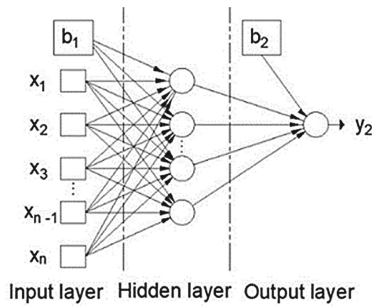


Fig. 1. Schematic presentation of a multilayer perceptron artificial neural network [8].

2.2 Regression Tree Ensembles

The binary tree depicted in Fig. 2 represents an example of a regression tree model, in which the entire set of data is started at the top of the tree, and observations are divided on each node based on whether or not they meet specified conditions in that node. Individual regions ($R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8$) correspond to the tree’s end nodes or leaves.

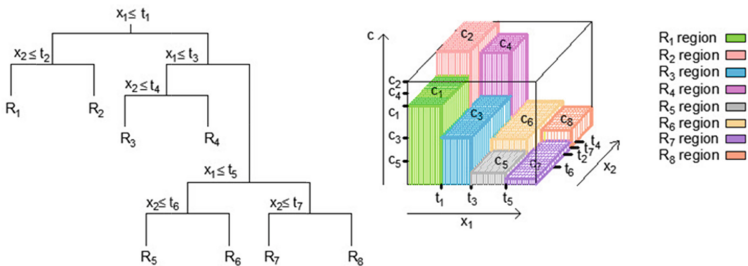


Fig. 2. Examples of space segmentation (left) and 3D regression surface in the regression tree (right) [11].

The regression model assigns a constant value c_m to each region based on the defined input variables:

$$\hat{f}(X) = \sum_{m=1}^8 c_m I\{(X_1, X_2) \in R_m\} \tag{3}$$

where the constant c_m represents the mean of the output variable for region R_m , and $I\{(X_1, X_2) \in R_m\}$ is an indicator function that has a value of one for exact statements and zero for all other statements.

Assuming that space segmentation is performed in M domains R_1, R_2, \dots, R_M , the output model with a constant value of c_m for each region in this case has the following form

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m). \tag{4}$$

The optimal value \hat{c}_m is the mean value y_i for the region R_m , as determined by the criterion of minimizing the sum of the squares $\sum (y_i - f(x_i))^2$:

$$\hat{c}_m = ave(y_i | x_i \in R_m). \tag{5}$$

where *ave* is the mean value.

A Greedy algorithm [12, 13] is used to find the input space’s binary segmentation point. When using this algorithm, only the best results from each step of the process are considered, with no consideration for subsequent steps.

Let us consider the variable j on which the split will be conducted, as well as the value or split point s , and define a pair of half-planes starting with the entire set of data:

$$R_1(j, s) = \{X | X_j \leq s\} \text{ i } R_2(j, s) = \{X | X_j > s\}. \tag{6}$$

The variable j , as well as the value of the split point s , must be determined to meet the following expression:

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]. \tag{7}$$

By using the mean y for the region R_1 and the mean y for the region R_2 as follows, the value in square brackets is minimized:

$$\hat{c}_1 = ave(y_i | x_i \in R_1(j, s)) \text{ and } \hat{c}_2 = ave(y_i | x_i \in R_2(j, s)). \tag{8}$$

The split point can be determined by examining all the model’s input variables. The operation continues with further splitting of the space until a specified stop criterion is met after finding the optimal point (which corresponds to the smallest value of the formula (7)). A similar approach is used for multidimensional input variables, with the exception that it is a matter of space segmentation in so-called hyperboxes in that case. There is a chance that the produced regression tree has very good performance on the training set but poor generalization on the test data set if the technique for constructing a regression tree model is followed.

The bootstrap sampling method [14–16], i.e., repeated sampling for the same training dataset, can solve the aforementioned difficulty. Each of these sets can be used to train a regression tree on its own. The ensemble is made up of the trees that have been generated in this manner.

All input variables are considered potential splits in the Treebager method, but just a subset of the input variables is used for tree splitting in the Random Forest method. If the prediction function for the b -th bootstrap training set is $\hat{f}^{*b}(x)$ at the point x , then an ensemble model [17] is created by averaging all B models (Fig. 3), with the prediction function specified by the expression:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x). \tag{9}$$

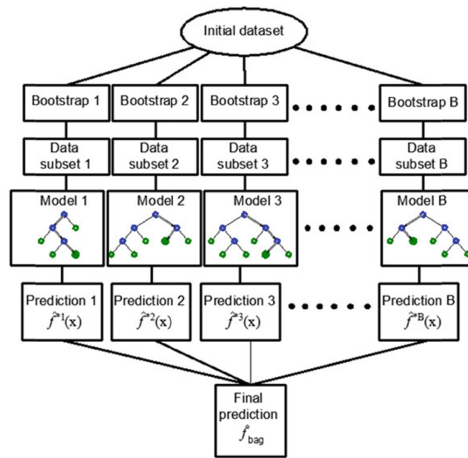


Fig. 3. Regression tree ensembles with Bootstrap aggregation-bagging [11].

The Boosting approach employs sequential training of models, with new regression trees that are added increasing the preceding tree collection’s performance [17]. For quadratic error function (Gradient Boosting method), each sub-sequent step (Fig. 4) adds a new sub-model to the basic model that best guesses the residuals of the preceding model. By applying an iterative approach to the addition of a model, a definite model is defined. Such model is an ensemble of the models that were obtained in previous steps.

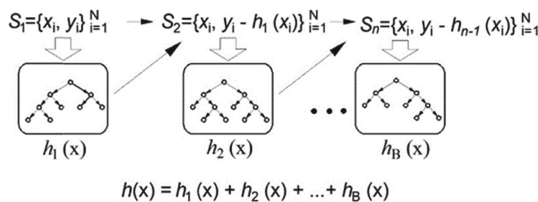


Fig. 4. Regression tree ensembles using gradient boosting [11].

2.3 Support Vector Regression (SVR)

Consider the following training datasets: $\{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \in \mathbb{R}^n \times \mathbb{R}$, where $x_i \in \mathbb{R}^n$ is the n-dimensional vector expressing the inputs and y_i are the responses. The approximation function can be written as:

$$f(x) = \sum_{i=1}^l (\alpha_i^* - \alpha_i) \mathbf{K}(\mathbf{x}_i, \mathbf{x}) + b. \tag{10}$$

The empirical risk function needs to be minimized. The problem of minimizing of the empirical risk function can be solved by using the Vapnik’s linear loss function with ϵ -insensitivity zone (Fig. 5), which is defined as follows [18, 19]:

$$|y - f(\mathbf{x}, \mathbf{w})|_\epsilon = \begin{cases} 0 & \text{if } |y - f(\mathbf{x}, \mathbf{w})| \leq \epsilon \\ |y - f(\mathbf{x}, \mathbf{w})| - \epsilon & \text{otherwise.} \end{cases} \tag{11}$$

As a result, the problem can be simplified to minimization of the equation below:

$$R = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l |y_i - f(\mathbf{x}_i, \mathbf{w})|_\epsilon. \tag{12}$$

By introducing slack variables ξ and ξ^* , which are shown in Fig. 5, minimization of R is equivalent to the minimization the following:

$$R_{w, \xi, \xi^*} = \frac{1}{2} \left[\|\mathbf{w}\|^2 + C \left(\sum_{i=1}^l \xi + \sum_{i=1}^l \xi^* \right) \right], \tag{13}$$

To implement the model, the LIBSVM software [20, 21] was utilized within the MATLAB program.

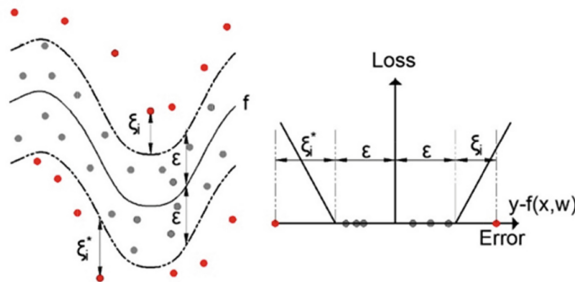


Fig. 5. Nonlinear SVR with the ϵ -insensitivity zone [11].

2.4 Gaussian Process Regression (GPR)

Limited collections of the random variables are characterized by a multivariate normal distribution. In other words, a finite linear combination of random variables is normally distributed in a Gaussian process. Gaussian process regression model is a probability

distribution over alternative functions, which fit a set of observed points. Let us consider the following nonlinear regression problem:

$$y = f(x) + \varepsilon, \varepsilon \sim N(0, \sigma^2). \tag{14}$$

where $f(\cdot) : R^n \rightarrow R$ is an unknown function that must be estimated, y is the target and x are the input variables while ε is an additive noise that is normally distributed.

The Gaussian process-type regression [22] implies that $f(\cdot)$ has a Gaussian distribution with a mean function $\mu(\cdot)$, and a covariance function $k(\cdot, \cdot)$. Here, the observations in an arbitrary data set $\mathbf{y} = \{y_1, \dots, y_n\}$ may always be thought of as a sample from a multivariate Gaussian distribution:

$$(y_1, \dots, y_n)^T \sim N(\boldsymbol{\mu}, K), \tag{15}$$

where $\boldsymbol{\mu} = (\mu(x_1), \dots, \mu(x_n))^T$ is the mean vector, K is the $n \times n$ covariance matrix with the (i, j) element defined as $K_{ij} = k(x_i, x_j) + \sigma^2 \delta_{ij}$, and the Kronecker delta function is represented by δ_{ij} . Let x^* be any of the test points and y^* be the response. The joint distribution (y_1, \dots, y_n, y^*) is a $(n + 1)$ variate normal distribution $(y_1, \dots, y_n, y^*) \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}^* = (\mu(x_1), \dots, \mu(x_n), \mu(x^*))^T$ and covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} K_{11} & K_{12} & \dots & K_{1n} & K_{1*} \\ K_{21} & K_{22} & \dots & K_{2n} & K_{2*} \\ \dots & \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & K_{nn} & K_{n*} \\ K_{*1} & K_{*2} & \dots & K_{*n} & K_{**} \end{bmatrix} = \begin{bmatrix} K & K^* \\ K^{*T} & K^{**} \end{bmatrix} \tag{16}$$

where $K^* = (K(x^*, x_1), \dots, K(x^*, x_n))^T$ and $K^{**} = K(x^*, x^*)$.

Given $\mathbf{y} = (y_1, \dots, y_n)^T$, the conditional distribution of y^* is $N(\hat{\mathbf{y}}^*, \hat{\boldsymbol{\sigma}}^{*2})$ with

$$\hat{\mathbf{y}}^* = \mu(x^*) + K^{*T} K^{-1} (\mathbf{y} - \boldsymbol{\mu}) \tag{17}$$

$$\hat{\boldsymbol{\sigma}}^{*2} = K^{**} + \sigma^2 - K^{*T} K^{-1} K^*. \tag{18}$$

Using the automatic relevance determination (ARD), hyperparameters may be utilized to define the inputs that are more relevant than the others. Let us consider the squared exponential covariance function and variable length scale parameters used for every input (ARD SE):

$$k(x_p, x_q) = v^2 \exp \left[-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_p^i - x_q^i}{r_i} \right)^2 \right] \tag{19}$$

where r_i is the covariance function’s length scale along input dimension i . The input is reduced if r_i is very large [10, 17].

The maximum likelihood technique can be used to estimate the noise variance σ^2 and the hyperparameters of covariance functions. The training data’s log-likelihood is given by:

$$L(v, r_1, \dots, r_n, \sigma^2) = -\frac{1}{2} \log \det K - \frac{1}{2} \mathbf{y}^T K^{-1} \mathbf{y} - \frac{n}{2} \log 2\pi. \tag{20}$$

3 Dataset

The proposed material consumption methods are based on the creation of a data collection that comprises project and contract documentation for prestressed concrete bridges built in Serbia's Corridor X (Fig. 6) [10]. On the eastern and southern branches of Corridor X, the bridge data set comprises complete data on 75 PC bridges (prefabricated or cast on site). The choice of input variables is critical in predicting material consumption in the construction of PC bridges because the bridge design is often influenced by a number of variables.

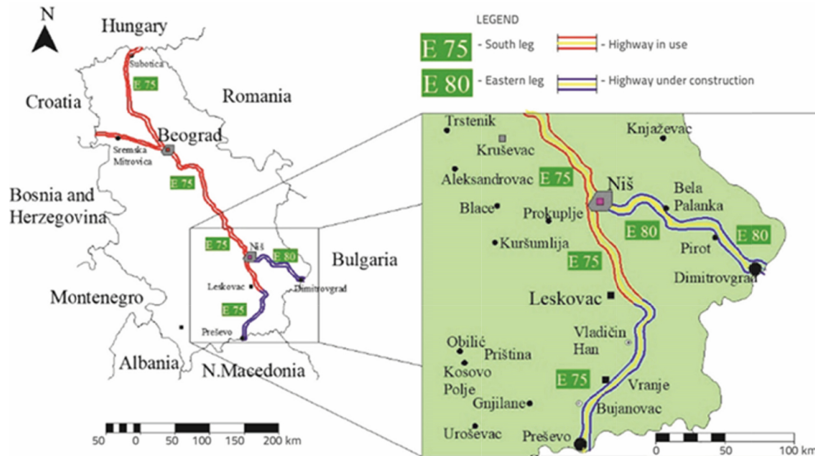


Fig. 6. Eastern and southern legs of Corridor X in Serbia [10].

To estimate the amount of prestressed steel, the following factors were used as independent variables in the model: x_1 – maximum individual bridge span, x_2 – average span, x_3 – total bridge span, x_4 – bridge width (Table 1).

Table 1. Mean, minimum and maximum values of variables in the model used to estimate the prestressed steel consumption per m^2 of the bridge superstructure

Variable	Average value	Minimum value	Maximum value
Max. individual bridge span [m]	31.58	18.00	49.00
Average bridge span [m]	30.74	17.60	44.91
Total bridge span [m]	161.33	21.20	628.74
Bridge width [m]	12.81	8.40	17.80

Based on project documents, the dependent variable is y - mass in kg of prestressed steel per m^2 [23] of the bridge superstructure. The properties of prestressing steel ropes are listed in the Table 2.

4 Evaluation and Performance Measures

To assess the model quality, the following errors were analyzed: RMSE (root mean square, MAE (mean absolute), R (Pearson’s Linear Correlation Coefficient), and MAPE (mean absolute percentage). The RMSE, described by Eq. (21), is a measure of the model’s overall accuracy, and is given in the same units as the modeled quantity:

$$RMSE = \sqrt{\frac{1}{N} \sum_{k=1}^N (d_k - o_k)^2} \tag{21}$$

where d_k is the actual value (the target value), o_k is the modeled output (forecast), and N is the number of the used training samples.

Table 2. Geometrical and mechanical rope characteristics [23].

Abbreviated name	Class	Nominal values			Guaranteed values			Maximum relaxation at a force of 0.7 Fpk after 1000 h [2.5%]
		Diameter Ø [mm]	Tensile strength [N/mm ²] fpk	Elastic modulus [kN/mm ²] E	Section area [mm ²] Apk	Characteristic breaking force [kN] Fpk	Characteristic 0.1% proof-stress of prestressing steel [kN] Fp0.1k	
Y1770S7	A	15.2	1770	195	140	248	213	2.5
Y1860S7	B	15.2	1860		140	260	224	
Y1770S7	A	16.0	1770		150	265	228	
Y1860S7	B	16.0	1860		150	279	240	

The MAE is also a measure of the model’s absolute accuracy. It is used to estimate the model’s mean absolute error as follows:

$$MAE = \frac{1}{N} \sum_{k=1}^N |d_k - o_k| \tag{22}$$

The R coefficient, defined by expression (23), is a relative criterion for the evaluation of the model’s accuracy:

$$R = \sqrt{\left[\sum_{k=1}^N |(d_k - \bar{d})(o_k - \bar{o})| \right]^2 \times \left[\sum_{k=1}^N \left| (d_k - \bar{d})^2 (o_k - \bar{o})^2 \right| \right]^{-1}} \tag{23}$$

where \bar{o} is the predicted mean that is obtained by the related model, while \bar{d} is the mean target value.

The MAPE criterion, which is defined by the Eq. (24), is a measure of the relative accuracy of the model prediction.

$$MAPE = \frac{100}{N} \sum_{k=1}^N \left| \frac{d_k - o_k}{d_k} \right| \tag{24}$$

The model cross-validation approach is used in the analysis. The approach was chosen because it lowers the bias associated with random data partition into training and training models, which should, in theory, have the same statistical features. The ten-fold cross-validation procedure, for example, must be performed by dividing the data set randomly into ten disjoint sections of the same size (10 folds) [24].

5 Results and Discussion

Because it is a regression problem, the output layer neurons number is one in the MLP model, which has an architecture with the same number of input layer neurons as predictors, i.e., with four input layer neurons. For the hidden layer, the number of neurons was determined experimentally by observing the upper value obtained using Eqs. (1) and (2). The Levenberg-Marquardt algorithm [25] was employed to train the MLP-ANN that had one hidden layer. The maximum epochs number of 1000, the minimum gradient magnitude of 10^{-5} , or the mean square error of 0 were used to determine when the training should be stopped. Prior to training, the normalization of the input data was done in the range $[-1, 1]$.

In this situation, network architectures with up to 8 neurons in the hidden layer must be examined. In terms of accuracy criteria, two similar models can be seen (Fig. 7). According to the RMSE and R criterion, the model with three neurons within the hidden layer is superior, while the model with four neurons within the hidden layer is better in terms of MAE and MAPE (Table 4).

The parameters of the model were adjusted using the cross-validation procedure for developing a model for estimating the amount of prestressed steel utilizing ensembles of regression trees, and in order to attain appropriate predictive performance. The following methods were used to generate the model:

- Bagging method (TreeBagger),
- Random Forests (RF) method,
- Boosted Trees method.

Alternative values of adjustable model parameters were analyzed when using the TreeBagger (TB) approach, as follows:

- The number of the generated trees, B . The maximum number of generated trees is limited to 500. The number of created trees is set to 100 by default in the MATLAB software that is used to apply the Bagging method. The method of bootstrap aggregation was employed to construct trees, in which samples of the same size as the original samples were formed, i.e., a total of 75 samples were formed throughout each iteration to generate a tree model.
- Within the tree, the min leaf size, i.e., the minimum number of samples that are assigned to the leaf. Values ranging from one to ten samples were evaluated, with a step of one per tree leaf. When using the Bagging method in MATLAB, the default setting takes 5 samples per tree leaf for regression problems. Within this study, a somewhat greater range for the number of data per tree leaf was explored, as well as the impact of the number of data on the model's generalization.

The RF method randomly selects a subset of input variables that are split in a tree. Different values of adaptive model parameters were analyzed when this method was implemented, as follows:

- The number of variables in the tree that are split. According to L. Breiman’s paper Random Forests [26], the subset m of the variables subjected to the splitting should be approximately \sqrt{p} or $p/3$ predictors, depending on whether the task is a classification or regression problem. The number of input variables or predictors in this study is 4, therefore when the second regression criterion is used, the subset of variables on which splits can be is 1 or 2. In this study, a slightly broader range of models is examined (Fig. 8), namely models with two, three, and eventually four variables as a subset of the variables subjected to the splitting. The TB method is obtained after analyzing the splitting with all four variables.
- Within the tree, the min leaf size, i.e., the minimum number of samples that are assigned to the leaf. We considered the values from 1 to 10 samples (Fig. 8) with the step of 1 per tree leaf. When using the RF method in MATLAB, the default configuration is to take 5 samples per tree for regression issues. In this study, it is analyzed a slightly greater range for the number of data per tree leaf.

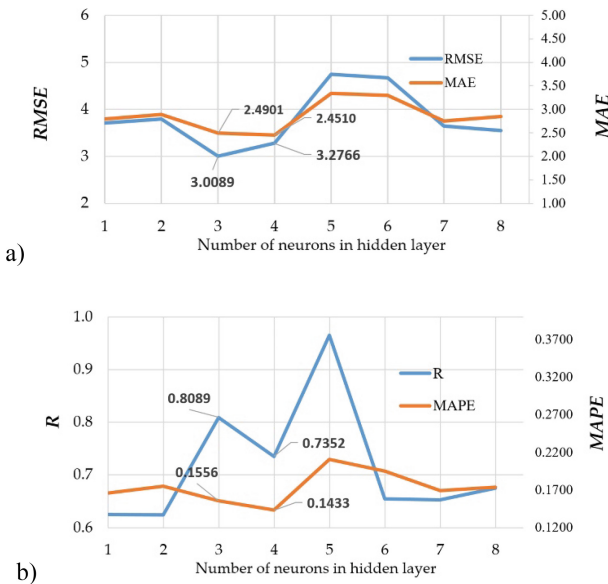


Fig. 7. Performance measures compared using MLP-ANNs with varying number of neurons within the hidden layer: a) the RMSE and MAE criteria, b) the R and MAPE criteria.

The parameters of the TB and RF procedures were defined using the grid-search method. The impact of the number of splitting variables and the minimal number of data for terminal leaf were investigated. The ideal model (Fig. 8) was found to be the RF model, in which two input variables are separated and the number of the data per leaf is kept to a minimum of one.

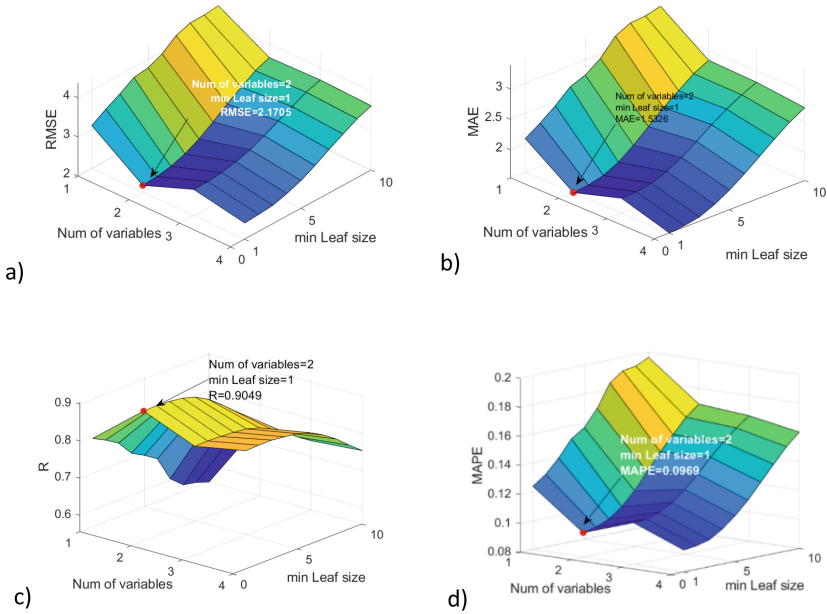


Fig. 8. Comparison of performance measures for tree-based ensemble methods a) RMSE, b) MAE, c) R, d) MAPE.

In both the Boosting and other methods, the cross-validation procedure was utilized to identify the parameters of the best model. The model parameters for the Boosting Trees approach are as follows:

- The number of generated trees, B . When using the Boosting approach, there is a risk of model overtraining if too many trees are formed. Since the number of the analyzed models in this research is large, the number of trees (base models) in the ensemble is limited to 100.
- Reduction parameter λ (learning rate). The λ parameter defines the model's training speed. Although the usual values are 0.01 and 0.001, we analyzed several values: 0.001; 0.01; 0.1; 0.5; 0.75, and 1.0.
- The number of splits in a tree, d . The values of the number of splits with an exponential increase were analyzed. Tree models were created with a number of splits limited to $2^0 = 1, 2^1, 2^2, 2^3, 2^4, 2^5, 2^6 = 64$.

Figure 9 shows the influence of the adopted parameters on the model’s accuracy, i.e., the influence of the maximum number of splits and reduction parameters on the model’s accuracy in terms of the MSE value obtained by cross-validation. The boundary cases are modeled with maximum number of splits equal to 64. The optimal model with the lowest MSE value for the evaluated ensemble of 100 base models is shown in Fig. 9 and is represented by a yellow line.

The following parameter values were determined for the optimal model: a total of 100 trees, learning rate of 0.1, and a number of splits limited to 8.

The use of various SVM kernel functions is examined in this study to discover the optimal one. SVR models using RBF, linear, and sigmoid kernels were investigated. Before training and testing the model, all input data were transformed into the range (0, 1) via normalization. For all kernels, the best model was found using the grid search technique ($C = 0.6208$, $\epsilon = 0.0642$ for the linear kernel; $C = 1.2263$, $\epsilon = 0.0302$, $\gamma = 7.9800$ for the RBF kernel; $C = 243.9853$, $\epsilon = 0.0656$, $\gamma = 0.0028$ for the sigmoid kernel).

A comparison examination of several SVR models reveals that the models have different accuracy for different kernel functions. According to several criteria, models with the linear and sigmoid kernels exhibit equivalent accuracy. Considering all criteria functions, the SVR model with the RBF kernel has significantly highest accuracy.

The application of several covariance functions was investigated when analyzing the consumption of prestressed steel per m^2 of superstructure using GPR. The data used in modeling are standardized. The so-called Z-score standardization [17] procedure was performed on each column. Models with a constant base function were analyzed. The parameter values are defined by maximization of the log marginal probability.

Table 3 shows the model’s parameters for different covariance functions. Table 4 shows the parameters of the model with ARD covariance functions. Table 5 shows the results of different machine learning models.

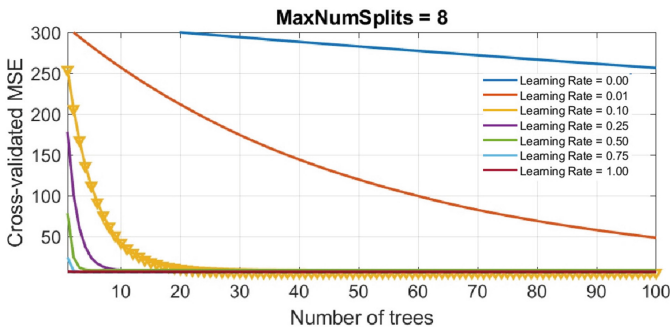


Fig. 9. Dependence of the MSE value on the reduction parameter λ and the number of trees (base models) used in the Boosted Trees method when the maximum number of splits is limited to 8.

Table 3. The GPR model covariance functions parameters

GP model covariance function	Covariance function parameters		
Exponential	$k((x_i, x_j \Theta)) = \sigma_f^2 \exp\left[-\frac{1}{2} \frac{r}{\sigma_l^2}\right]$		
	$\sigma_l = 5.4002$	$\sigma_f = 9.9766$	
Squared exponential	$k((x_i, x_j \Theta)) = \sigma_f^2 \exp\left[-\frac{1}{2} \frac{(x_i - x_j)^T (x_i - x_j)}{\sigma_l^2}\right]$		
	$\sigma_l = 2.4565$	$\sigma_f = 10.7092$	
Matern 3/2	$k((x_i, x_j \Theta)) = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{\sigma_l}\right) \exp\left[-\frac{\sqrt{3}r}{\sigma_l}\right]$		
	$\sigma_l = 5.1291$	$\sigma_f = 13.3442$	
Matern 5/2	$k((x_i, x_j \Theta)) = \sigma_f^2 \left(1 + \frac{\sqrt{5}r}{\sigma_l} + \frac{5r^2}{3\sigma_l^2}\right) \exp\left[-\frac{\sqrt{5}r}{\sigma_l}\right]$		
	$\sigma_l = 3.6042$	$\sigma_f = 11.8095$	
Rational quadratic	$k((x_i, x_j \Theta)) = \sigma_f^2 \left(1 + \frac{r^2}{2a\sigma_l^2}\right)^{-\alpha}$		
	$\sigma_l = 2.6978$	$a = 2.8872$	$\sigma_f = 11.3713$

where $r = \sqrt{(x_i - x_j)^T (x_i - x_j)}$.

In a model, the relevance of specific predictors or variables can be seen utilizing ARD covariance (Automatic Relevance Determination) functions. Higher ARD covariance function parameter values indicate less significance of the variable to which they refer. Models using an ARD covariance function have higher accuracy than models with only one length scale parameter, as shown by the values of the criteria (Table 5).

The study of the length scale parameters values may be used to see the significance of individual variables (Fig. 10) for the model with the highest accuracy. This means that our analysis will be performed on a model with an ARD exponential covariance function.

The variable x_1 , which indicates the maximum individual span in the model with ARD exponential function, has the largest value (the value of the variable from the aspect of the analyzed model is inversely proportional to the length scale parameter) when the values of the length scale parameters are examined. This can be explained by the fact that bridges are usually made up of similar individual spans with little variation, so the variable x_2 , which represents the mean value of the spans, implicitly contains some information about the maximum span with little variation, and this information is hence included in the model.

Table 4. The GPR ARD model covariance functions parameters.

Parameters of covariance function			
σ_1	σ_2	σ_3	σ_4
ARD exponential:			
$k((x_i, x_j \Theta)) = \sigma_f^2 \exp(-r); \sigma_f = 7.9176; r = \sqrt{\sum_{m=1}^d \frac{(x_{im}-x_{jm})^2}{\sigma_m^2}}$			
327822.2661	0.4200	8.4734	170.7902
ARD squared exponential:			
$k((x_i, x_j \Theta)) = \sigma_f^2 \exp\left[-\frac{1}{2} \sum_{m=1}^d \frac{(x_{im}-x_{jm})^2}{\sigma_m^2}\right]; \sigma_f = 8.4546$			
3,3532	0,8216	2,8218	11,5585
ARD matern 3/2:			
$k((x_i, x_j \Theta)) = \sigma_f^2 (1 + \sqrt{3}r) \exp[-\sqrt{3}r]; \sigma_f = 5.7735$			
0,5812	0,9767	71499.7664	0,1014
ARD matern 5/2:			
$k((x_i, x_j \Theta)) = \sigma_f^2 (1 + \sqrt{5}r + \frac{5r^2}{3}) \exp[-\sqrt{5}r]; \sigma_f = 7.1089$			
1,8478	0,3117	4243.2352	4,8180
ARD rational quadratic:			
$k((x_i, x_j \Theta)) = \sigma_f^2 \left(1 + \frac{1}{2\alpha} \sum_{m=1}^d \frac{(x_{im}-x_{jm})^2}{\sigma_m^2}\right)^{-\alpha}; \alpha = 0.1673; \sigma_f = 8.8932$			
1,6509	0,3324	7441.4850	5.2814

where $r = \sqrt{\sum_{m=1}^d \frac{(x_{im}-x_{jm})^2}{\sigma_m^2}}$.

Table 5. Results of different machine learning models

Model	RMSE	MAE	R	MAPE/100
NN 4-3-1	3.0089	2.4901	0.8089	0.1556
NN 4-4-1	3.2766	2.4510	0.7352	0.1433
TreeBagger	2.4138	1.5349	0.8736	0.1433
Random forest	2.1705	1.5326	0.8849	0.0969
Boosted trees	2.3822	1.6650	0.8896	0.1025
SVR-Lin. kernel	3.2536	2.5971	0.7906	0.1498
SVR-RBF kernel	2.4980	1.8839	0.8762	0.1234

(continued)

Table 5. (continued)

Model	RMSE	MAE	R	MAPE/100
SVR-Sig. kernel	3.2484	2.5866	0.7890	0.1490
Exponential	2.6536	1.4834	0.8565	0.0912
ARD-exponential	2.3800	1.3302	0.8870	0.0815
Squared exponential	2.8456	2.0747	0.8396	0.1293
ARD-Sq. exponential	2.4239	1.7913	0.8822	0.1099
Matern 3/2	3.6520	2.2463	0.7113	0.1331
ARD-matern 3/2	3.0418	1.9661	0.8087	0.1254
Matern 5/2	3.1981	2.2096	0.7851	0.1327
ARD-matern 5/2	3.1297	1.9085	0.7937	0.1147
Rational quadratic	2.9855	2.0910	0.8180	0.1289
ARD rational quadratic	2.6568	1.7374	0.8565	0.1069

The length of the bridge, x_3 , is variable with a larger value of the length scale parameter or less relevance in comparison to the variable x_2 , which can be explained by the fact that the output variable reflects prestressed steel consumption expressed in m^2 . The significance of the variable x_3 would be great if the output variable was, for example, the total amount of prestressed steel. For the variable x_4 , i.e., the influence of bridge width, a similar explanation can be given. Steel consumption per m^2 of bridge superstructure is between 12.5 and 22 kg/m^2 in most investigated bridges, which is consistent with the approximate values stated in the literature [27]. Figure 11 depicts the modeled values in respect to the model's target values. Figure 12 shows a regression plot of the modeled and target values in the optimal model. The model that uses the

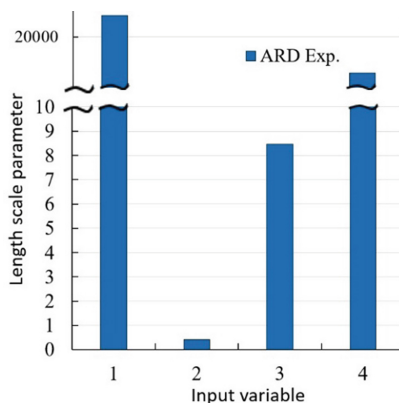


Fig. 10. Length scale parameter values for models with ARD exponential covariance function for prestressed steel consumption models.

most relevant variables is considered further in the analysis. In some circumstances, a model with the same or superior accuracy can be obtained by employing a narrowed set of variables.

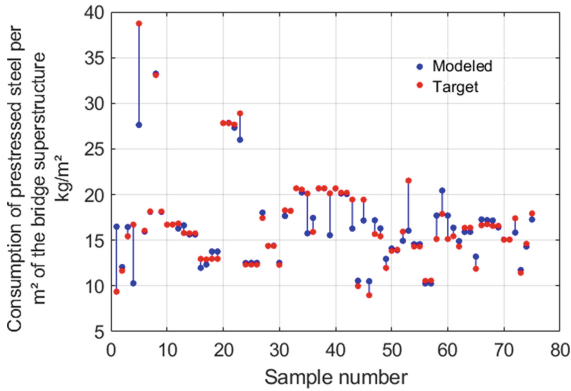


Fig. 11. The GPR ARD Exponential model – target and modeled values.

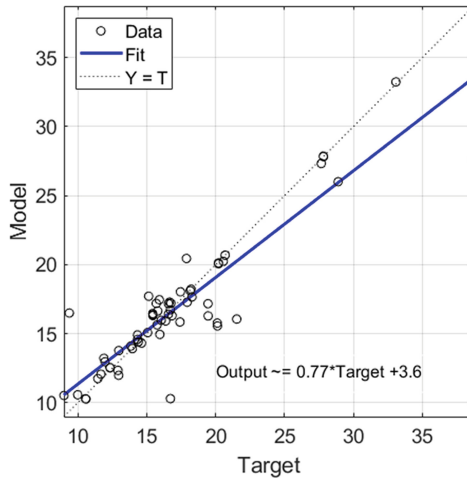


Fig. 12. The GPR ARD Exponential model – the target and modeled values’ regression plot.

Three more models were investigated further:

- a. Variables 2, 3, and 4 are included in the model.
- b. Variables 2 and 3 are included in the model.
- c. Only variable 2 is included in the model.

Table 6 compares alternative sets of input variables to the model of prestressed steel consumption per m² of bridge superstructure. In the table, a binary value (0 or 1) indicates

that a variable is used in the model. By analyzing Table 5, it can be seen that by reducing the number of input variables in a particular scenario, the model can be simplified while also improving its accuracy. In the case of estimating prestressed steel consumption per m^2 in Model 2, improved values for the accuracy criteria RMSE, MAE, MAPE, and R were obtained by eliminating variables x_1 and x_4 from a total of four studied input variables. The values of RMSE and MAE indicate that the model is accurate, with the value of RMSE being significantly higher than the value of MAE, indicating slightly lower accuracy in predicting extreme values of steel consumption. The R and MAPE values also imply that the model is accurate enough.

Table 6. Comparative analysis of the model of prestressed steel consumption per m^2 of the bridge superstructure with alternative sets of input variables

Model	x_1	x_2	x_3	x_4	RMSE	MAE	R	MAPE
1	0	1	1	1	2.8232	1.3881	0.8368	8.2593
2	0	1	1	0	2.1676	1.0993	0.9096	6.5537
3	0	1	0	0	3.0318	1.7853	0.8088	10.0922
4	1	1	1	1	2.3800	1.3302	0.8870	8.1541

6 Conclusions

In order to decide whether or not to begin construction of the PC bridge, an early techno-economic analysis must be conducted during the preparatory phase of the investment project, which must also contain the planned resources. The early planning and optimization of the resources provides a better basis for later easier and less-costly maintenance.

A number of prediction models for early estimation of prestressed steel consumption per m^2 were developed, examined, and verified in this paper, based on artificial intelligence approaches. A database of design characteristics for 75 PC bridges was created for the assessment model. Cross-validation was used to train and test all models under same conditions.

Individual neural network models were used, and the results were unsatisfactory in terms of accuracy across all criteria. Hence, the employment of regression tree ensembles was proposed. The Bagging, Random Forests, and Boosting methods were investigated. For each of the approaches, hyperparameters are defined, as well as a procedure for optimizing them.

A comparison of multiple SVR models revealed that, depending on the kernel function, the models have varying accuracy for alternative criteria. The procedure for parameter optimization has been established. The grid search procedure produced an optimal answer by optimizing parameters through a rough and detailed search. In terms of all criterion functions, the model with RBF kernel function showed much superior accuracy. The results indicate that the RBF kernel would be our recommendation for solving this and similar regression problems.

The application of the Gaussian process to the problem of forecasting prestressing steel consumption is also discussed. The usage of various covariance functions or kernel functions is examined in this context. ARD covariance functions, which apply different length scale parameters to input variables, have proved to be superior to covariance functions with a single distance parameter. The results show that using the ARD exponential covariance function is the optimum solution for the problem.

An accuracy of $RMSE = 2.1676$ was achieved using the ARD exponential covariance function, together with $MAE = 1.0993$, $R = 0.9096$, and $MAPE = 6.55\%$, which is the best result in terms of accuracy of all analyzed models. The application of the ARD covariance function allows the relevance of individual input variables to be considered. The model's achieved accuracy can be considered satisfactory.

Based on the suggested model, a MATLAB-based software was created with a graphical user interface for the input of basic variables, from which estimated prestressing steel quantities can be derived. The use of such software for the estimation of the optimal amount of steel will in a general case lead to the construction of higher-quality structures that will require less maintenance on a long-term basis.

References

1. Marcous, G., Bakhom, M.M., Taha, M.A., El-Said, M.: Preliminary quantity estimate of highway bridges using neural networks. In: Proceedings 6th International Conference on the Application of Artificial Intelligence to Civil and Structural engineering, Stirling, Scotland (2001)
2. Fragkakis, N., Lambropoulos, S., Tsiambaos, G.: Parametric model for conceptual cost estimation of concrete bridge foundations. *J. Infrastruct. Syst.* **17**(2), 66–74 (2011). [https://doi.org/10.1061/\(ASCE\)IS.1943-555X.0000044](https://doi.org/10.1061/(ASCE)IS.1943-555X.0000044)
3. Mučenski, V., Peško, I., Trivunić, M., Dražić, J., Čirović, G.: Optimizacija neuronske mreže za procenu potrebnih količina betona i armature u višespratnim objektima. *Građevinski materijali i konstrukcije* **55**(2), 27–46 (2012)
4. Fragkakis, N., Marinelli, M., Lambropoulos, S.: Preliminary cost estimate model for culverts. *Procedia Eng.* **123**, 153–161 (2015). <https://doi.org/10.1016/j.proeng.2015.10.072>
5. Marinelli, M., Dimitriou, L., Fragkakis, N., Lambropoulos, S.: Non-parametric bill of quantities estimation of concrete road bridges superstructure: an artificial neural networks approach. In: Proceedings 31st Annual ARCOM Conference, Lincoln, United Kingdom (2015)
6. Antoniou, F., Konstantinitis, D., Aretoulis, G.: Cost analysis and material consumption of highway bridge underpasses. In: Eighth International Conference on Construction in the 21st Century (CITC-8), Changing the Field: Recent Developments for the Future of Engineering and Construction, Thessaloniki, Greece (2015)
7. Dimitriou, L., Marinelli, M., Fragkakis, N.: Early bill-of-quantities estimation of concrete road bridges - an artificial intelligence-based application. *Public Work Manage. Policy* **23**(2), 127–149 (2018). <https://doi.org/10.1177/1087724X17737321>
8. Kovačević, M., Ivanišević, N., Dašić, T., Marković, L.: Application of artificial neural networks for hydrological modelling in Karst. *Građevinar* **70**, 1–10 (2018). <https://doi.org/10.14256/JCE.1594.2016>
9. Beale, M.H., Hagan, M.T., Demuth, H.B.: *Neural Network Toolbox*. The Mathworks, Inc (2010)
10. Kovačević, M., Ivanišević, N., Petronijević, P., Despotović, V.: Construction cost estimation of reinforced and prestressed concrete bridges using machine learning. *Građevinar* **73**, 1–13 (2021). <https://doi.org/10.14256/JCE.2738.2019>

11. Kovačević, M., Lozančić, S., Nyarko, E.K., Hadzima-Nyarko, M.: Modeling of compressive strength of self-compacting rubberized concrete using machine learning. *Materials* **14**, 4346 (2021). <https://doi.org/10.3390/ma14154346>
12. Black, P.: Dictionary of Algorithms and Data Structures, NISTIR (1998). <http://www.nist.gov/dads>. Accessed 11 Oct 2021
13. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. The MIT Press, London (2009)
14. Hastie, T., Tibsirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Heidelberg (2009)
15. Breiman, L., Friedman, H., Olsen, R., Stone, C.J.: Classification and Regression Trees. Chapman and Hall/CRC, Wadsworth (1984)
16. Breiman, L.: Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996). <https://doi.org/10.1007/BF00058655>
17. Kovačević, M.: Model for forecasting and assessment of construction cost of reinforced-concrete bridges, Doctoral dissertation, University of Belgrade, Faculty of Civil Engineering, Belgrade, Serbia (2018). <https://doi.org/10.13140/RG.2.2.24025.65129>
18. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
19. Kecman, V.: Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. The MIT Press, Cambridge (2001)
20. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**(3) (2011). <https://doi.org/10.1145/1961189.1961199>
21. LIBSVM-A Library for Support Vector Machines. <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>. Accessed 21 Feb 2021
22. Rasmussen, C.E., Williams, C.K.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)
23. IMS: Sertifikat sistema kvaliteta broj 170101, SPB SUPER Sistem za prenaprezanje. IMS, Centar za prednaprezanje, Beograd (1999)
24. Chou, J.S., Lin, C.W., Pham, A.D., Shao, J.Y.: Optimized artificial intelligence models for predicting project award price. *Autom. Constr.* **54**, 106–115 (2015). <https://doi.org/10.1016/j.autcon.2015.02.006>
25. Hagan, M.T., Demuth, H.B., Beale, M.H., Jesus, O.D.: Neural Network Design. Oklahoma State University, Stillwater (2014)
26. Breiman, L.: Random forests. *Mach. Learn.* **45**, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
27. Radić, J., Šavor, Z., Puž, G.: Tipizacija mostova za autoceste. *Građevinar* **52**(6), 321–330 (2000)