Primljen / Received: 16.6.2019.
Ispravljen / Corrected: 9.12.2020.
Prihvaćen / Accepted: 19.12.2020.
Dostupno online / Available online: 10.2.2021.

Construction cost estimation of reinforced and prestressed concrete bridges using machine learning

Authors:

Assist.Prof. **Miljan Kovačević**, PhD. CE
University of Prishtina
Faculty of Technical Sciences
miljan.kovacevic@pr.ac.rs
Corresponding author

Prof.dr.sc. **Nenad Ivanišević**, PhD. CE
University of Belgrade, Serbia
Faculty of Civil Engineering
nesa@grf.bg.ac.rs

Assist.Prof. **Predrag Petronijević**, PhD. CE
University of Belgrade, Serbia
Faculty of Civil Engineering
pecap@grf.bg.ac.rs

Vladimir Despotović, PhD. EI
University of Luxembourg
Department of Computing
vladimir.despotovic@uni.lu

Original scientific paper

Miljan Kovačević, Nenad Ivanišević, Predrag Petronijević, Vladimir Despotović

Construction cost estimation of reinforced and prestressed concrete bridges using machine learning

Seven state-of-the-art machine learning techniques for estimation of construction costs of reinforced-concrete and prestressed concrete bridges are investigated in this paper, including artificial neural networks (ANN) and ensembles of ANNs, regression tree ensembles (random forests, boosted and bagged regression trees), support vector regression (SVR) method, and Gaussian process regression (GPR). A database of construction costs and design characteristics for 181 reinforced-concrete and prestressed-concrete bridges is created for model training and evaluation.

Key words:
reinforced concrete bridges, prestressed concrete bridges, machine learning, construction costs

Izvorni znanstveni rad

Miljan Kovačević, Nenad Ivanišević, Predrag Petronijević, Vladimir Despotović

Procjena troškova izgradnje AB i prednapetih betonskih mostova primjenom strojnog učenja

U ovom radu istraženo je sedam najnovijih postupaka strojnog učenja za procjenu troškova izgradnje armiranobetonskih i prednapetih betonskih mostova, uključujući umjetne neuronske mreže (ANN) i ansamble ANN, ansamble regresijskih stabala (eng. random forests, boosted and bagged regresijska stabla), metodu potpornih vektora za regresiju (SVR) i Gausov regresijski proces (GPR). Stvorena je i baza podataka o troškovima izgradnje i projektnim karakteristikama za 181 armiranobetonski i prednapeti betonski most za treniranje i ocjenu modela.

Ključne riječi:
armiranobetonski mostovi, prednapeti betonski mostovi, strojno učenje, troškovi izgradnje

Wissenschaftlicher Originalbeitrag

Miljan Kovačević, Nenad Ivanišević, Predrag Petronijević, Vladimir Despotović

Schätzung der Baukosten für Stahlbeton- und Spannbetonbrücken durch maschinelles Lernen

In dieser Arbeit werden sieben kürzlich durchgeführte maschinelle Lernverfahren zur Schätzung der Kosten für den Bau von Stahlbeton- und Spannbetonbrücken untersucht, darunter künstliche neuronale Netze (ANN) und ANN-Ensembles, regressive Baumensembles (Random Forests, Bagging und Boosting bei Regressionsbäumen), die Methode der Support-Vektor-Maschine für Regression (SVM) und der Gaußsche Regressionsprozess (GPR). Außerdem wurde eine Datenbank zu Baukosten und Planungsmerkmalen für 181 Stahlbeton- und Spannbetonbrücken für die Modellschulung und -bewertung erstellt.

Schlüsselwörter:
Stahlbetonbrücken, Spannbetonbrücken, maschinelles Lernen, Baukosten

1. Introduction

There are currently more than two million bridges in operation worldwide, and their number is constantly increasing [1]. According to the American Road and Transportation Builders Association (ARTBA), the total investment costs of bridges in the USA were estimated at US$ 27 billion in 2014 [2]. In the European Union, 20.4 billion euros are planned for the construction of Trans-European Networks (TENS) within the transport sector (Connecting Europe Facility) for the 2014-2020 period [3]. This trend will certainly continue in the oncoming years, hence the estimation of construction costs, which are the most significant part of total investment costs, is of utmost importance [4]. Predicting construction costs is one of the most important preliminary steps in any construction project, since cost prediction is crucial to avoid construction delays and ensure successful project completion [5]. The main problem in estimation of transport infrastructure project costs is significant deviation between the estimated costs and the real, actual construction costs, due to intentional underestimation in the initial project phases, when the costs are evaluated in order to decide whether the transport infrastructure should be built. Based on the analysis of 258 transport infrastructure projects worth $90 billion (U.S.), it was found that in the vast majority of projects actual costs were significantly higher than initially estimated, e.g. 34 % higher on an average for bridges and tunnels [6]. This underestimation is obviously not an error, it is prone to subjectivity, and may potentially introduce biases in the decision making process [6]. Therefore, being able to objectively forecast these costs is highly desirable. The estimation of construction costs in transport infrastructure is a complex process influenced by a variety of factors, uncertainty, and imprecision. Methods based on machine learning have shown promising results, enabling automation of the construction costs estimation process, and eliminating the biases introduced by human factor. Hegazy and Ayed designed an artificial neural network (ANN) model for the assessment of highway construction costs [7]. Backpropagation, simplex optimization and genetic algorithm (GA) were used for network training. The network was trained using a set of eighteen highway projects constructed in Newfoundland, Canada. Marcous et al. used ANN with backpropagation learning algorithm to predict the volume of concrete and the weight of prestressing steel in bridge superstructure [8]. A set of twenty-two prestressed concrete bridges over the Nile in Egypt was used for network training. Marinelli et al. used the feed-forward ANN model to predict the quantities of superstructure material (concrete, prestressed steel, and reinforcing steel) using the project data from 68 highway bridges constructed in Greece [9]. Mostafa used multiple regression analysis to estimate the costs of 54 bridges and 72 culverts [10]. Using the multiple regression analysis, Hollar et al. assessed the costs of preliminary engineering of bridges, determined as a percentage of construction costs [11]. The dataset consists of bridge projects in North Carolina, USA, between 2001 and 2009. Cheng and Wu applied support vector machines (SVM) to predict construction costs using the set of twenty-nine construction projects as training cases, with an average prediction error of less than 10 % [12]. Kim and Kim studied preliminary cost estimations using case-based reasoning (CBR) and GA [13]. Fragkakis et al. presented a prediction model for bridge foundation costs that predicted material quantities for various types of foundations, and estimated the total foundation costs using the backward stepwise regression [14]. Cirilovic et al. studied prediction models based on multiple regression analysis and ANNs for the unit costs of road reconstruction works, using a dataset of 200 contracts from 14 countries in Europe and Central Asia signed between 2000 and 2010 [15]. Pesko et al. conducted a similar research on the estimation of traffic infrastructure reconstruction costs in urban areas using ANNs [16]. Elfaki et al. reviewed methods for estimating construction costs including machine learning, rule-based systems, evolutionary systems, agent-based systems, and hybrid systems [5]. Chou et al. studied models based on multiple regression analysis, CBR and ANNs, to predict bid prices for bridge construction projects in Taiwan [17]. The best prediction was obtained using ANN model, with MAPE - used as performance criterion - amounting to 13.09 %.

It can be concluded from previous research and studies that researchers have used cost estimation models, the advantage of which being that a wider professional community is familiar with such models. The disadvantage is that most researchers use either linear regression models based on the assumption of linearity, which makes the whole estimation process biased, or neural network models that are significantly more complex to interpret (black box models) and require a more extensive database, or use hybrid models which are even more complicated. This paper offers a comprehensive comparative analysis of seven state-of-the-art machine learning techniques for the estimation of construction costs of RC and PC bridges. Some of the proposed models, such as GPR, have not been, to the best of our knowledge, previously used for estimating construction costs of transport infrastructure projects.

2. Methodology

State-of-the-art machine learning techniques for estimating construction costs of RC and PC bridges, including ANNs and ensembles of ANNs, regression tree ensembles, support vector regression, and Gaussian process regression, are briefly outlined in this section, and details of their implementation are given.

2.1. Multilayer perceptron artificial neural network (MLP- ANN)

The multilayer perceptron (Figure 1) is a feed-forward neural network that consists of at least three layers: input layer, hidden layer and output layer.

Each layer is composed of one or more processing units called neurons, where each neuron in one layer is connected to each neuron of the next layer. Multiple neuron layers with nonlinear transfer functions allow the network to learn nonlinear relationships between input and output vectors [18]. MLP with one hidden layer with bipolar sigmoid activation function and an output layer with linear activation function can approximate arbitrary multidimensional function for a given dataset, given sufficient number of neurons in the hidden layer [19].

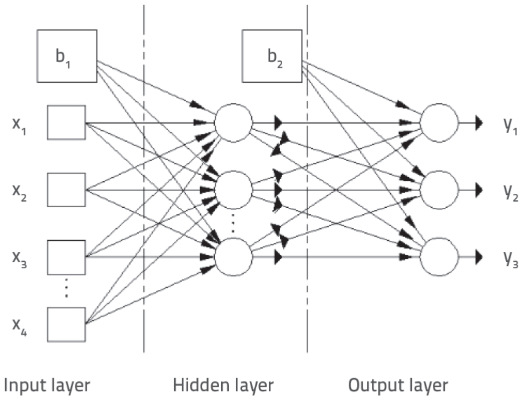

Figure 1. Multilayer perceptron artificial neural network

The number of neurons in the hidden layer can be determined experimentally for the given dataset, with the upper limit calculated by:

$$N_H \le 2\times N_i \tag{1}$$

$$N_H \le \frac{N_S}{N_i+1} \tag{2}$$

where N_i represents the number of inputs in the neural network, and N_s represents the number of instances used for training. It is suggested to accept the lower value of the number of neurons in the hidden layer given by (1) and (2) [20, 21].

Neural network ensembles can be used to improve generalization of ANN, where many neural networks are used together to predict the unseen data. The components that form an ensemble are denoted as base models, or submodels, and each submodel is allowed to have different number of neurons in the hidden layer. MLP neural networks, with early stopping of training to avoid overfitting ,are used as submodels in an ANN ensemble.

2.2. Regression tree ensembles

Linear regression represents a global model, where a single formula describes the relations between the inputs and the outputs of the model over the entire data space. It is very hard to design a single global model when there are many features interacting in nonlinear ways. An alternative approach is to divide the data space into smaller partitions, where the modelling of these interactions is easier to achieve. These partitions can be further divided into even smaller regions, until finally one gets the data space cells where simple models can be applied. This is called recursive partitioning.

Regression trees use the tree to represent the recursive partition. It splits the input data space in partitions and assigns a prediction value to each partition. The terminal nodes of the tree, denoted as leaves, represent these partition cells. In order to determine to which leaf the input data belongs, and to assign it the prediction value, the algorithm starts from the root node and asks successive binary questions. Depending on the outcome of the question, the sub-branch of the tree is chosen. Eventually, the algorithm arrives at the leaf node, where the prediction is made. This prediction is found as an average of all training data instances which reach that leaf node.

Suppose a training dataset D = $\{(x_i,y_i) \in \mathbb{R}^n \times \mathbb{R}, i = 1, 2, \ldots, l\}$ which consists of l training pairs (x_1,y_1), (x_2,y_2), ..., (x_l,y_l) , where $x_i \in \mathbb{R}^n$ is the n-dimenzional vector denoting model's inputs and y_i are the observed responses to these inputs (model's outputs). Suppose further a division of the input data space into M partitions R_m, m = 1, 2, ..., M, , where the response is modelled as a constant in each partition:

$$f(x)=\sum_{m=1}^{M} c_m I\left(x\in R_m\right) \tag{3}$$

where $I\{x \in R_i\}$ is a binary function that takes the value 0 or 1 depending on the outcome of the question at the tree split point [22]. Constant c_i can be determined as an average of responses y_i in the region R_i. Greedy algorithm is used in order to determine the split point [23, 24]. Regression trees can be combined in an ensemble, which represents a predictive model composed of a weighted combination of multiple regression trees. Various algorithms can be used for ensemble learning, such as, for instance, bagging and boosting.

2.2.1. Bagging

A major problem with regression trees is high variance, which occurs due to the fact that only a minor change in the data can cause significantly different tree structures. This happens because the error in one of the top splits propagates all the way down to the leaves. In bootstrap aggregation, or bagging, multiple data subsets D_i are created from the training dataset D, by sampling from randomly and with replacement [25]. Each of these subsets is called a bootstrap sample (or simply bootstrap). Since replacement is allowed, the bootstraps might have duplicated data instances, or some of them may be omitted, resulting in bootstraps different from the initial dataset. Each of these bootstraps is used to build a single regression tree, which might have a different number of leaves and different structure in comparison to the original tree. All individual trees are further combined in an ensemble (see Figure 2). The predictions are averaged over all trees in the ensemble, thus decreasing the variance and improving prediction.

2.2.2. Random Forests

Random forests represent an extension of bagging that reduces the correlation between the individual trees, thus building an ensemble of decorrelated trees.

Suppose that training dataset D is composed of *l* observations and *n* features. First, a sample from the training dataset is taken randomly with replacement and bootstrap is created. Before each split, $m \leq n$ features are randomly selected as candidates for splitting. The best feature (split point) among features is used to split the node iteratively [22, 26]. Single tree is grown for each bootstrap and predictions are averaged over all trees in the forest. Typical values for *m* are approximately $\sqrt{n}$ [20, 24]. Reducing *m* reduces the correlation between any pair of trees in the ensemble, thus reducing the variance of the average.

While in bagging the random data subsets are sampled from the initial dataset for each tree, in random forests, in addition to this, the feature subsets are also randomly selected, instead of using all features to grow the trees. Many random trees form random forests.

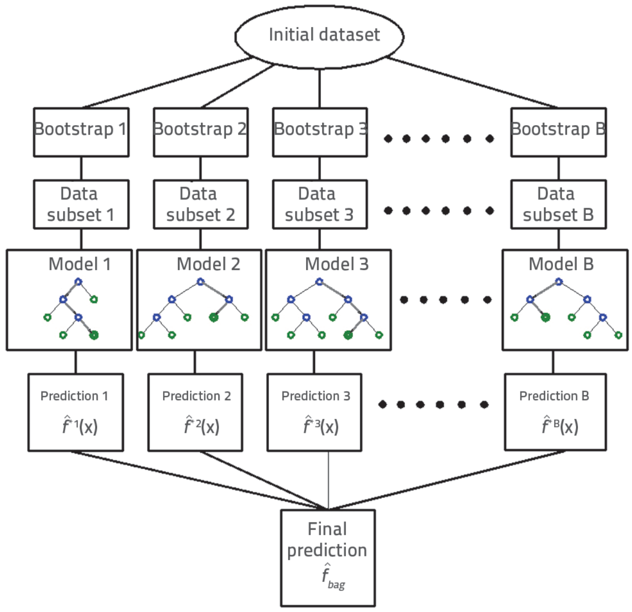

Figure 2. Bootstrap aggregation (bagging) in regression tree ensembles

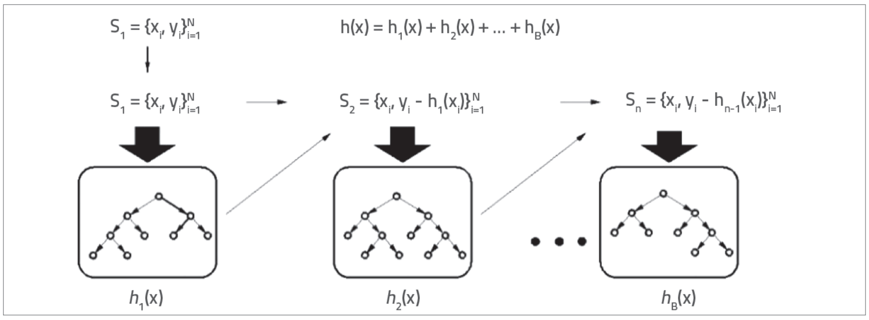
$S_1 = \{x_i, y_i\}_{i=1}^N$
$h(x) = h_1(x) + h_2(x) + \ldots + h_B(x)$
$S_1 = \{x_i, y_i\}_{i=1}^N$
$S_2 = \{x_i, y_i - h_1(x_i)\}_{i=1}^N$
$S_n = \{x_i, y_i - h_{n-1}(x_i)\}_{i=1}^N$
$h_1(x)$
$h_2(x)$
$h_B(x)$

Figure 3. Gradient boosting in regression tree ensembles

2.2.3. Boosting

Boosting is an ensemble technique where predictors are created sequentially, rather than independently, as in bagging. The rationale behind this is that each subsequent predictor learns from the mistakes committed by previous predictors [22]. When gradient boosting is applied to regression tree ensembles, the first regression tree is the one that maximally reduces the loss function for the selected tree structure and the given training dataset. The residual (prediction error) is then calculated. It represents the mistake committed by the predictor model (the first regression tree). In the next step, a new tree is fitted to the residuals of the first tree. In each step, a new tree is added to the model, which is fitted to the residuals of the previous one. The residual values are usually multiplied by the learning rate (value less than 1) to avoid overfitting. The final model obtained by boosting is simply a linear combination of all trees (usually hundreds or thousands of trees), as shown in Figure 3.

The main idea of boosting is that, instead of using a complex single regression tree, which is easily overfitted, a much better fit is produced if many simple regression trees are trained iteratively, each of them improving the prediction performance of the previous ones [22].

2.3. Support vector regression (SVR)

Suppose a training dataset $\{(x_1,y_1), (x_2,y_2), \ldots, (x_l,y_l) \in \mathbb{R}^n \times \mathbb{R}\}$ is given, where $x_i \in \mathbb{R}^n$ is the n-dimenzional vector denoting model's inputs and y_i are the observed responses to these inputs (model outputs). SVR tries to find an approximating function $f(x)$ with deviation ε from the observed response y_i for all training data *x*. This approximating function for the nonlinear SVR [27] equals to

$$f(x) = \sum_{i=1}^{l} \left(\alpha_i^* - \alpha_i\right) \mathbf{K}\left(\mathbf{x_i}, \mathbf{x}\right) + b \qquad (4)$$

In Eq. (4) **K** denotes the kernel function, α_i^*, α_i and *b* are the parameters derived by the objective function minimization

$$\|w\|^2 + C\left(\sum_{i=1}^{l}\xi_i + \sum_{i=1}^{l}\xi_i^*\right)$$

for the given constraints (Figure 4.). ξ_i and ξ_i^* denote the slack variables which allow the regression errors to cope to a certain extent with otherwise infeasible constraints of the optimization problem.

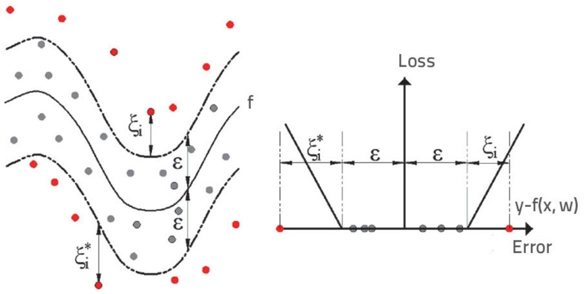

Figure 4. Nonlinear SVR

The constant $C > 0$ is the parameter chosen by user that denotes the amount od deviation larger than ε that can be tolerated. An increase in C penalizes larger errors. Another parameter chosen by the user is the required precision ε. RBF kernel function is used in this paper [28].

$$K\left(x_i, x\right) = \exp\left(-\gamma\left\|x_i - x\right\|^2\right),\ \gamma > 0 \tag{5}$$

2.4. Gaussian process regression

The GP method represents a non-parametric method that is defined as an infinite set of random variables such that every finite subset follows a multivariate Gaussian distribution. By expanding multivariate Gaussian distribution to an infinite set of random variables, it is possible to observe GP as the posterior distribution over random functions, while the Bayes' rule is applied to determine the probability distribution from the training data in a supervised machine learning setup. Consider a problem of nonlinear regression:

$$y = f(x) + \varepsilon,\ \varepsilon \sim N(0,\sigma^2) \tag{6}$$

Where the function $f(\cdot) : \mathrm{R}_n \rightarrow \mathrm{R}$ is an unknown and needs to be estimated, y_i is the target variable, x are input variables and ε is the normally distributed additive noise. The Gaussian process regression [29] assumes that $f(\cdot)$ follow Gaussian process with mean function $\mu(\cdot)$ and covariance function $k(\cdot\ \cdot)$. The n observations in an arbitrary data set $\mathbf{y} = \{y_1, \ldots, y_n\}$, can always be imagined as a sample from some multivariate (n variate) Gaussian distribution

$$(y_1, \ldots, y_n)^T \sim N(\boldsymbol{\mu}, K) \tag{7}$$

where $\boldsymbol{\mu} = (\mu(x_1), \ldots, \mu(x_n))^T$ is the mean vector, and K is n × n covariance matrix of which the $(i, j)^{\text{th}}$ element $K_{ij} = k(x_i, x_j) + \sigma^2\delta_{ij}$. Here δ_{ij} is the Kronecker delta function. Let x^* be any test point and y^* be the corresponding response value. The joint distribution of $(y_1, \ldots, y_n, y^*)$ is an $(n + 1)$ variate normal distribution $(y_1, \ldots, y_n, y^*) \sim \mathrm{N}(\mu^*, \Sigma)$, where $\boldsymbol{\mu}^* = (\mu(x_1), \ldots, \mu(x_n), \mu(x^*))^T$, and the covariance matrix is:

$$\Sigma = \begin{bmatrix} K_{11} & K_{12} & \cdots & K_{1n} & K_{1*} \\ K_{21} & K_{22} & \cdots & K_{2n} & K_{2*} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ K_{n1} & K_{n2} & \cdots & K_{nn} & K_{n*} \\ K_{*1} & K_{*2} & \cdots & K_{*n} & K_{**} \end{bmatrix} = \begin{bmatrix} K & K^* \\ K^{*T} & K^{**} \end{bmatrix} \tag{8}$$

where $K^* = (k(x^*, x_1), \ldots, k(x^*, x_n))^T$ and $K^{**} = k(x^*, x^*)$.
The conditional distribution of y^*, given $\mathbf{y} = (y_1, \ldots, y_n)^T$ is then $N\left(\hat{y}^*, \hat{\sigma}^{*2}\right)$ with

$$\hat{y}^* = \mu\left(x^*\right) + K^{*T}K^{-1}\left(y - \mu\right) \tag{9}$$

$$\hat{\sigma}^{*2} = K^{**} + \sigma^2 - K^{*T}K^{-1}K^* \tag{10}$$

The covariance is a crucial part of the model specification. Various covariance functions are used in the experiments. Each of these covariance functions depends on hyperparameters whose values also need to be tuned. For some covariance functions, hyperparameters can be used to determine which inputs are more relevant than others, using the automatic relevance determination (ARD). For example, consider squared exponential covariance function with different length scale parameters for each input (ARD SE)

$$k\left(x_p, x_q\right) = v^2 exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_p^i - x_q^i}{r_i}\right)^2\right] \tag{11}$$

where r_i denotes the length scale of the covariance function along the input dimension i. If r_i is very large, relative importance of the i-the input is smaller [29]. The hyperparameters $\{v, r_1, \ldots, r_n\}$ and the noise variance σ^2 can be estimated by the maximum likelihood method. The log-likelihood of the training data is given by (12):

$$L\left(v, r_1, \ldots, r_n, \sigma^2\right) = -\frac{1}{2}\log\det K - \frac{1}{2}y^T K^{-1} y - \frac{n}{2}\log 2\pi \tag{12}$$

3. Dataset

The proposed cost estimate methods rely on the development of a dataset that includes project and contract documentation of RC and PC bridges constructed at the Corridor X, which is one of the most important Pan-European transport corridors connecting Austria, Hungary, Slovenia, Croatia, Serbia, Bulgaria, Republic of North Macedonia, and Greece (Figure 5).
The current bridge dataset includes complete data on 181 constructed highway bridges, including 104 bridges with cast

in situ RC superstructure and 77 bridges with PC superstructure (prefabricated or cast in situ), located at the eastern and southern legs of Corridor X in Serbia. Out of the total 181 bridges, 148 are bridges carrying the motorway, and 33 are overpasses not carrying the motorway. The total contract value of all bridges included in the dataset is over EUR 100 million.

Analysing the project costs in the created dataset, it can be concluded that 77.41 % of all construction costs were the costs that are related to steel and concrete. Contracts were signed for all bridges between September 2009 and June 2014. The cost of labour, material, and quarried aggregate increased significantly over that period. Average gross salaries increased by approximately 30 %, quarried aggregates price increased by 16 %, and the price of steel increased by 27 % (maximum values compared to September 2009).

The literature review [7-17] shows that a significant number of models start from a particular assumption about the model. In this paper, an attempt is made to obtain the model from the experimental data without making any prior assumptions about the model, using a narrow set of data that are available at preliminary stages of project development.

Bridge design is generally affected by many variables; hence selecting the input variables plays a crucial role in modelling construction costs of RC and PC bridges. As concrete and metal works are the most cost intensive (accounting for, on an average, almost 80 % of all costs), variables that are directly related to the amount of concrete works and the amount of metal works are adopted as input variables of the model. In this regard, the following variables were considered: *Total bridge span length*, *Bridge width*, *Average pier height, Foundation type*.

In cases when the *Total bridge span length* is the same, regardless of whether it is composed of a small number of large individual spans or a large number of short individual spans, a new variable *Average bridge span* was introduced, which better characterizes the bridge length. It can be obtained by dividing the *Total bridge span length* with the number of bridge spans.

According to [30], the costs related to formwork and scaffolding can amount to up to 20 % of the total construction costs. In order to consider the potential impact of these costs, a variable *Type of bridge construction* was introduced in this paper.

The variables *Gross salary*, *Quarried aggregate price index* and *Steel price index* allow comparison of construction costs of bridges that have been contracted with a different base date.

Table 1. Input variables used for modelling construction costs of RC and PC bridges

Variable ID	Variable name
x_1	Average bridge span
x_2	Total bridge span length
x_3	Bridge width
x_4	Type of bridge construction
x_5	Average pier height
x_6	Foundation type
x_7	Gross salary*
x_8	Quarried aggregate price index*
x_9	Steel price index*

*Statistical Office of the Republic of Serbia (http://www.stat.gov.rs/)

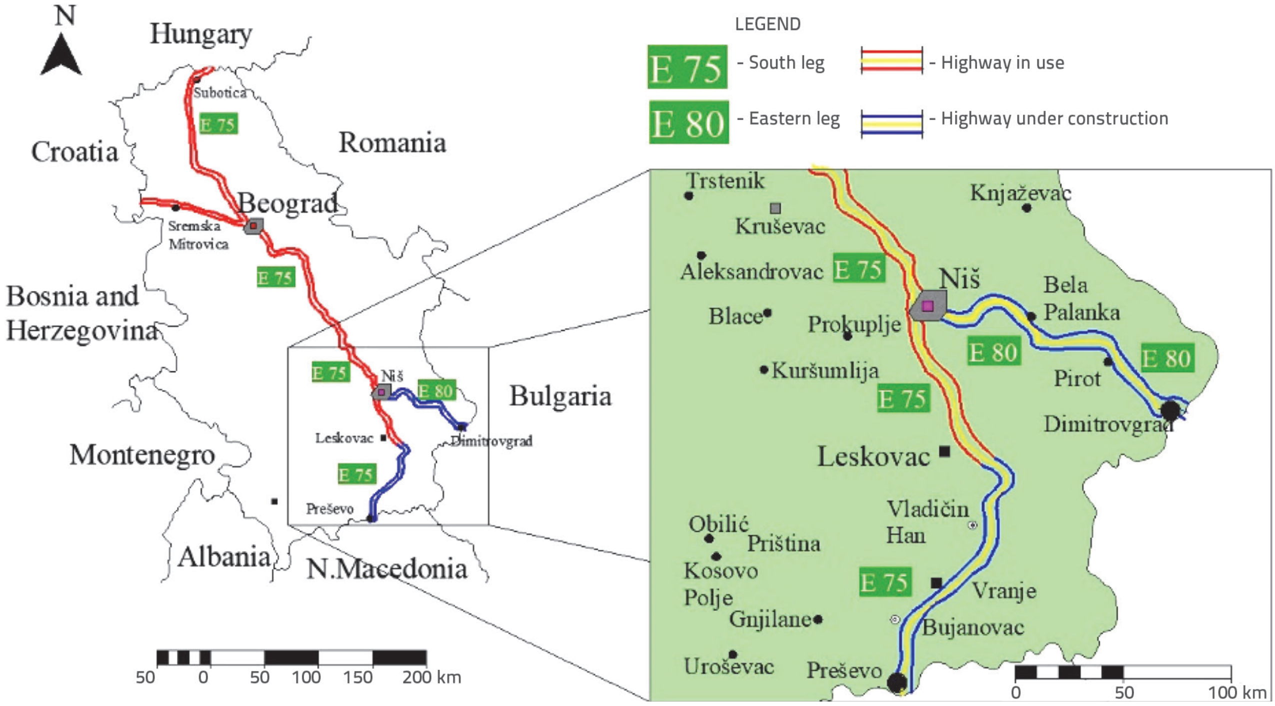

Figure 5. Eastern and southern legs of Corridor X in Serbia

Table 2. Average, minimum and maximum values of input and output variables used for modelling construction costs of RC and PC bridges

Variable	Average value	Minimum value	Maximum value
Average bridge span [m]	21.25	6.52	49.00
Total bridge span length [m]	84.24	6.52	628.74
Bridge width [m]	13.43	7.90	19.91
Average pier height [m]	9.60	3.28	35.01
Gross salary	44608	38427	51248
Quarried aggregate price index	109.85	100	115.99
Steel price index	123.37	100	127.58
Construction costs [EUR/m^2]	593.65	310.83	1335.39

Nine input variables used for all models are shown in Table 1. The variable "Type of bridge construction" is binary, and takes value 1 for PC span superstructure (prefabricated or cast in situ), or value 0 for cast in situ RC span superstructure. The variable "Foundation type" is also binary, and takes value 1 for deep foundations, or value 0 for shallow foundations. Quarried aggregate and steel prices were converted to indexes, using prices applicable in September 2009 as base indexes. The variable "Gross salary" denotes an average gross salary in construction industry defined in contract documentation (note that gross salaries vary significantly from September 2009 to June 2014).

Average, minimum and maximum values of input variables (excluding binary variables) are given in Table 2. The binary variable *Type of bridge structure* has a value of 1 for 77 bridges with PC superstructure (prefabricated or cast in situ) and 0 for 104 bridges with cast in situ RC superstructure. The binary variable *Foundation type* has a value of 1 for 136 bridges and 0 for 45 bridges. The variable *Construction cost* is the estimated output variable given in €/m^2.

4. Evaluation and performance measures

In this study, the performance assessment of models was done using both absolute and relative statistical performance criteria, as suggested by Legates and McCabe [31]. he considered statistical measures were root mean square error (RMSE) and mean absolute error (MAE) as absolute measures, and Pearson's linear correlation coefficient (R) and the mean absolute percentage error (MAPE) as relative measures.

RMSE is the measure of differences between values predicted by the model o_k and the actually observed (measured) values d_k. It is the measure of general accuracy of the model.

$$RMSE=\sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(d_k-o_k\right)^2} \tag{13}$$

MAE is used to represent the mean absolute error of the model according to equation:

$$MAE=\frac{1}{N}\sum_{k=1}^{N}\left|d_k-o_k\right| \tag{14}$$

R is a measure of linear correlation between values predicted by the model o_k and the actually observed (measured) values d_k:

$$R=\sqrt{\left[\sum_{k=1}^{N}(d_k-\bar{d})(o_k-\bar{o})\right]^2\cdot\left[\sum_{k=1}^{N}\left(d_k-\bar{d}\right)^2\left(o_k-\bar{o}\right)^2\right]^{-1}} \tag{15}$$

where $\bar{d}$ represents the mean of d_k and $\bar{o}$ represents the mean of o_k, $k = 1, 2, ..., N$, and N is the number of instances in the dataset.

MAPE is a percentage-based measure of prediction accuracy. It is calculated as an average of the absolute percentage error.

$$MAPE=\frac{100}{N}\sum_{k=1}^{N}\left|\frac{d_k-o_k}{d_k}\right| \tag{16}$$

The machine learning methods used in this paper were evaluated using 10-fold cross-validation, where the dataset is randomly partitioned into 10 subsets, 9 of them being used for training the model and the remaining one for model validation (testing). The cross-validation procedure is repeated 10 times, with each of the subsets used exactly once for validation, and 10 obtained results are then averaged to produce a single estimation.

5. Results and discussion

Several state-of-the-art machine learning techniques for estimation of construction costs of RC and PC bridges are compared in this section, including ANNs and ensembles of ANNs, regression tree ensembles (random forests, boosted and bagged regression trees), SVR, and GPR. The results are obtained using the dataset developed for this purpose, containing construction costs and project characteristics for 181 RC and PC bridges at the Pan-European Corridor X. Root mean square error (RMSE), mean absolute error (MAE), Pearson's linear correlation coefficient (R), and mean absolute percentage error (MAPE) were used as performance measures. For all machine learning methods, all input variables were used as features for modelling construction costs, as shown in Table 2. Construction cost expressed in €/m^2 is the output variable that needs to be predicted.

MLP-ANN with one hidden layer was trained using the Levenberg-Marquardt algorithm [32].The criterion to stop the training was either the maximum number of epochs (set to 1000), the minimum gradient magnitude (set to 10^{-5}) or the network performance (measured as the mean square error and set to 0). All input data are normalized in the range [-1,1] prior to training. The number of neurons in the input layer is determined by the number of input variables, i.e. it consists of 9 neurons, while there is only one neuron in the output layer. The maximum number of neurons in the hidden layer was determined experimentally using Eq. 1 and 2 and equals 18.

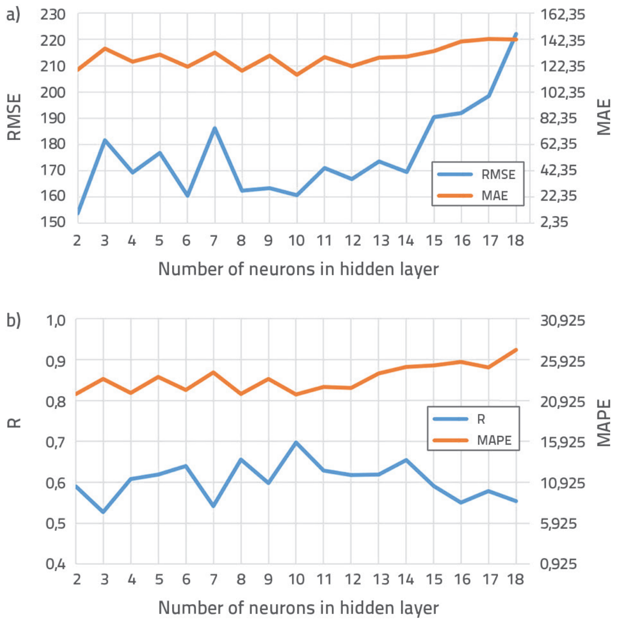

Figure 6. Comparison of performance measures for estimating construction costs using MLP-ANNs with different configurations: a) RMSE and MAE, b) R and MAPE

Figure 6a shows the performance obtained using RMSE and MAE as absolute measures, while Figure 6b presents results using R and MAPE as relative measures. The best performing model using MAE, R and MAPE as performance measures, is MLP-ANN with 10 neurons in the hidden layer. In order to further improve model performance, the ensembles of MLP-ANNs with early stopping were analysed, with base models having up to 18 neurons in the hidden layer. Each base model is allowed to have different number of neurons in the hidden layer. Optimal base models that form an ensemble are determined based on the minimum RMSE. Ensembles with 1 and up to 100 base models were tested, as shown in Figure 7. Learning curves presenting RMSE and MAE vs. number of base models in the ensemble (see Figure 7a) show that performance improves as the number of base models increases, but the curves saturate at approximately 40 base models; hence there is no point to further add base models in the ensemble, as this would increase the model complexity without significant improvement in performance. Similar behaviour can be observed in Figure 7b, where R and MAPE are used as performance criteria.

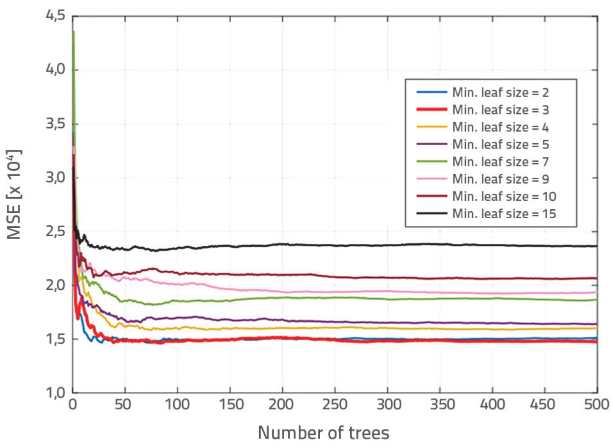

Figure 8. MSE vs. number of trees in the ensemble for different minimum leaf sizes using regression tree ensembles realized with bootstrap aggregation (bagging)

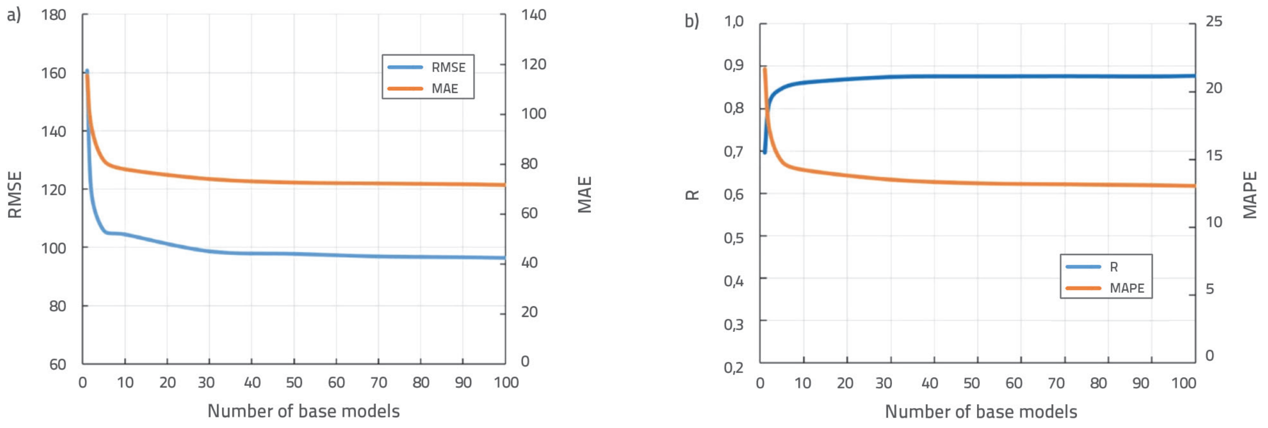

Figure 7. Comparison of performance measures for estimating construction costs using ensembles of MLP-ANNs with different number of base models: a) RMSE and MAE, b) R and MAPE

Table 3. Performance of GPR with various covariance functions for prediction of construction costs of RC and PC bridges using RMSE, MAE, R and MAPE as performance measures

Criteria Covariance function	RMSE	MAE	R	MAPE
Exponential	117.91	75.89	0.83	13.97
ARD Exponential	**95.98**	**63.25**	**0.89**	**11.60**
Squared Exponential	121.85	80.02	0.82	14.94
ARD-Squared Exponential	108.75	69.23	0.86	12.25
Matern 3/2	119.99	77.25	0.83	14.27
ARD-MATERN 3/2	105.46	67.72	0.87	12.62
Matern 5/2	120.36	78.03	0.82	14.47
ARD-Matern 5/2	99.40	64.81	0.88	11.83
Rational Quadratic	118.54	76.42	0.83	14.09
ARD Rational Quadratic	122.68	76.88	0.82	14.22

Regression tree ensembles realized using bootstrap aggregation (bagging) are optimized for different model parameters, including number of trees in an ensemble that is limited to 500 and to minimum leaf size ranging from 2 to 15. Grid search is used for optimization. Learning curves presenting MSE vs. number of trees in the ensemble for different minimum leaf sizes are shown in Figure 8. Minimum leaf size of 2 and 3 gives the best performance measured by MSE. There is no need to use more than 50 trees in the ensemble, as no improvement is observed with further increase in the number of trees.

Random Forests are analysed for different model parameters, including number of trees in an ensemble limited to 500, minimum leaf size ranging from 2 to 10, and the number of randomly selected features as candidates for splitting. The rule of thumb is that $m = n/3$ features should be used as candidates for splitting for regression problems [24]. Values of $m = 2$, $m = 3$ and m = 4 are considered in this paper. Grid search is used for optimization.

Regression tree ensembles realized using boosting are optimized for different model parameters, including number of trees in an ensemble, learning rate, number of splits and number of observations per parent node. Learning rate determines the training speed. Learning rates equal to 0.001; 0.01; 0.1; 0.5; 0.75 and 1.0 are analysed in this paper. Number of splits is exponentially increased, starting from 2^0 = 1 to 2^7 = 128. Number of observations per parent node changes between 5 and 20. Optimal model is obtained using 64 splits and 11 observations per parent node. Grid search is used for optimization. Learning curves presenting MSE vs. number of trees in the ensemble for different learning rates are shown in Figure 9. The learning rate equal to 0.1 gives the best performance measured by MSE. There is no need to use more than 30 trees in the ensemble, as no improvement is observed with further increase in the number of trees.

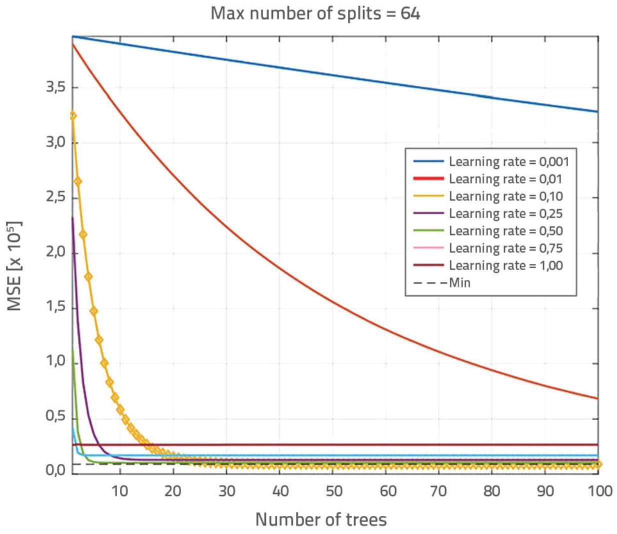

Figure 9. MSE vs. number of trees in the ensemble for different learning rates using regression tree ensembles realized with boosting (max. 64 splits and 11 observations per parent node)

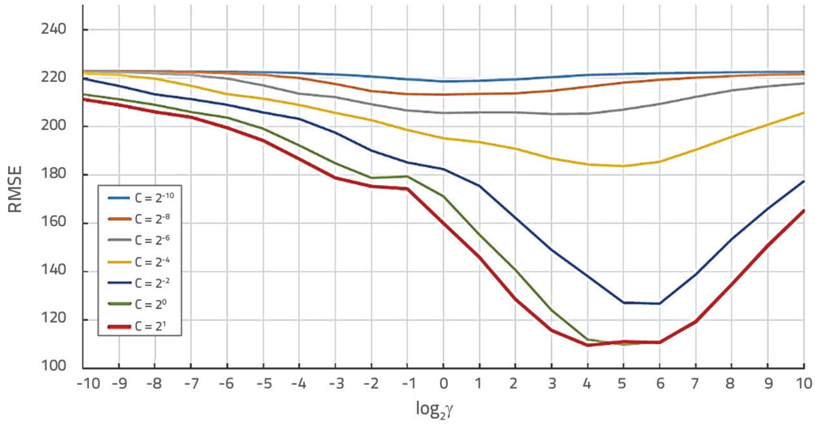

Figure 10. RMSE vs. hyperparameters C i γ for $\varepsilon = 2^{-6}$ using SVR with RBF kernel

Table 4. Performance of machine learning methods for prediction of construction costs of RC and PC bridges using RMSE, MAE, R and MAPE as performance measures

Criteria Model	RMSE	MAE	R	MAPE
MLP-ANN-9-10-1	160.75	115.48	0.7	21.66
MLP-ANN ensemble	96.45	71.71	0.88	13.04
Bagging	121.50	88.72	0.80	15.76
Random Forest	129.05	93.53	0.79	16.58
Gradient Boosting	96.03	67.15	**0.89**	12.03
SVR-RBF	109.32	68.25	0.86	12.03
GPR ARD-Exponential	**95.98**	**63.25**	**0.89**	**11.60**

SVR was analysed using the RBF kernel function. The LIBSVM library was used for SVR implementation [33]. The normalization, which scales all input data into the range [0,1], was done prior to training and testing. The model hyperparameters C, γ and ε were first roughly tuned using grid search, as shown in Figure 10. The SVR model was then fine-tuned into a more accurate position by iteratively narrowing down the search area, leading to optimum hyperparameters C = 1.7271, γ = 18.7334 and ε = 0.0157. The number of iterations is limited to 100.

A Gaussian process is completely defined by its mean and covariance function, and so GPR algorithm is tested using different covariance functions, such as exponential, squared exponential, Matern, and rational quadratic, as well as their equivalent ARD covariance functions that have a separate length scale for each input variable (see Table 3). All inputs and targets are normalized to have zero mean and unit variance. The mean of the Gaussian process is set to zero and the covariance function parameters are determined by maximizing the log marginal likelihood.

Table 4 shows summarized results of the prediction of construction costs of RC and PC bridges using all machine learning algorithms considered in this paper. The best performing model is highlighted. The worse prediction is obtained using single MLP-ANN, which is expected as most of the competing models are ensemble based models. On the other hand, the ensemble of MLP-ANNs has one of the best performances according to both absolute and relative performance measures. Figure 7 shows that at least 10 base models are needed to achieve sufficient generalization; however, an ensemble with 100 base models is adopted as the representative and used in further experiments.

Regression tree ensembles using bagging, as well as random forests, have shown to be relatively poor predictors for the given dataset, unlike regression tree ensembles using boosting which perform substantially better. SVR using RBF kernel have shown a solid performance with R = 0.86 and MAPE = 12.03 %. We also tested linear and sigmoid kernels, but the prediction was poor. Finally, the best prediction performance was obtained using GPR with ARD-exponential covariance functions, with R = 0.89 and MAPE = 11.60 %.

The additional benefit of GPR models is the fact that their training time is significantly lower in comparison to any of the ensemble methods. As GPR with ARD exponential covariance function performs best according to all performance measures, it will be used for future comparisons.

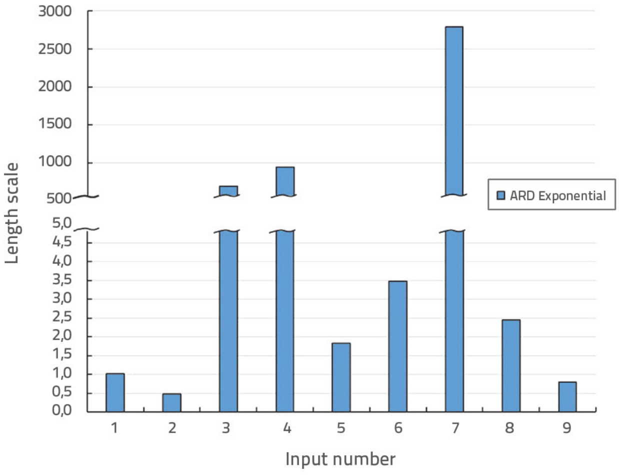

Figure 11. Feature selection using ARD exponential covariance function

Parameters of ARD covariance functions can be used to decide which inputs (features) are relevant for predicting a particular output, and removing less relevant inputs. The analysis of inputs relevance using ARD exponential covariance function is shown in Figure 11 using the length-scale of the covariance function hyperparameters as the criterion. As the values of the length-scale hyperparameter are higher, the particular input becomes less relevant. Note that the inputs 3 (bridge width), 4 (type of bridge construction) and 7 (gross salary) have significantly higher values of the length-scale parameter; therefore, they can be considered less relevant.

This can be explained by the fact that the quarried aggregate price is dependent on the gross salary, and might carry more informative information than the gross salary itself. Hence, gross salary is implicitly represented by the quarried aggregate price. Regarding the bridge width, the output variable

Table 5. Prediction of construction costs of RC and PC bridges using GPR with ARD exponential covariance function

Model	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	RMSE	MAE	R	MAPE
1.	1	1	0	0	1	1	0	1	1	93.55	61.62	**0.90**	11.38
2.	1	1	1	0	1	1	0	1	1	**92.51**	**59.59**	**0.90**	**10.86**
3.	1	1	1	1	1	1	0	1	1	95.21	62.63	0.89	11.53
4.	1	1	1	1	1	1	1	1	1	95.58	63.25	0.89	11.60

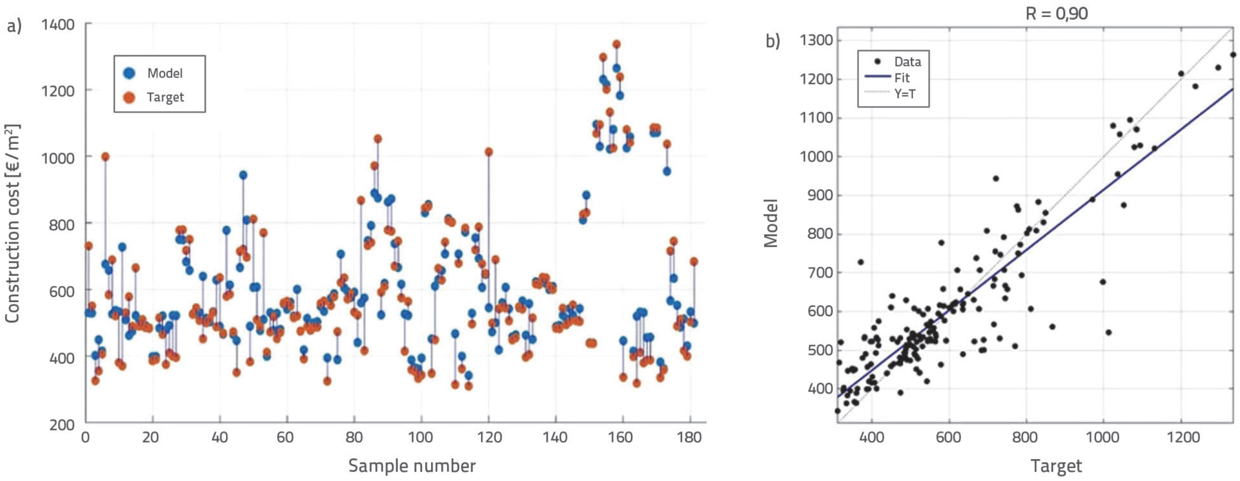

Figure 12. a) Modelled and targeted values for an optimal model with ARD exponential function, b) Regression plot of modelled and targeted values

"Construction costs" is defined in EUR per square meter of the bridge superstructure, which might influence the lower relevance of the bridge width as a feature. The numerical values of the variable width of the bridges in the considered dataset are within a narrower range, as the bridges carrying the motorway dominate (148), while the number of overpasses is smaller (33). This is one of the reasons why the variable width of bridges has less relevance to the model. By expanding the dataset in future research with a significant number of bridges of different widths, the influence of this variable could be determined more accurately. A minor relevance of the variable *Type of bridge structure* may be explained by slightly higher costs of making a PC span superstructure, although there are potential savings in the assembly work. The construction of RC bridges is cheaper, but scaffolding is more expensive. In both cases, the impact of the project implementation time frame on costs was not analysed. No significant difference between these two construction methods was observed using the proposed model.

Table 5 presents results obtained for different combinations of inputs (features) used for modelling. The binary value 0 or 1 denotes whether a particular feature is omitted from the model or not. Note that all the models with reduced number of features outperform the one with the full set of features. The benefit is not only in performance gain, but also in smaller complexity and faster training of the model.

The best performing model is model 2 in Table 5 (see regression plot of modelled and targeted values in Figure 12), which depends on the following input variables: average bridge span, total bridge span length, bridge width, average pier height, foundation type, quarried aggregate price index, steel price index. The performance improves in comparison to the model with the full set of features by 0.8 %, measured by MAPE, leading to MAPE equal to 10.86 %. The improvement is also observed for all other performance measures.

6. Conclusions

In order to make a decision about the need to build transport infrastructure that includes RC or PC bridges, it is necessary to estimate the cost of construction as accurately as possible in the early phase of project implementation. The estimation of construction costs of RC or PC bridges is a complex process that is influenced by a variety of factors. This paper gives a comprehensive overview of the state-of-the-art machine learning methods that can be used for estimating these costs, including MLP-ANN, ensembles of MLP-ANNs, regression tree ensembles (random forests, boosted and bagged regression trees), SVR with RBF kernel, and GPR with exponential, squared exponential, Matern, and rational quadratic covariance functions.

In order to train and assess the models, a dataset was created that includes project and contract documentation for 181 RC and PC bridges constructed on Pan-European Corridor X. All models were trained and tested under equal conditions using the 10-

fold cross validation. According to the relevant performance measures, most of the tested models are able to capture very well complex interrelations between the input features, and demonstrate strong generalization capability. Although ensemble methods, such as ensembles of ANNs, regression tree ensembles using boosting and SVR with RBF kernel, perform well, they require a considerable amount of time to train the models, especially if the number of base models in the ensemble is high. On the other hand, the complexity of models based on Gaussian processes is substantially lower, but they are still able to outperform the ensemble models. Moreover, feature reduction is easy to combine with Gaussian process regression using ARD, leading to models with better performance and even lower complexity. Two out of nine input features can be reduced without any negative influence on model performance. To the best of our knowledge, no results were previously reported for implementation of Gaussian process regression in estimating construction costs.

The research carried out in this paper has confirmed that methods based on machine learning eliminate the biases introduced by human factor, and offer a fast and reliable tool for the construction industry to estimate construction costs of concrete bridges, even in early implementation stages, when only the basic technical and economic characteristics are available. Further research might be aimed at improving the dataset used for model training and evaluation by including additional relevant data about both the existing and new bridges. The problem of estimating construction costs is considered as a regression problem. However, it can be also observed as a classification problem if the costs are divided into groups. In that case, classification algorithms can be applied. The developed models can also be applied, with some modifications, to other costs during the project life cycle.

LITERATURA

[1] Pržulj, M.: Mostovi, Udruženje "Izgradnja", Beograd, 2014.

[2] Antoniou, F., Konsantinidis, D., Aretoulis, G.: Analytical formulation for early cost estimation and material consumption of road overpass bridges, Research Journal of Applied Sciences, Engineering and Tehnology, 12 (2016) 7, pp. 716-725.

[3] Financing the Trans-european networks, https://www.europarl.europa.eu/factsheets/en/sheet/136/financing-the-trans-european-networks

[4] Locatelli, G., Mikic, M., Kovacevic, M., Brookes, N., Ivanisevic, N.: The Successful Delivery of Megaprojects: A Novel Research Method, Project Management Journal, 48 (2017) 5, pp. 78-94.

[5] Elfaki, A.O., Alatawi, S., Abushandi, E.: Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey, Advances in Civil Engineering, 2014.

[6] Flyvbjerg, B., Skamris, H., Buhl, S.: Underestimating Costs in Public Works Projects: Error or Lie?, Journal of the American Planning Association, 68 (2002) 3, pp. 279-295.

[7] Hegazy, T., Ayed, A.: Neural Network Model for Parametric Cost Estimation of Highway Projects, Journal of Construction Engineering and Management, 124 (1998) 3 , pp. 210-218.

[8] Marcous, G., Bakhoum, M.M., Taha, M.A., El-Said, M.: Preliminary quantity estimate of highway bridges using neural networks, Procedings of the Sixth International Conference on the Application of Artificial Inteligence to Civil and Structural engineering, Stirling, Scotland, 2001.

[9] Marinelli, M., Dimitriou, L., Fragkakis, N., Lambropoulos, S.: Non-parametric bill of quantities estimation of concrete road bridges' superstructure: an artificial neural networks approach, Proceedings 31st Annual ARCOM Conference, Lincoln, United Kingdom, 2015.

[10] Mostafa, E.M.: Cost analysis for bridge and culvert, Seventh International Water Technology Conference IWTC7, Cairo, 2003.

[11] Hollar, D.A., Rasdorf, W., Liu, M., Hummer, J.E., Arocho, I.M.: Preliminary Engineering Cost Estimation Model for Bridge Projects, Journal of Construction Engineering and Management, 139 (2013) 9, pp. 1259-1267.

[12] Cheng, M.Y., Wu, Y.W.: Construction Conceptual Cost Estimates Using Support Vector Machine, 22nd International Symposium on Automation and Robotics in Construction ISARC 2005, Ferrara, Italy, 2005.

[13] Kim, K.Y., Kim, K.: Preliminary Cost Estimation Model Using Case-Based Reasoning and Genetic Algorithms, Journal of Computing in Civil Engineering, 24 (2010) 6, pp. 499-505.

[14] Fragkakis, N., Lambropoulos, S., Tsiambaos, G.: Parametric Model for Conceptual Cost Estimation of Concrete Bridge Foundations, Journal of Infrastructure Systems, 17 (2011) 2, pp. 66-74.

[15] Cirilovic, J., Vajdic, N., Mladenovic, G., Queiroz, C.: Developing Cost Estimation Models for Road Rehabilitation and Reconstruction: Case Study of Projects in Europe and Central Asia, Journal of Construction Engineering and Management, 140 (2013) 3.

[16] Pesko, I., Trivunic, M., Cirovic, G., Mucenski, V.: A preliminary estimate of time and cost in urban road construction using neural networks, Technical Gazette, 20 (2013) 3, pp. 563–570.

[17] Chou, J.S., Lin, C.W., Pham, A.D., Shao, J.Y.: Optimized artificial intelligence models for predicting project Automation in Construction, 54 (2015), pp. 106-115.

[18] Kovacevic, M., Ivanisevic, N., Dasic, T., Markovic, L.: Application of artificial neural networks for hydrological modelling in karst, Građevinar, 70 (2018) 1, pp. 1-10, https://doi.org/10.14256/JCE.1594.2016

[19] Beale, M.H., Hagan, M.T., Demuth, H.B.: Neural network toolbox, The Mathworks Inc., 2010.

[20] Kingston, G.B.: Bayesian Artificial Neural Networks in Water Resources Engineering, doctoral dissertation, University of Adelaide, Australia: School of Civil and Environmental Engineering, Faculty of Engineering, 2006.

[21] Matić, P.: Short-term forecasting of hydrological inflow by use of the artificial neural networks [Ph.D. thesis], Split: University of Split, Faculty of Electrical Engineering, Mechanical Engineering And Naval Architecture, Split, Croatia, 2014.

[22] Hastie, T., Tibsirani, R., Friedman, J.: The Elements of Statistical Learning, Springer, 2009.

[23] Black, P.E.: Dictionary of Algorithms and Data Structures, U.S. National Institute of Standards and Tecnology (NIST), 2012.

[24] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, London: The MIT Press, 2009.

[25] Breiman, L.: Bagging Predictors, Machine Learning, 24 (1996) 2, pp. 123-140.

[26] Breiman, L.: Random Forests, Machine Learning, 45 (2001), pp. 5-32.

[27] Kecman, V.: Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models, Cambridge, Masachusetts: MIT Press, 2001.

[28] Smola, A.J., Sholkopf, B.: A tutorial on support vector regression, NeuroCOLT Techincal Report TR-98-030, 2003.

[29] Rasmussen, C.E., Williams, C.K.: Gaussian Processes for Machine Learning, Cambridge, Massachusetts: The MIT Press, 2006.

[30] Menn, C.: Prestressed concrete bridges, Basel-Boston-Berlin, Birkhauser Verlag, 1990.

[31] Legates, D., McCabe, J.: Evaluating the use of goodnes-of-fit measures in hydrlologic and hydroclimatic model validation, Water Resources Research, 35 (1999) 1, pp. 233-241.

[32] Hagan, M.T., Demuth, H.B., Beale, M.H., Jesus, O.D.: Neural network design, Martin T. Hagan, 2014.

[33] Chang, C.C., Lin, C.J.: LIBSVM : a library for support vector machines, ACM Transactions on Intelligent Systems and Technology, 2 (2011) 3.