09 Research paper I Оригинални научни рад
DOI 10.7251/STP2215083S
ISSN 2566-4484

Nevena Simić, University of Belgrade, nsimic@grf.bg.ac.rs
Predrag Petronijević, University of Belgrade, pecap@grf.bg.ac.rs
Aleksandar Devedžić, TX Services, aleksandar.devedzic93@gmail.com
Marija Ivanović, University of Belgrade, mapetrovic@grf.bg.ac.rs

PRELIMINARY QUANTITY ESTIMATION IN CONSTRUCTION USING MACHINE LEARNING METHODS

Abstract

This paper analyses the problem of estimating the required quantities of major work items in the construction of residential and residential-commercial buildings using machine learning algorithms. The goal is to form a model that will provide a fast and sufficiently accurate estimate of the quantities of major work items, based on a small amount of known information on the technical characteristics and the environment of future residential and residential-commercial buildings. The case study included 71 projects of residential and residential-commercial buildings construction realised on the territory of the Republic of Serbia. Several models have been developed, and the paper presents those models that had the best performances. The models developed in this way can significantly contribute to resource planning and the accuracy of cost estimates in the early project phases.

Keywords: quantity estimation, cost estimation, machine learning, artificial intelligence

ПРЕЛИМИНАРНА ПРОЦЕНА КОЛИЧИНА У ГРАЂЕВИНАРСТВУ ПРИМЕНОМ МЕТОДА МАШИНСКОГ УЧЕЊА

Сажетак

У овом раду је анализиран проблем предвиђања потребних количина главних радова код изградње стамбених и стамбено-пословних објеката коришћењем алгоритама машинског учења. Циљ је формирање модела који ће на основу малог броја познатих информација о техничким карактеристикама и окружењу будућих стамбених и стамбено-пословних објеката пружити брзу и довољно прецизну процену количина главних радова. Студија случаја је укључивала 71 пројекат изградње стамбених и стамбено-пословних објеката реализованих на територији Републике Србије. Развијен је већи број модела, а у раду су приказани они модели који су имали најбоље перформансе. Овако развијени модели могу значајно допринети планирању ресурса и тачности процена трошкова у раним пројектним фазама.

Кључне ријечи: процена количина, процена трошкова, машинско учење, вештачка интелигенција

1. INTRODUCTION

The construction industry in the Republic of Serbia has recorded significant growth in recent years, and this is mostly due to the increase in the number of residential and residential-commercial buildings that represent the dominant type in the field of building construction, both in the world and in our country. According to the data from the Statistical Office of the Republic of Serbia, the number of apartments built in the Republic of Serbia increased several times in the period from 2015 to 2020 (Figure 1) [1]. Since there has been a constant growth in this branch of construction in the last few years, the question is whether the process of quantities and cost planning in the early stages of project development can be improved and accelerated. Additionally, many investors who finance the construction of residential buildings do not have the technical knowledge in the field of construction to be able to plan their resources in the early stages of investment. Therefore, the question arises as to how this problem can be overcome.

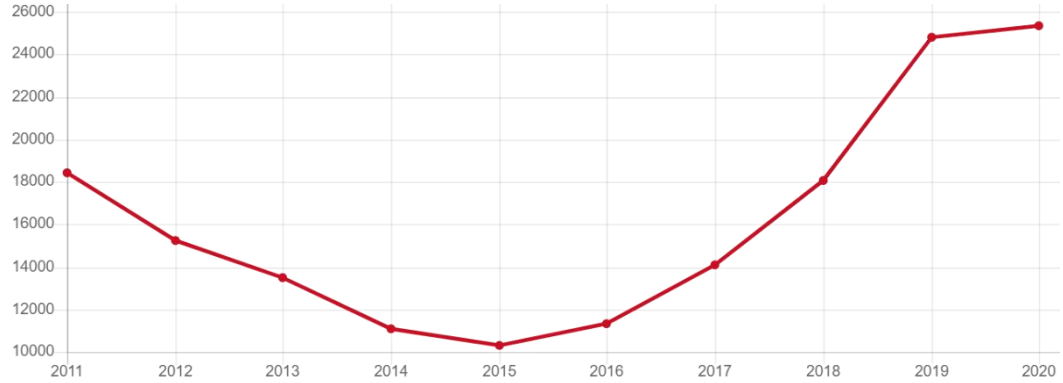

Figure 1. *Number of apartments built in the Republic of Serbia by years [1]*

In addition to the construction industry, the area that has encountered substantial growth in the world in recent years is the development of artificial intelligence systems. The main goal of artificial intelligence is to develop a system that will solve a problem in an intelligent way. Such systems are suitable for application in areas where a large amount of data is available, and as construction is one of the areas in which each project consists of a multitude of data, the collaboration of artificial intelligence and the construction industry is completely justified. Artificial intelligence finds great application in overcoming numerous problems in construction project management [2] [3] [4] [5]. There are a number of cost estimation models developed using different methods of artificial intelligence. Two different terms can often be found in this field, namely artificial intelligence and machine learning. Machine learning is an area of artificial intelligence in which conclusions are based on previous experience.

Garcia de Soto et al. [6] developed a methodology for estimating material quantities in the early project stages using MRA, ANN, and CBR. The authors point out that the developed methodology can increase the accuracy of cost estimates and provide estimates in a shorter time. Beljkaš et al. [7] developed an ANN model that showed a high level of accuracy, with a mean absolute percentage error of 8.56% and 17.31% for concrete and reinforcement consumption, respectively. Chou et al. [8] developed a cost estimation system based on estimating the quantity of individual cost items in highway construction projects based on information known in the early project stages. The main objective for the quantity estimation on the item-level instead of estimating at the project level is to improve accuracy by separating unit prices from the quantity estimation. Quantity estimates have been developed for the major work items in the WBS. A statistical parametric model for quantities estimation was developed for each major work item. The developed models are integrated into the cost estimation model in order to estimate the costs of individual work items and the total costs of the project. Petroutsatou et al. [9] used neural networks to estimate the cost of tunnel construction. The development of the neural network model consisted of two steps. In the first step, the quantity of works is estimated, while in the second step, the final costs are estimated based on the quantities that are the result of the first step, and which now represent the input parameters. Models were developed using MLFN and GRNN. The application of GRNN gave better results in terms of model accuracy.

This paper analyses the problem of predicting the required quantities of major work items in the construction of residential and residential-commercial buildings by using machine learning algorithms. The goal is to form a model that will provide a fast and sufficiently accurate estimate of

the quantities of major work items based on a small amount of known information on the technical characteristics and the environment of future residential and residential-commercial buildings. Estimated work quantities can provide a preliminary cost estimate by multiplying them by corresponding unit prices and final summirising. This approach, which is based on a work breakdown structure, is more useful for decision makers because it can separate uncertainties related to quantities from uncertainties related to cost, but also uncertainties related to different types of work.

2. METHODOLOGY

The methodology of the work (Figure 2) consists of four phases:

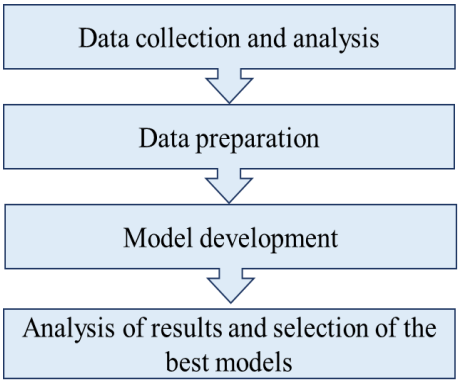

Figure 2. *Methodology of the model development*

3. DATA COLLECTION AND ANALYSIS

In the first phase, data on the realised projects for the construction of residential and residential-commercial buildings were collected. Data were collected by contacting contractors and design firms. The requested data included a technical description of a building, architectural design, and bill of quantities, and estimates of works. The total number of collected projects was 71, and the projects were implemented on the territory of the Republic of Serbia in the period from 2012 to 2020. In machine learning, the quality and size of the database are crucial for the success of prediction [10] so data collection was approached with special care.

After the data collection, there was an analysis of them and their limitations. The amount of data available in the early stages of project development is limited. For this reason, at the beginning of the research, the information and limitations available to the investor and the contractor in the early stages were analysed. These data represent the parameters on which the future cost and quantity estimates are based.

The first step in data processing is the analysis of the physical characteristics of the building such as the measures of the plot area, gross and net building area, and the vertical projection area of the building. The gross floor area ranges from 600 to 25,000 m2, the number of underground floors ranges from 0 to 2 while the number of above-ground floors ranges from 2 to 9. Another parameter that has been taken into account is the population density in the area in which the building is situated. The population density parameter is classified into three categories: medium density, dense, and very dense. The buildings analysed have five different types of facades: plastered, demit, demit-stone, demit-ventilated, and aluminum facades. There are two types of slabs on the buildings: FERT system slabs and reinforced concrete slabs. As for the method of foundation, buildings differ in whether they are built on piles or without them. It was also taken into account whether a building has an underground garage or not.

Quantity data were analysed only for those types of works that have standardized units of measure, namely earthworks, concrete works, reinforcement works, ceramic tiles works, hardwood floor works, and insulation works. Quantities for earthworks and concrete works were calculated in cubic meters, quantities for ceramic tilework, hardwood, and insulation work in square meters, and quantities of reinforcement work in tons.

In machine learning, data that significantly deviates from the other data are called outliers and can be removed automatically using the algorithm called Isolation Forest, but in this case, the use of this algorithm is not necessary due to a small amount of data and projects representing outliers are

manually eliminated. Outliers can significantly influence the outcome of machine learning model training, so this step is extremely important for further work.
The physical characteristics of the building were analysed in more detail in order to see their distribution and identify possible extreme values. Projects with a gross floor area of more than 15,000 m2 and a net floor area of more than 13,000 m2 are extreme values in terms of the area measurements. Also, extreme values include projects whose plot area exceeds 4,000 m2 and whose vertical projection exceeds 2,000 m2. This data can affect the accuracy of the future model and is therefore eliminated.
After the elimination of outliers, the final number of projects for model development was 52.

4. DATA PREPARATION

4.1. INPUT AND OUTPUT VARIABLES IDENTIFICATION

The first step in creating a model is to define input (independent) and output (dependent) variables. The quantities of major works were adopted as the output variables. Parameters that define the location and technical characteristics of the building were adopted as input variables of the model. The input and output variables are shown in Table 1.

Table 1. Input and output variables

Input variables		**Output variables**
Object type	Number of above-ground floors	Earthworks quantity (m^3)
Plot area	Number of underground floors	Concrete works quantity (m^3)
Gross area of the building	Slab type	Reinforcement works quantity (t)
Net area of the building	Facade type	Ceramic tile works quantity (m^3)
Floor area	Vertical projection surface	Hardwood floor works quantity (m^3)
Occupancy of the plot	City	Insulation works quantity (m^3)
Foundation type	Density of population	
Underground garage		

4.2. CORRELATION ANALYSIS

For the successful development of a prediction model, it is of great importance to know the correlation between the input and output variables of a model, but also the correlation between different input variables. For this correlation to be described by an adequate mathematical function, it is necessary that the variables between which the correlation is described have numerical, not categorical values. Table 2 shows the correlation between certain variables that have numerical values. The higher the value in the table cell, the greater the correlation between the two quantities, i.e. as one measure increases, so does another one. Correlations between all examined variables have positive results, which means that with the increase of the area of the building or plot the quantities of works also increase.

Table 2. Correlation between input and output variables

	Plot area	Gross area of the building	Net area of the building	Floor area	Earthworks quantity	Quantity of concrete works	Reinforcement works quantity	Ceramic tileworks quantity	Hardwood floor works quantity	Insulation works quantity
Plot area	1	0,92	0,93	0,98	0,8	0,91	0,9	0,88	0,85	0,87
Gross area of the building	0,92	1	1	0,94	0,87	0,95	0,93	0,93	0,93	0,85
Net area of the building	0,93	1	1	0,94	0,88	0,95	0,94	0,92	0,93	0,86
Floor area	0,98	0,94	0,94	1	0,84	0,91	0,91	0,86	0,86	0,86
Earthworks quantity	0,8	0,87	0,88	0,84	1	0,88	0,88	0,78	0,74	0,73
Quantity of concrete works	0,91	0,95	0,95	0,91	0,88	1	0,95	0,9	0,86	0,85
Reinforcement works quantity	0,9	0,93	0,94	0,91	0,88	0,95	1	0,87	0,86	0,87
Ceramic tileworks quantity	0,88	0,93	0,92	0,86	0,78	0,9	0,87	1	0,89	0,89
Hardwood floor works quantity	0,85	0,93	0,93	0,86	0,74	0,86	0,86	0,89	1	0,85
Insulation works quantity	0,87	0,85	0,86	0,86	0,73	0,85	0,87	0,89	0,85	1

Figure 3 shows an example of the correlation between the input variables the gross area of the building and the area of the plot and the amount of earthworks.

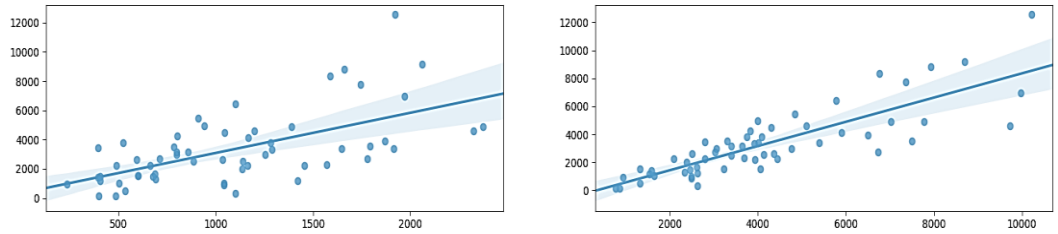

Figure 3. *Correlation between the quantity of earthworks and: a) the plot area, b) the gross area of the building*

4.3. ENCODING CATEGORICAL VARIABLES

Data such as the measures of the area or the number of floors are numerical data, so as such they can be passed directly to the model. However, data such as location, type of foundation, or façade type are categorical data, and as such must be encoded before being forwarded to the model.

Categorical data that has only two values, such as an underground garage (yes or no) or a type of foundation (with or without piles), can be easily encoded by replacing the category with a numeric value (1 or 0). However, data that contains more categorical values cannot simply be encoded with numbers because the question arises as to which value will get the higher score. In this case, the values of the categorical data are converted into columns, and the value of the data in the column corresponding to a certain categorical value will be 1, while in other columns representing the categorical values it will be 0 (Table 3).

Table 3. Encoding the categorical variables

Type of slab	Type of slab	Density of population	Medium-density	Dense	Very dense
RC	1	Medium-density	1	0	0
FERT	0	Dense	0	1	0
FERT	0	Very dense	0	0	1
RC	1	Dense	0	1	0
RC	1	Medium-density	1	0	0

4.4. FEATURE SCALING

The input parameters based on which the model will perform prediction can be quite different in terms of the size of numbers. Thus, for example, the parameter representing the number of above-ground floors ranges from 1 to 9, while the parameter representing the gross floor area of the building ranges up to 14,000. This difference in size can contribute to reducing the accuracy of the model, so the size of the input parameters must be scaled. Only numerical quantities are scaled, such as gross and net area, plot area, etc., and categorical quantities such as the type of facade are not scaled.
The most common types of data scaling are:

- standard scaling (scales sizes between -1 and 1 so that the average value is 0),
- robust scaling (scales sizes between two default values with eliminating outliers),
- normalisation (scales the sizes so that they tend to adapt to the normal distribution, i.e., Gaussian function),
- scaling between two quantities (scales sizes between two setpoints, usually 0 and 1).

In practice, there is no clearly defined instruction on which type of scaling to be used but in most cases, all four types of scaling are examined, which will be the case in this paper as well.

5. MODEL DEVELOPMENT

The total number of input parameters after encoding columns that contain categorical values is 15. Before the parameter values are passed to the model to begin model training, it is necessary to determine which combination of parameters will give the greatest accuracy. This problem is solved by iteration, which is automated using the Recursive Feature Elimination (RFE) function, and is available in the Scikit-Learn library, which is specialised in creating machine learning models.
The goal of RFE is to select the parameters that give the highest accuracy by recursively considering smaller and smaller sets of parameters. The model is first trained with one parameter from the set of all parameters and the accuracy of the model is determined through cross-validation. Then the number of parameters is increased and the combination of parameters that gives the highest accuracy of the model is sought. The disadvantage of this method is a large amount of time used due to numerous iterations. The number of input parameters that give the highest accuracy to the model varies from model to model.
The models were created and tested in the Python programming language with the help of the Scikit-Learn library, which specialises in machine learning, and the Keras library, which specialises in creating neural networks [11]. As it is not possible to know in advance which model will give the best performance, a total of 28 models were examined, so only those models that gave the best results will be considered in the following text. The following models were used (Table 4):

Table 4. Examined prediction models

1. Lasso Regression	15. K-Neighbors Regressor with K=2
2. Linear Regression	16. K-Neighbors Regressor with K=3
3. Passive Aggressive Regressor	17. K-Neighbors Regressor with K=5
4. Ridge Regression	18. K-Neighbors Regressor with K=7
5. SGD Regression	19. K-Neighbors Regressor with K=9
6. Decision Tree Regression	20. Ada Boost Regressor
7. Random Forest Regression	21. Ada Boost Regressor with Lasso
8. Support Vector Regression with linear kernel	22. Ada Boost Regressor with Linear Regression
9. Support Vector Regression with polynomial kernel	23. Ada Boost Regressor with Decision Tree Regression
10. Support Vector Regression with sigmoid kernel	24. Bagging Regressor with Extra Tree Regression
11. Support Vector Regression with RBF kernel	25. Bagging Regressor with Random Forest Regression
12. Extra Trees Regressor	26. Bagging Regressor with Lasso Regression
13. Gradient Boosting Regressor	27. Bagging Regressor with Linear Regression
14. K-Neighbors Regressor with K=1	28. Bagging Regressor with Decision Tree Regression

In addition to the above models, there was an examination of neural networks with a combination of different parameters such as:

- Number of hidden layers (between 2 and 8),
- Number of neurons (increased by 5 in the range up to 50-200),
- Type of activation function (ReLU, Softmax, and Tanh).

Finally, by comparing all the examined models of machine learning, both those that do not include neural networks and those that are based on neural networks, it can be concluded which models give the best results in terms of estimating the works quantities. The models that showed the best performance for estimating the required quantities are presented in Table 5.

Table 5. Quantity estimation models with the best performances

Required Quantity	**Model**	**Activation Function**	**Type of scaling**	**Hidden layers**	**Number of neurons**	**Train Score**	**Test Score**	**MAPE (%)**
Earthworks quantity	Neural Network	Relu	Standard Scaler	5	100	0.96	0.824	17,6
Concrete works quantity	Neural Network	Relu	Standard Scaler	5	150	0.927	0.883	11,7
Reinforcement works quantity	Neural Network	Relu	Standard Scaler	7	150	0.902	0.871	12,9
Ceramic tile works quantity	Neural Network	Relu	Standard Scaler	7	150	0.862	0.837	16,3
Hardwood floor works quantity	Neural Network	Relu	Standard Scaler	7	150	0.941	0.902	9,8
Insulation works quantity	K-Nearest Neighbors Regressor K=3	/	Robust Scaler	/	/	0.821	0.711	28,9

6. RESULTS AND DISCUSSION

By analysing the results, it can be concluded that the highest accuracy in estimating the required quantity of work was achieved by applying neural network models with the Relu activation function to estimate the quantity of earthworks, concrete, reinforcement, and hardwood works, while the use of classical models and neural networks achieved satisfactory accuracy in terms of predicting the quantity of ceramic and insulation works.

The evaluation of the accuracy of the models in this paper was examined by applying the measure MAPE for the mean absolute percentage error. MAPE is one of the most commonly used measures to assess the accuracy of the prediction [12] and has been used in a large number of studies related to the estimation of costs and quantities in construction (e.g. [13] [14] [15]). The mean absolute percentage error is defined by the following formula:

$$MAPE = \frac{1}{n} * \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| * 100 \quad (1)$$

where A_t represents the actual value and F_t is the predicted value.

The quantities of individual works do not depend on the quality of works or on the fact whether subcontractors are engaged, but mostly on the type of construction, building area, number of floors, etc., so models for estimating the quantities can give good results in the early stages of planning. Cases in which the quantity decreases with increasing dimensions of the building is practically impossible, and a larger difference in the quantity of individual works for two buildings that are similar in area and number of floors can be attributed to different types of construction (e.g., RC slab and FERT slab). Very poor performance of the model for estimating the quantity of insulation works can be attributed to the fact that there are different types of insulation products so, due to the quality of materials in some cases, higher than average consumption is required. This is one of the rare types of work in which the quality of the material can significantly affect the required quantity.

7. CONCLUSION

Suggested models showed high levels of accuracy, with MAPE ranging from 9,8% to 28,9%. According to PMI (*Project Management Institut*) [16], the accuracy of cost estimates in early phases of project development ranges from -25% to +75%, while with project progress through life cycle phases accuracy increases and can range from -5% to +10%.

After all the results presented in this paper, the question arises as to how they can be improved. Based on the prediction of the proposed models, decision-making on starting the project implementation in the early stages of project development can be significantly accelerated and improved, and further improvement can improve future planning steps. The very principle of machine learning is based on learning from a large number of data based on which it is necessary to draw conclusions. To create the previously mentioned models, data from 52 construction projects were used, which represents a minimum amount of data in the field of machine learning, and especially in the field of neural networks (deep learning). In neural networks, unlike classical models, in most cases, there is an increase in accuracy with the increasing of the dataset, so it can be concluded that increasing the number of projects would contribute to improving the performance of created models.

The source of data for creating the model in this research was the technical description and bill of quantities and estimates of works but they provide a limited amount of data. In order to get a realistic insight into the quantities of works, instead of data from bill of quantities and estimates, it is necessary to use data from the project of the constructed object. The bill of quantities and estimates of work is an assessment, which in itself carries a certain level of error, while the project of the constructed object contains actual data on the quantities and money spent, and as such is a much more relevant source of information.

Finally, it must be mentioned that the data for 52 objects based on which the analysis was performed were collected from more than 40 companies, which represents a big problem in terms of creating models due to high variability. Each design firm has its own design style but the project itself dictates the way the work is performed and that can lead to poorer model performance, especially if there is a small amount of data. If the predictions were made based on more than 500 projects carried out by 20 to 30 companies, the fact which company performed them would significantly affect the result. However, as there is no universal model that would have good performance for all types of projects,

the best performance of a model would be achieved if the model were created only based on the data provided by one company, i.e., if the model were made specifically for a particular company. This can substantially limit the application of such a system, especially in the Republic of Serbia, because the number of companies that have built more than 50 residential or residential-commercial buildings (which were the subject of research) is very small.

LITERATURE

[1] The Statistical Office of the Republic of Serbia, "The number of completed appartments per year in the Republic of Serbia," 2022. [Online]. Available: https://www.stat.gov.rs/oblasti/gradjevinarstvo/stanovi/. [Accessed: 25-Feb-2022].

[2] Y. Pan and L. Zhang, "Roles of artificial intelligence in construction engineering and management: A critical review and future trends," *Autom. Constr.*, vol. 122, no. November 2020, p. 103517, 2021.

[3] W. Eber, "Potentials of artificial intelligence in construction management," *Organ. Technol. Manag. Constr.*, vol. 12, no. 1, pp. 2053–2063, 2020.

[4] M. Kovacevic, N. Ivanišević, P. Petronijevic, and V. Despotovic, "Construction cost estimation of reinforced and prestressed concrete bridges using machine learning," *Gradjevinar*, vol. 73, no. 1, pp. 1–13, 2021.

[5] I. Peško *et al.*, "Estimation of Costs and Durations of Construction of Urban Roads Using ANN and SVM," *Complexity*, vol. 2017, 2017.

[6] B. García de Soto, B. T. Adey, and D. Fernando, "A hybrid methodology to estimate construction material quantities at an early project phase," *Int. J. Constr. Manag.*, vol. 17, no. 3, pp. 165–196, 2017.

[7] Ž. Beljkaš, M. Knežević, S. Rutešić, and N. Ivanišević, "Application of Artificial Intelligence for the Estimation of Concrete and Reinforcement Consumption in the Construction of Integral Bridges," *Adv. Civ. Eng.*, vol. 2020, 2020.

[8] J. Chou, M. Peng, K. R. Persad, and J. T. O. Connor, "Quantity-Based Approach to Preliminary Cost Estimates for Highway Projects," no. December, 2006.

[9] K. Petroutsatou, E. Georgopoulos, S. Lambropoulos, and J. P. Pantouvakis, "Early Cost Estimating of Road Tunnel Construction Using Neural Networks," *J. Constr. Eng. Manag.*, vol. 138, no. 6, pp. 679–687, 2012.

[10] M. Mohri, A. Rostamizadeh, and A. Talwalker, *Machine learning*, vol. 0. Cambridge, Massachusetts: The MIT Press, 2012.

[11] S. Raschka and V. Mirjalili, *Python Machine Learning - Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, 2nd editio., vol. 69, no. 4. Birmingham UK: Packt Publishing Ltd., 2017.

[12] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *Int. J. Forecast.*, vol. 32, no. 3, pp. 669–679, 2016.

[13] K. Tijanić, D. Car-Pušić, and M. Šperac, "Cost estimation in road construction using artificial neural network ´ 1," *Neural Comput. Appl.*, vol. 0123456789, pp. 9343–9355, 2020.

[14] M. Li, M. Baek, and B. Ashuri, "Forecasting Ratio of Low Bid to Owner's Estimate for Highway Construction," *J. Constr. Eng. Manag.*, vol. 147, no. 1, p. 04020157, 2021.

[15] B. J. Gardner, D. D. Gransberg, M. Asce, and J. A. Rueda, "Stochastic Conceptual Cost Estimating of Highway Projects to Communicate Uncertainty Using Bootstrap Sampling," *J. Constr. Eng. Manag.*, pp. 1–9, 2017.

[16] PMI, *A guide to the project management body of knowledge (PMBOK guide)*. Project Management Institut, 2017.