XV International Conference on Durability of Building Materials and Components
DBMC 2020, Barcelona
C. Serrat, J.R. Casas and V. Gibert (Eds)

Early Stage Construction Cost Prediction in Function of Project Sustainability

Diana Car-Pušić[1] **and Marko Mlađen**[2]

[1] Faculty of Civil Engineering, University of Rijeka, Radmile Matejčić 3, Rijeka, Croatia, diana.car.pusic@uniri.hr

[2] GT Trade Ltd., Spinčićeva 2d, Split, Croatia, marko.mladjen@gmail.com

Abstract. *Construction project costs often reach values higher than planned. Accuracy in project cost estimation is one of the most important criteria for project success, even for its sustainability.*
The main idea of this research is to examine the relationship between realized cost and contracted cost values for residential buildings. The aim of the research is to determine the mathematical relationship between realized and planned costs in the project implementation phase by using a few mathematical methods and some machine learning methods in comparison to linear regression. This would enable validation of methods themselves by comparing and evaluating the obtained relevant parameters. Comparison would be performed on two levels, based on its general characteristics, as well as on the results of their application on the basis of 24 building reconstructions and new buildings by comparing the mean absolute percentage error (MAPE) and the determination coefficient (R^2) using Predictive Modelling Software DTREG (pronounced D-T-Reg). The relationship of realized and planned costs will be determined for the building as a whole and for certain types of construction works. That relationship would enable more realistic budget planning of similar future projects. Cost overrun factors will be analysed for particular types of construction works, as well as the probability of their occurrence, and what measures should be undertaken to prevent or reduce them in similar future projects. The phenomenon known in project planning as "optimism bias" will be analysed in the context of research focus of exceeding the construction cost.

Keywords: *Building Construction, Contracted Cost, Realized Cost, Predictive Modelling Software, Machine Learning.*

1 Introduction

In the construction practice, the contracted construction cost overrun is a very common occurrence, which is a very frequent and undesired situation. The goal is to avoid or minimize cost overruns, which can be achieved by accurately estimating costs during project preparation, before signing a construction contract. This can be achieved by using scientifically based models obtained by applying adequate mathematical methods and by browsing the finished projects database. In contrast, experience shows that costs are usually calculated by method of unit price calculations and calculated quantities, but also estimated on a flat-rate basis, on the basis of experience, that is, without adequate scientifically or experientially based budgets. This often happens under deadline pressure as the cost information must be provided so that further project activities can continue. Thus, inherent uncertainties and risks, which are characteristic for all construction projects to a greater or lesser extent, are neglected, resulting in possible generation of higher costs of those estimated in the project.

Project cost estimation is a complex and challenging task, as evidenced by the cost overruns of numerous projects all over the world (Flyvbjerg *et al.*, 2002, Le-Hoai *et al.*, 2008; Žujo, 2008, Le-Hoai and Lee, 2009, Car-Pušić and Radujković, 2009; Žujo *et al*., 2010,). Scientific knowledge (Nikić, 1998, Radujković, 1999, Flyvbjerg *et al.*, 2002, Žujo, 2008; Le-Hoai *et.al.*, 2008, Žujo *et al.,* 2010; Car-Pušić and Radujković, 2009; Alshamrani, 2017, Petrusheva *et al.,* 2017, Juszczyk *et al.*, 2018) and experience show that the successful cost estimation requires:

1. Monitoring and registering data on constructed structures
2. Sufficient information about the new project
3. Application of multiple assessment methods and selection of the most accurate one.

Due to the diversity of projects, numerous risks involved and the complexity of cost estimation, researchers have often addressed this problem and developed a number of linear regression models to analyze costs depending on different variables, e.g. construction time (Žujo *et al.*, 2010) structure type, building area, number of floors, floor height (Alshamrani, 2017), geotechnical and construction variables (Petroutsatou *et al.*, 2006). Linear regression models using construction time, which is the most common independent variable, as a predictor (Žujo *et al.*, 2010) could be considered an inverse problem compared to the well-known "time-cost" ("TC") model used in the 1960s and founded by Bromilow (1969). Although this approach may be tempered by over-simplification because only one predictor is considered, numerous further studies have been conducted, which resulted in establishing of country-specific models, because the specific economic circumstances determine the parameters (Chan and Kumaraswamy, 1999; A.P.C. Chan, 1999; Car-Pušić, 2004). In some studies, this model has been used as the basis for the development of hybrid cost estimation models, which combine regression and neural networks (Petrusheva *et al.*, 2017; Petrusheva *et al.*, 2019).

The application of artificial neural networks for cost estimation with one independent variable, the construction time, has been researched by some authors (Petrusheva *et al.*, 2017, Juszczyk *et al.*, 2018; Tijanić and Car-Pušić, 2019; Tijanić *et al*., 2019).

The research analyzes completed projects of building construction without the material deviation from the usual one (Mlađen, 2017). The relationship between the realized and contracted costs is investigated. As these costs most often vary, the previous thesis is confirmed stating that often flat-rate and superficial assessment without applying scientifically proven methods and without considering the possible risks are employed. The difference in costs, expressed relatively, is known in the literature as "optimism bias". Available data are data on contracted and realized cost values and data on years of construction/reconstruction (Mlađen, 2017). Different methods seek to obtain the most accurate estimation model for this type of structure. Since these are the projects implemented in the territory of the Republic of Croatia in the County of Istria, the model is best applied for future projects in the area. This methodology can be applied to any type of facility and area. The cost difference obtained in this way can be considered a risk response measure.

2 The Main Research Goals

The main research goal is to define relation between the real and the contracted investment cost in building construction applying mathematical methods – linear regression and artificial neural

networks. The goal is to identify the model that provides the most accurate cost estimate and implementation guidelines.

3 Methodology

The set research hypothesis will be tested by applying linear regression and neural networks on data on constructed or reconstructed structures and by comparing the model accuracy.
The conditions of application for the most accurate cost estimation model will be analyzed.

3.1 Research Hypothesys

The value of construction works in building construction (reconstructions and new constructions) is not contracted in real terms, but usually below real values. The contracted value of works is the total value of works from the accepted best offer, which is usually obtained by method of unit price calculations and calculated quantities. The thesis is that the real value of the works most often exceeds the contracted value. There is a dependence between these values, which can be mathematically determined with satisfactory accuracy.

4 Research Methods, Data Base and Data Processing

Using the documentation review and analysis, data on contracted and realized prices of public and private building construction structures (new construction, reconstruction and rehabilitation), 24 in total, built on the territory of the County of Istria in the Republic of Croatia were obtained. (Mlađen, 2017).The investments were realized in the period from 2006 to 2017. The average overrun of the realized project values is 12.15%. The minimum overrun is 1.61% and the maximum is 41.49%. The standard deviation is 11.87%.

Predictive Modelling Software DTREG is used for data processing and modelling, which supports a wide range of models (regression models, decision trees for regression, neural networks, support vector machine, etc.). The database must be an ASCII file in Comma Separated Value (CSV) format with values for one case per row and one column for each variable. One of the variables is the target variable, while one or more other variables are predictors. DTREG analyzes the data and generates a model that gives the best estimate of the target variable based on the available predictor values. The software allows you to select model parameters from the multiple options offered. Moreover, the iteration offers the expected optimal parameters (Sherrod, 2014). In this research, the target variable is the value of the real construction cost C_R, and the predictor is the value of the expected, planned and contracted construction costs.

Considering the research goals and scientific hypothesis, the following methodological procedure is applied:

1. Application of linear regression
2. Application of neural networks and Support Vector Machine (SVM) model
3. Comparison of results for the validation data, discussion, application proposal.

4.1 Linear Regression

First, the simple regression analysis method was applied as a common and simple, well-understood, widely used method (Sherrod, 2014), and the adequacy and accuracy of the model

obtained was determined by appropriate indicators. The general form of the applied regression function is:

$$y = \beta_0 + \beta_1 x \tag{1}$$

4.1.1 Output generated for linear regression

The following parameters were obtained using simple regression analysis using DTREG software:

Table 1. Results for linear regression.

Variable	Coefficient	Std. Error	t	Prob.(t)
C_C	1.128	0.0137	82.22	<0.00001
Constant	14.877,5	1.532e+004	0.97	0.34209

The regression function is:

$$C_R = 1{,}128C_C + 14877{,}5 \tag{2}$$

The t Statistic value is 82.22 with a probability <0.00001. The coefficient R^2 as a proportion of variance explained by the model is 0.98849 (98.849%). The correlation coefficient R between actual and predicted values is 0.99684 (99.684%). Mean Absolute Percentage Error (MAPE) is 13.824% for validation data.

4.2 Neural Networks and SVM Model

In an attempt to obtain a more accurate cost estimation model, i.e. a lower value of the MAPE parameter, the next step includes the application of General Regression Neural Network (GRNN), Radial Basis Function (RBF) and Support Vector Machine (SVM). These models were selected because they are suitable for continuous variables and for smaller databases. The Multilayer Perceptron (MLP) network has proven to be inappropriate due to too little data was excluded from further consideration.

4.2.1 Output generated for GRNN, SVM, RBF

For all models, the expected optimal initial parameters provided by the software were used.

For GRNN, the Gaussian kernel function was applied, σ values were calculated for each predictor variable (min σ 0,0001 and max σ 10, with 20 search steps) and Leave One Out (LOO) validation method of evaluating σ values during the optimization process was used. Each training model is built with all training rows except one and then, the error was evaluated. This was repeated for all rows, and the error was averaged (Sherrod, 2014).

The model has been reduced to 7 neurons when the minimum error occurred. Despite the small number of retained neurons, the reduction of neurons should be performed because it significantly contributes to the accuracy of the model. In this case, the MAPE was reduced from 75.488% to 35.130%. The coefficients R^2 and R are shown in Table 2.

The initially defined type of SVM model Epsilon-SVR model was used for the SVM. Another possible option is the V-SVR model. The differences that may arise are small and the other model has not been applied. Radial Basis Function (RBF) has been implemented as the kernel function, which in most cases produces the best results. Kernel function transforms the

input data into an n-dimensional space where a hyperplane can be constructed to partition the data (Sherrod, 2014). The optimal approximation linear function will be obtained in the new space (Petrusheva *et al.,* 2017).

Using the RBF, the SVM model generates the RBF network architecture. The difference is in the number and position of nodes. At SVM, their position is on the support vectors (Sherrod, 2014). 24 support vectors are used by the model.

Model validation is performed by Cross validation method with 10 cross-validation folds. At SVM, low values were obtained for R and R^2 as well as for MAPE 14.814% (Table 2).

After three iterations, the three-neuron RBF network gave the best result of 10.976% for MAPE, but lower values of R^2 and R (Table 2). The Cross validation method with 10 cross-validation folds was also applied as the validation method.

Table 2. Results for linear regression and neural networks for source data.

Parameter	LR	GRNN	SVM	RBF
R^2	0.988	0.605	0.345	0.556
R	0.997	0.919	0.644	0.796
MAPE /%/	13.824	35.130	14.814	10.976

4.3 Application of Input Variable Natural Algorithms

Following the time-cost model, the following model is assumed:

$$C_R = E \times C_C^F \tag{3}$$

C_R = realized price
C_C = contracted price
E = model parameter that shows the average real price for monetary value construction
F = model parameter that shows real cost dependence of contracted cost changes

The model can be expressed as a logarithm to obtain the following form:

$$lnC_R = lnE + FlnC_C \tag{4}$$

It means:

$$\left.\begin{array}{rcl} lnE & = & \beta_0 \\ E & = & e^{\beta_0} \end{array}\right\} \tag{5}$$

$$F = \beta_1 \tag{6}$$

4.3.1 Output generated for linear regression for logharitmized data

The natural logarithms of C_C predictors and the target variables C_R were calculated. Applying regression analysis on natural logarithms using DTREG software, the following parameters were obtained:

Table 3. Results for linear regression.

Variable	Coefficient β_1	Std. Error	t	Prob.(t)
lnC_C	1.03367	0.0165	62.64	<0.00001
Constant β_0	0.0700096	0.02752	2.54	0.01849

Using the information in the previous table (Table 3), the regression function reads:

$$lnC_R = 0{,}07 + 1{,}03367 lnC_C \tag{7}$$

$$lnE = 0{,}07$$

$$E = 1{,}07$$

$$F = 1{,}03367$$

The model which is obtained is:

$$C_R = 1{,}07 \cdot C_C^{1{,}03367} \tag{8}$$

The value of t Statistic is 62.64 with a probability of p (t) <0.00001. The values of R^2 and R are 0.993 and 0.997 respectively. These values confirm the regression dependence. The MAPE is 11.816%. (Table 4).

4.3.2 Output generated for GRNN, SVM, RBF for logharitmized data

The Gaussian kernel function was applied for GRNN, σ values were calculated for each predictor variable (min σ 0.0001 and max σ 10, with 20 search steps) and Leave one out (LOO) validation method was used. The model has been reduced to 18 neurons when the minimum error occurred. In this case, the MAPE is 12.291%. The coefficients R^2 and R are shown in Table 4.

After four iterations, the three-neuron RBF network earns the MAPE value of 23.810%, and the R^2 and R values shown in Table 4. As a validation method, the Cross validation method with 10 cross-validation folds was also applied.

The initially defined Epsilon-SVR model was used for the SVM. Radial Basis Function (RBF) was applied as the kernel function. 24 support vectors are used by the model. Model validation was performed by Cross validation method with 10 cross-validation folds.

Table 4. Results for linear regression and neural networks for logharitmized data.

Parameter	LR	GRNN	SVM	RBF
R^2	0.993	0.962	0.995	0.693
R	0.997	0.985	0.997	0.839
MAPE /%/	11.816	12.291	6.469	23.810

MAPE is 6.469 %, R^2 and R values are 0.995 and 0.997 respectively. By applying the SVM model to logarithmic values, the lowest model error value is obtained.

5 Discussion of Results

By applying the linear regression model to the original data, a MAPE of 13.824% was obtained. In order to obtain a more accurate cost estimation model, the neural network models GRNN and RBF and the SVM model were applied. RBF gives a more favorable MAPE of 10.976%.

By applying the linear regression and RBF, GRNN and SVM to the natural logarithms of the input data (following the Bromilow "time-cost" model), (Bromilow, 1969) the lowest MAPE was obtained using SVM. The MAPE is 6.469%. Compared to the lowest MAPE obtained by processing the original data, this one is lower by 4.507%.

6 Conclusions

- The real costs of new constructions and reconstructions of structures in building construction field often exceed contracted values. The accepted tender price, obtained from the bill of quantities and unit prices (C_C) is taken as the contracted price. The aim of the research was to obtain a model that accurately defines the functional dependence of the real price C_R and the planned price C_C. Thus obtained real price should be taken as the contract price. This would reduce the risk of the contract price being exceeded. Although the database consists of a relatively small amount of data, the possibility of applying linear regression models, neural networks and SVM models to estimate the real cost of construction and reconstruction of high-rise buildings is presented. These models were applied to the original values C_R and C_C as well as to the values of natural logarithms. Namely, following the Bromilow "time-cost" model, the model $C_R = E \times C_C^F$ was established and tested. By expressing it in logarithms, the linear regression function is obtained.
- By applying linear regression and neural networks to the values of the natural logarithms of the variables, it was found that more accurate results can be obtained compared to models with original data. The best results, i.e. the lowest MAPE of 6.469%, are obtained by the SVM model applied to natural logarithm values of the predictor and the target variable.
- Despite the small database, the obtained insight could be relevant to solving the real cost prediction problem when working with larger databases. The results should also be investigated for civil engineering databases.

Acknowledgements

This work has been fully supported by the University of Rijeka under the project uniri-tehnic-18-125.

ORCID

Diana Car-Pušić: https://orcid.org/0000-0003-2555-335X
Marko Mlađen: -

References

Alshamrani, O.S. (2017). Construction Cost Prediction Model for Conventional and Sustainable College Buildings in North America. *Journal of Taibah University of Science, 11(2),* 315-323. doi:10.1016/j.tusci.2016.01.004

Bromilow, F.J. (1969).Contract Time Performance Expectations and the Reality. *Building* Forum, *1(3),* 70-80.

Car-Pušić D. (2004). *Metodologija planiranja održivog vremena građenja* (in Croatian), PhD Thesis, Građevinski fakultet Sveučilišta u Zagrebu, Zagreb, Croatia.

Car-Pušić, D. and Radujković, M. (2009). Construction Time-Cost Model in Croatia. *International Journal for Engineering Modelling, 22(1–4),*63–70.

Chan, W.M.D. and Kumaraswamy, M.M. (1999). Forecasting Construction Durations for Public Housing Projects Hong Kong Perspective. *Building and Environment, 34(5),* 633-646. doi: 10.106/s0360-1323(98)00040-7

Chan, A.P.C. (1999). Time-cost Relationship of Public Sector Projects in Malaysia. *International Journal of Project Management, 19(4),* 223-229. doi:10.1016/S0263-7863(99)00072-1

Flyvbjerg, B., Holm, M.S. and Buhl, S.L. (2002). Underestimating Costs in Public Works Projects: Error or Lie? *Journal of the American Planning Association, 68(3),* 279–295. doi:10.1080/01944360208976273

Juszczyk, M., Leśniak, A. and Zima, K. (2018). ANN Based Approach for Estimation of Construction Costs of Sports Fields. *Complexity, Article ID 7952434, 11 pages*. doi:10.1155/2018/7952434

Le-Hoai L., Lee Y.D. and Lee, J.Y. (2008). Delay and Cost Overruns in Vietnam Large Construction Projects: a

Comparison with Other Selected Countries. *KSCE Journal of Civil Engineering, 12(6),*367–377. doi: 10.1007/s12205-008-0367-7

Le-Hoai L. and Lee Y.D. (2009). Time-Cost Relationships of Building Construction Project in Korea. *Facilities, 27 (13/14)*, 549–559. doi: 10.1108/02632770910996379

Mlađen, M. (2017). *Analiza uzroka i vjerojatnosti troškovnih odstupanja u projektima visokogradnje*, (in Croatian), Master's Thesis, Građevinski fakultet Sveučilišta u Rijeci, Rijeka, Croatia.

Nikić, R. (1998). *Upravljanje rizicima kod građevinskih projekata zemlje u tranziciji* (in Croatian), Master's Thesis, Građevinski fakultet Sveučilišta u Zagrebu, Zagreb, Croatia.

Tijanić, K., Car-Pušić, D. and Šperac, M. (2019). Cost Estimation in Road Construction Using Artificial Neural Network, *Neural Computing and Applications, accepted for publishing.* doi: 0.1007/s00521-019-04443-y

Petroutsatou, C., Lambropoulos, S. and Pantouvakis, J.P. (2006). Road Tunnel Early Cost Estimates Using Multiple Regression Analysis. *Operational Research 6(3),* 311-322. doi: 10.1007/BF02941259

Petrusheva S., Zileska-Pancovska, V., Žujo, V. and Brkan-Vejzović, A., (2017). Construction Costs Forecasting: Comparison of the Accuracy of Linear Regression and Support Vector Machine Models, *Technical Gazette, 24(5),* 1431-1438. doi:10.17559./TV-20150116001543

Petrusheva, S., Car-Pušić, D. and Zileska-Pancovska, V. (2019). Support Vector Machine Based Hybrid Model for Prediction of Road Structures Construction Costs. *IOP Conference Series: Earth and Environmental Science 222 (1755-1307).* doi:10.1088/1755-1315/222/1/012010.

Radujković, M. (1999). Upravljanje rizikom i resursima kod građevinskih projekata, znanstveno istraživački project, MZITRH (Ministarstvo znanosti i tehnologije Republike Hrvatske) [Construction Project Risk and Resource Management, scientific research work of MST (Ministry of Science and Technology)], Građevinski fakultet, Sveučilište u Zagrebu, Faculty of Civil Engineering, University of Zagreb. (in Croatian)

Sherrod, P.H. (2014). Manual DTREG Predictive Modeling Software

Tijanić, K. and Car-Pušić, D. (2019).The Assessment of School Operational Costs by Using Artificial Neural Networks. In *Proceedings of the VII Gathering of Young Researchers in the Field of Civil Engineering and Related Technical Sciences Common found 2019*, Rijeka, Croatia, 126-131.

Žujo, V. (2008). *Upravljanje građevinskim projektima kroz planiranje vremena građenja* (in Bosnian), PhD Thesis, Građevinski fakultet Univerziteta Džemal Bijedić u Mostaru, Mostar, Federation of Bosnia and Herzegovina.

Žujo, V., Car-Pušić, D. and Brkan-Vejzović, A. (2010). Contracted price overrun as contracted construction time overrun function. *Technical Gazette, 17(1),* 23-29.