Predicting buildings construction cost overruns on the basis of cost overruns structure

Key words: cost overrun, quantities of works, unforeseen works, additional works, linear regression analysis, soft computing methods

Introduction

Construction costs are one of the main criteria for decision making in the early stages of the construction process, and therefore their prediction is of interest to all project participants (Ambrule & Bhirud, 2017). Experience has shown that very often there are discrepancies between the estimated costs in relation to the realized costs of the construction project (Peško Trivunić, Cirović & Mučen- ski, 2013) and discrepancies occur due to lack of data and information in the conceptual phase (Al-Zwainy & Aidan, 2017). The aim is to avoid or minimize cost overruns, which can be achieved by accurate cost estimation during project preparation before signing a construction contract (Car-Pušić & Mladen, 2020).

In previous studies, various mathematical methods and tools have been used to solve the problem of predicting construction costs and cost overruns in construction projects, such as simple and multiple linear regression, “soft computing” methods, such as: neural networks, machine learning, fuzzy logic, etc. (Plebankiewicz, 2018). It was found that the experience of contractors in previous construction projects is an important element that can help to avoid mistakes and increase the chances of success of future projects in the construction phase. Construction cost data collected from previous projects can be useful for estimating costs in different phases of the project life cycle using linear regression and “soft computing” methods (Tijanić, Car-Pušić & Šperac, 2019).

Researchers have developed several linear regression models to analyse costs and cost estimation depending on different variables, such as construction time (Žujo, Car-Pušić & Brkan-Vejzo-
Predicting buildings construction cost overruns on the basis of cost overruns structure

vić, 2010), structure type, building area, number of floors, floor height (Alshamrani, 2017), geotechnical and construction variables (Petroutsatou, Lambropoulos & Pantouvakis, 2006), etc. Cost estimation models based on linear regression and construction time as an independent variable (predictor) (Žujo et al., 2010) could be considered as an inverse problem compared to the well-known Bromilow time–cost (TC) model (Bromilow, 1969). While this approach can be criticized for its simplicity due to only one independent variable, numerous other studies have been carried out which led to the establishment of country/area specific models with high accuracy (Chan & Kumaraswamy, 1999; Chan, 2001; Car-Pušić, 2004). In some studies, this inverse Bromilow model has been used as a basis for the development of hybrid cost estimation models combining regression and neural networks (Petrusheva, Zileska-Pancovska, Žujo & Brkan-Vejzović, 2017; Petrusheva, Car-Pušić & Zileska-Pancovska, 2019; Car-Pušić & Mladen, 2020).

The use of artificial neural networks for cost estimation with construction time as an independent variable has been investigated by several authors (Petrusheva et al., 2017; Tijanić & Car-Pušić, 2019; Tijanić et al., 2019). Some other authors (Hegazy & Ayed, 1998; Attala & Hegazy, 2003; El-Kholy, 2015) developed models for predicting the extent of cost overruns in construction projects based on regression analysis and neural networks.

The main research goal

The main research goal is to identify the discrepancy between realized construction costs and contractually agreed costs for construction projects and to analyse the reasons for the discrepancies, based on the available data for the high-rise buildings (Mladen, 2017). The discrepancy is usually due to quantity variances compared to the contractually agreed quantities, which are due to unforeseen and sometimes additional work as a result of change orders by the customer. According to the Croatian Chamber of Economy and Croatian Employers Association, “additional works are those which have not been contractually agreed and are not necessary for the performance of the contract, but which are required by the client”, and “unforeseen works are urgent works which the contractor has to carry out without the consent of the client in order to maintain the stability and safety of the building, the environment and persons or for the smooth regular execution of the works, and here are other unforeseen works which are necessary for technological and/or functional reasons for the regular execution of the contractually agreed works” (Hrvatska Gospodarska Komora, 2017).

The above definition of unforeseen works emphasizes, inter alia, that unforeseen works are carried out to preserve the environment and the people living in it, which is one of the components of sustainable construction. Construction is an important area for achieving the objectives of sustainability (sustainable development), as it concerns all three aspects of sustainability: economic development, social development, and environmental protection. The first step towards sustainability of buildings is to think about the life cycle of a building, whereby sustainable construction should be seen as a comprehensive process.
capable of understanding and meeting the needs and requirements of users while reducing environmental impact and life cycle costs (Vezilić Strmo, Senjak & Štulhofer, 2014; Biolek, Hanak & Maro-vić, 2017).

As construction costs are a significant part of the life cycle of any building, it is very important to plan them properly, avoid overruns and achieve a sustainable project outcome. A sustainable project outcome has positive benefits and long-term social, economic, and environmental impacts. A project is successful if it achieves the objectives or needs defined in the acceptance criteria within an agreed time frame and budget. The main criteria for measuring the success of project implementation methods are cost, quality, time, safety, and how the project ultimately achieves its intended purpose (Fong, Avetisyan & Cui, 2014).

Taking all these aspects into account, the main goal of this work is to model the relationship between the realized and contractually agreed construction costs with the best possible accuracy by applying linear regression and “soft computing” methods, and to contribute in this segment to bringing the construction project to a sustainable level.

Research hypothesis

According to the available data, the realized construction costs very often exceed the contractually agreed costs. Therefore the following research hypothesis shall be tested: Three types of works cause the cost discrepancy with different effects on them. Further, there is a relationship between realized and contractually agreed construction costs and this can be modelled with acceptable accuracy.

Methodology

The data for this study were collected by interviewing the site managers responsible for the individual construction projects. The site managers also provided the author with project documentation. By studying and analysing the documentation and the information obtained in interviews, a database was created, which includes 24 public and private high-rise buildings – new construction, renovation and reconstruction – built between 2006 and 2017 in Istria County, Republic of Croatia. The database contains basic data on the constructed buildings, such as the type of intervention, the year of construction and data on the contractually agreed and realized construction costs. Data were also collected on the types of works that caused the cost differences: differences in quantity, unforeseen works (Mladen, 2017). Projects of similar or identical nature and complexity were taken into account to make the research result as credible as possible.

A summary of the data collected is presented in Table 1. More detailed information on the projects is available from the authors of this paper and at Mladen (2017).

For the data collected, the main statistical indicators of total cost overruns and cost overruns caused by differences in quantities, unforeseen works and additional works were calculated. A model was then developed to estimate the reali-
Predicting buildings construction cost overruns on the basis of cost overruns structure

zed construction costs with the lowest error value using linear regression, neural networks, and support vector machine.

A review of the literature has shown that these methods give very good results when applied to similar data sets as in this paper (Attala & Hegazy, 2003; Car-Pušić, 2004; Žujo et al., 2010; El-Kholy, 2015; Petrusheva et al., 2017; Tijanić & Car-Pušić, 2019; Tijanić et al., 2019). In a comparable case, the authors of this paper decided to investigate the applicability of the selected methods in estimating construction costs to obtain models that could be helpful in future cost estimates. Furthermore, the methods proved to be suitable given the amount of data collected.

### Statistical indicators

The basic statistical indicators for the collected database, which are presented in Table 2, are calculated. The average cost overrun of the contractually agreed construction costs is 12.15% with a standard deviation of 11.87%.

According to the relevant Croatian legislation accompanying the construction works, as it said, there are three types of possible differences between contractual and realized works, which may lead to cost differences: differences in quantity, additional works, and unforeseen works. The quantities of work may be larger or smaller, resulting in higher or lower costs ($Q$). They are calculated using the

### TABLE 1. Summary of the data collected (Mladen, 2017)

<table>
<thead>
<tr>
<th>Project type</th>
<th>Projects number</th>
<th>Average construction year</th>
<th>Contracted construction costs</th>
<th>Realized construction costs</th>
<th>Costs incurred by differences in quantities</th>
<th>Costs incurred by additional works</th>
<th>Costs incurred by unforeseen works</th>
</tr>
</thead>
<tbody>
<tr>
<td>New private investment</td>
<td>10</td>
<td>2014</td>
<td>127 164.81</td>
<td>149 995.80</td>
<td>4 340.57</td>
<td>16 514.00</td>
<td>1 976.42</td>
</tr>
<tr>
<td>Renovation and reconstruction of private investment</td>
<td>8</td>
<td>2014</td>
<td>52 752.97</td>
<td>57 588.75</td>
<td>–668.43</td>
<td>3 626.03</td>
<td>1 878.18</td>
</tr>
<tr>
<td>Renovation and reconstruction of public investment</td>
<td>6</td>
<td>2012</td>
<td>38 723.57</td>
<td>43 189.77</td>
<td>2 197.03</td>
<td>1 960.05</td>
<td>309.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Specification</th>
<th>$OB%$</th>
<th>$Q$(EUR)</th>
<th>$Q%$</th>
<th>$A$(EUR)</th>
<th>$A%$</th>
<th>$U$(EUR)</th>
<th>$U%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>12.15%</td>
<td>–2 135.02</td>
<td>–48.36%</td>
<td>8 579.52</td>
<td>97.83%</td>
<td>1 526.85</td>
<td>50.53%</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>11.87%</td>
<td>17 747.40</td>
<td>184.70%</td>
<td>19 537.05</td>
<td>112.67%</td>
<td>2 942.24</td>
<td>168.63%</td>
</tr>
</tbody>
</table>

$A$ – costs incurred by additional works, $U$ – costs incurred by unforeseen works, $OB$ – optimism bias.
same unit prices agreed in the contract as for the contracted quantities. Unforeseen works are important because they are essential for the stability of the structure, safety, and environmental protection. They give rise to additional costs \( (U) \), especially as they are calculated using the new unit prices which are not laid down in the basic contract. Additional works are not agreed in the basic contract without contractually agreed unit prices. There are works that are required by the client and are not indispensable, which generates additional costs \( (A) \). There is no doubt that they are all the result of poor preliminary planning.

The “optimism bias” is defined as the tendency for a project’s costs to be underestimated and/or benefits to be overestimated. It is expressed as the percentage difference between the estimate at appraisal and the final outturn (MacDonald, 2002). Formula (1) is used for calculation:

\[
\text{optimism bias} = 100 \times \frac{\text{realized value} - \text{contracted value}}{\text{contracted value}} \% 
\]  

\( (1) \)

The range of optimism bias regarding contracted and realized costs is between 1.6% and 41.5%, which in practice can be regarded as frequent and usual cost overruns. Cost overruns of 50–60% and more, sometimes even 100%, can be considered very high, with negative effects on the quality of the samples and the final model.

Figure 1 shows the structure of the deviations, which indicate that the main cause of cost overruns is additional work that is not necessary but is carried out at the request of the client after the construction contract has been signed. They account for 97.83% of the total cost overrun, with a standard deviation of 112.67%. This shows that the planning of the project budget by the client at the project preparation stage is not appropriate.

Unforeseen work contributes to the exceedance with 50.53% with a standard deviation of 168.63%. This indicates insufficient and inadequate previous investigation work, which can be considered a serious deficit in terms of sustainable construction. In this study, the quantity deviation speaks in favour of quantity differences, i.e. it is negative (−48.36%) with a standard deviation of 184.70%. The contractually agreed quantities are on average higher than the services provided. Although these statistical indicators may seem strange at first glance, the explanation lies in the number of cases in the database, the variety of properties and the type of construction work. These are the reasons for such large standard deviations.
Model for predicting the realized construction costs

The modelling of the relationship between realized and contractually agreed costs based on available data was carried out in three steps for two groups of variable cases, as shown in Table 3.

As the statistical indicators show, the largest cost overrun is caused by additional work that is difficult to predict and sometimes unjustified. For this reason, modelling was also carried out for the value of the target variable \( CR - A \). Linear regression was used, as well as the general regression neural network (GRNN), the support vector machine (SVM) and the radial basic function neural network (RBF). A multilayer perceptron (MLP), which is very often used for data estimates, was not applicable because the database did not have enough data.

The predictive modelling software the DTREG was used, which is a powerful statistical analysis program that generates neural networks and other techniques (support vector machine, gene expression programming, discriminant analysis, linear and logistic regression model, etc.) that describe data relationships and can be used to predict values for future observations (Sherrod, 2014).

TABLE 3. Steps and groups of variables in modelling the ratio of realized and contracted costs (own studies)

<table>
<thead>
<tr>
<th>Step</th>
<th>Data nature</th>
<th>First group of variables pairs</th>
<th>Second group of variables pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>predictor</td>
<td>target</td>
</tr>
<tr>
<td>1</td>
<td>original data</td>
<td>( C_C )</td>
<td>( C_R )</td>
</tr>
<tr>
<td>2</td>
<td>original optimism bias data</td>
<td>( OB )</td>
<td>( C_R )</td>
</tr>
<tr>
<td>3</td>
<td>natural logarithms</td>
<td>( \ln C_C )</td>
<td>( \ln C_R )</td>
</tr>
</tbody>
</table>

\( C_C \) – contracted cost of construction, \( C_R \) – realized cost of construction, \( C_R - A \) – realized cost of construction minus costs of additional works, \( OB \) – optimism bias, \( \ln C_C \) – natural logarithm of contracted cost of construction, \( \ln C_R \) – natural logarithm of realized cost of construction, \( \ln(C_R - A) \) – natural logarithm of realized cost of construction minus natural logarithm of costs of additional works.
The DTREG software is characterized by the self-optimization of the model parameters to give them the smallest error in the estimation (Tijanić et al., 2019).

The natural logarithms of the variables were modelled based on Bromilow’s time–cost model (Bromilow, 1969). In this study the model in formula (2) was assumed as follows:

\[ C_R = E \cdot C_C^F \]  

(2)

where:
- \( E \) – model parameter that shows the average real price for monetary value of construction,
- \( F \) – model parameter that shows real cost dependence of contracted cost changes.

Taking the logarithm results in formula (3):

\[ \ln C_R = \ln E + \ln C_C \]  

(3)

By checking the value of t-statistics (62.64 with \( p < 0.00001 \), \( DF = 1 \) and \( R^2 = 0.993 \)), it was found that a regression function can be applied (Car-Pušić & Mladen, 2020).

Results and discussion

The results for coefficient of determination (\( R^2 \)) and mean absolute percentage error (\( MAPE \)) are shown below. The \( MAPE \) and \( R^2 \) are most often used estimators of the accuracy of the model (Petrusheva, Car-Pušić & Zileska-Pancovska, 2016). The \( MAPE \) is a measure of prediction accuracy and it is defined by formula (4):

\[ MAPE = \frac{1}{N} \sum \frac{\text{realized value} - \text{contracted value}}{\text{realized value}} \times 100 \% \]  

(4)

Coefficient of determination indicates how well data points match the approximation function which is obtained from the model – it is a measure of the general match of the model. The value \( R^2 = 0.9700 \) can be interpreted as: 97% of the variation in the response can be explained by the predictor variables. The remaining 3% can be attributed to unknown variables or inherent variability (Petrusheva et al., 2016).

The results obtained by data processing from Table 3 are given below. Given the nature of the data used, the results are shown in steps 1 to 3. Within each step, cost estimation models are developed by modifying the target variable and the predictor variable, all with the aim of obtaining a model with the lowest possible estimation error.

Step 1. Use of original data:
- Model I: Target variable \( C_R \) is function of predictor variable \( C_C \) \([C_R = f(C_C)]\);
- Model II: Target variable \( C_R - A \) is function of predictor variable \( C_C \) \([C_R - A = f(C_C)]\).

Table 4 shows the results of the mentioned indicators for validation data. The best result for \( RBF \) is for the original \( C_C \) and \( C_R \) data, but with a low \( R^2 \) value. The reduction of additional works did not result in a more accurate model.

Step 2. Use of original optimism bias data:
- Model III: Target variable \( OB \) for values \( C_R \) is function of predictor variable \( C_C \) \([OB(C_R) = f(C_C)]\);
Predicting buildings construction cost overruns on the basis of cost overruns structure 373

Model IV: Target variable \(OB\) for values \(CR - A\) is function of predictor variable \(CC\) \([OB(\text{CR} - A) = f(\text{CC})]\).

Using these models and selected estimation methods, very poor results were obtained. The highest \(R^2\) value obtained is only 0.136 for \(SVM\), while all MAPE values are extremely high (the lowest value is over 70%).

Step 3. Using data in the form of natural logarithms:
- Model V: Target variable \(\ln CR\) is function of predictor variable \(\ln CC\) \([\ln CR = f(\ln CC)]\);
- Model VI: Target variable \(\ln (CR - A)\) is function of predictor variable \(\ln CC\) \([\ln (CR - A) = f(\ln CC)]\).

Using natural logarithms, the highest values of \(R^2\) and the lowest values of MAPE were obtained, as can be seen from Table 5. Good results were obtained when \(\ln CC\) and \(\ln CR\) were used as variables instead of the source values using linear regression and \(SVM\). The best model accuracy is \(MAPE = 0.522\%\), with \(R^2 = 0.994\). These values indicate a model of acceptable accuracy. This proves the second part of research hypothesis. Again, the accuracy of the model was not improved by excluding the value of the additional work. Nevertheless, it is still claimed that the cost overruns caused by additional work are unreasonably high and indicate poor planning and project management by the client at the design stage of the project. This can be facilitated by the application of a realistic project cost planning model and the commitment of the project manager by the client.

### Conclusions

The experience of construction practice indicates that the construction costs are overrun frequently in construction projects. By analysing the structure of the works that lead to the construction cost overrun, this research has determined that the main cause lies in additional works as the result of client’s variation orders, which are not necessarily

---

**TABLE 4. Results for linear regression and neural networks for Model I and II (own studies)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Statistic</th>
<th>LR</th>
<th>GRNN</th>
<th>SVM</th>
<th>RBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>(R^2)</td>
<td>0.988</td>
<td>0.605</td>
<td>0.345</td>
<td>0.556</td>
</tr>
<tr>
<td></td>
<td>MAPE%</td>
<td>13.824</td>
<td>35.130</td>
<td>14.814</td>
<td>10.976</td>
</tr>
<tr>
<td>II</td>
<td>(R^2)</td>
<td>0.950</td>
<td>0.398</td>
<td>0.222</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>MAPE%</td>
<td>18.700</td>
<td>68.858</td>
<td>18.291</td>
<td>13.407</td>
</tr>
</tbody>
</table>

**TABLE 5. Results for linear regression and neural networks for Model V and VI (own studies)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Statistic</th>
<th>LR</th>
<th>GRNN</th>
<th>SVM</th>
<th>RBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>(R^2)</td>
<td>0.994</td>
<td>0.952</td>
<td>0.992</td>
<td>0.875</td>
</tr>
<tr>
<td></td>
<td>MAPE%</td>
<td>0.522</td>
<td>1.268</td>
<td>0.529</td>
<td>1.267</td>
</tr>
<tr>
<td>VI</td>
<td>(R^2)</td>
<td>0.990</td>
<td>0.928</td>
<td>0.992</td>
<td>0.963</td>
</tr>
<tr>
<td></td>
<td>MAPE%</td>
<td>0.643</td>
<td>1.320</td>
<td>0.578</td>
<td>1.022</td>
</tr>
</tbody>
</table>
needed but are carried out at the request of the investor after signing the construction contract. Additional works account for 97.83% of the total cost overruns, thus confirming the research hypothesis. This is the result of poor planning by the client at the conceptual stage of the project.

Adequate planning of the project budget using the appropriate model will undoubtedly help to reduce construction cost overruns. A more accurate budget estimate contributes to the successful implementation of the project and is one of the construction sustainability criteria. Therefore, it is justified and important to keep exploring the possibilities of the different methods of cost forecasting. In this research, linear regression analysis and several “soft computing” methods have been applied to source data and natural logarithms. The natural logarithms of the variables were modelled on the basis of Bromilow’s time–cost model, i.e. by hybrid modelling (by using linear regression and “soft computing” methods). The results of the analysis indicate that better results have been achieved by using hybrid modelling comparing to source data. Despite the small database, the assumption that hybrid models and the use of logarithmic data can provide better indicators of the accuracy of cost prediction models has been confirmed, thus confirming the research hypothesis. Given that it has been affirmed several times (Petrusheva et al., 2016; Petrusheva et al., 2017; Petrusheva et al., 2019), it is recommended to try to model the cost-time relationship in construction projects in the way presented.

Excluding the value of additional work from the total construction costs has not improved the accuracy of the design model. However, this does not mean that these works are justified, but rather that they should be avoided with as much planning as possible in the conceptual phase of the project.

Acknowledgements

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

References


Car-Pušić, D. & Mladen, M. (2020). Early stage construction cost prediction in function of project sustainability. In 15th International Conference on Durability of Building Mater-
Summary

Predicting buildings construction cost overruns on the basis of cost overruns structure. In construction practice, contractually agreed costs are often exceeded, which interferes with the sustainable realization of construction projects. The research described in this paper covers 24 new construction, renovation and reconstruction projects in the Republic of Croatia realized in the years 2006 to 2017, in order to analyse the occurrence of cost overruns more precisely with regard to the source of the overruns. It was found that additional work is the main source of cost overruns: firstly, additional work as a result of the client’s change orders and then unforeseen construction work as a result of unforeseen circumstances. As for the additional works, they are carried out at the client’s request and are not necessary for the safety and stability of the building. Using linear regression and “soft computing” methods, the possibility of modelling the relationship between contractually agreed and realized construction costs with satisfactory accuracy was tested. The model with the values of the natural logarithms of the variables, modelled according to the time–cost model of Bromilow, proved to be of the highest accuracy.

Authors’ address:
Diana Car-Pušić
(https://orcid.org/0000-0003-2555-335X)
University of Rijeka
Faculty of Civil Engineering
Radmile Matejčić, 3, HR51000, Rijeka
Croatia
e-mail: diana.car.pusic@uniri.hr