Site Overhead Cost Index Prediction Using RBF Neural Networks

Michał JUSZCZYK¹,a,*, Agnieszka LEŚNIAK²,b

¹,²Cracow University of Technology, Faculty of Civil Engineering, Poland

ᵃmjuszczyk@izwbit.pk.edu.pl, ᵇalesniak@izwbit.pk.edu.pl

*Corresponding author

Keywords: Site Overhead Cost, RBF Neural Networks, Construction Cost Management.

Abstract. Cost estimation is one of the key tasks in the process of construction project management. Total costs incurred during the construction stage of the project by a contractor include direct costs that are related to works execution and indirect costs that accompany the delivery. This paper presents Artificial Neural Network (ANN) approach for prediction index of site overhead cost which is significant part of indirect cost. Applicability of Radial Basic Function (RBF) networks was investigated. A quantitative study on the factors conditioning site overhead costs of polish construction projects was completed. Moreover actual site overhead costs incurred by enterprises during project implementation were investigated. This research phase resulted in completion of a data set which covered 143 real-life cases of building projects. On the basis of the neural modelling the authors stated that the RBF networks can be a promising solution in the regression problem of site overhead cost index prediction.

Introduction

The building contractor’s overhead costs are divided into two categories: site overhead costs and company’s overhead costs [7]. Site overhead costs include items that can be identified with a particular job, but not materials, labour, or production equipment. Company’s overhead costs are items that represent the cost of doing business and often are considered as fixed expenses that must be paid by the contractor. The majority of contractors in Poland calculate overhead costs jointly, that is company’s (general) overhead costs and project overhead costs, as a percentage index from labour and equipment [8]. Contractors apply the method by convention, since it is quick and convenient; however, cost estimation it computes is low. Yet the issue of sufficiently reliable overheads estimation is vital for the potential contractor. The influence of improper calculation of the overhead costs can be significant for the financial situation of the company [8].

Artificial neural networks (ANN) belong to a dynamically developing field of knowledge involved in a number of scientific domains. ANNs indeed make a good tool to solve the problems involved in classification, prediction or regression [6]. Due to the structure imitating biological neural networks they possess unusual features, including [6]: ability to learn and generalize the acquired knowledge, ability to adapt to changing conditions and small sensitivity to errors in the input data.

The ANN model for estimating construction projects site overhead cost has been already proposed in [2]. This study involved a survey that investigated the factors affecting project's site overhead cost for building construction projects in Egypt. A Neural Network model was developed through fifty-eight experiments for predicting the percentage of site overhead costs. The model was based on MLP network. This model consists of one input layer with ten neurons (nodes) and one hidden layer having thirteen hidden nodes with a sigmoid transfer function and one output layer. With increasing frequency, construction cost management problems employ the artificial intelligence [3,4,8,9].

The aim of this article is to propose a model based on the artificial neural networks involving Radial Basis Functions (RBF) for prediction site overhead cost index of the construction projects. The RBF (Radial Basic Function) networks which belong to a special group of neural networks reveal certain advantages when compared with other networks, such as a Multi-Layer Perceptron (MLP) [6]:
- First, RBF can model any non-linear function by means of a single latent layer, which eliminates the need to decide on the number of latent layers.
- Second, a simple linear transformation performed in the output layer may be fully optimized by means of traditional techniques of linear modelling which are quick and do not involve such problems as local minima, found in MLP training.

For this reason RBF networks can be trained within a considerably short period of time.

The General Form of the Model

Regression analysis of the site overhead cost index with use of RBF neural networks required appropriate set of training data. A quantitative study on the determination of factors conditioning site overhead costs of polish construction projects was undertaken. The study resulted in determination of ten factors influencing sited overhead costs. To collect a data set useful in the process of networks training, a survey was conducted among Polish contractors. As a result of these studies 143 real-life cases of building projects was collected. The research included quantitative studies of the factors proposed, influencing site overhead costs in relation to the construction works under analysis and the actual site overhead costs of the projects. The set of 10 factors that influenced the site overhead cost has been adopted as input parameters of the neural network model (Table 1).

Table 1. The input parameters of the neural network model.

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>works type</th>
<th>general construction works,</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_2$</td>
<td>works type</td>
<td>installation works,</td>
</tr>
<tr>
<td>$X_3$</td>
<td>works type</td>
<td>engineering works,</td>
</tr>
<tr>
<td>$X_4$</td>
<td>construction site localization</td>
<td>in city centre,</td>
</tr>
<tr>
<td>$X_5$</td>
<td>construction site localization</td>
<td>outside the city centre,</td>
</tr>
<tr>
<td>$X_6$</td>
<td>construction site localization</td>
<td>non-urban spaces,</td>
</tr>
<tr>
<td>$X_7$</td>
<td>distance between the construction site and the company’s office,</td>
<td></td>
</tr>
<tr>
<td>$X_8$</td>
<td>works implementation time,</td>
<td></td>
</tr>
<tr>
<td>$X_9$</td>
<td>relations between the amount of works performed in winter to the total amount of works,</td>
<td></td>
</tr>
<tr>
<td>$X_{10}$</td>
<td>relations of the amount of works performed by subcontractors to the total amount of works.</td>
<td></td>
</tr>
</tbody>
</table>

The primary assumption was a covert implementation of the functional dependency $F$ by the artificial neural network. The prediction of the general construction index (the value of the variable described of the model), as represented by formula (1):

$$\hat{y}^{(p)} = F (x^{(p)})$$

where:
- $y^{(p)}$ – function $F$ value (predicted value of the general construction costs index) for the p-th vector of describing variables,
- $F$ – correlation function binding the describing variables and the described variable of the model implemented by the neural network,
- $x^{(p)}$ – p-th vector of describing variables $x_j$.

The authors proposed to establish the value of the described variable of the model $Y$, that is overhead site costs index, denoted hereinafter by $OSC_{ind}$, on the basis of the following formula:

$$Y = OSC_{ind} = \frac{OSC}{LC + MC + EC + SC}$$

where:
- $Y$ – described variable of the model,
- $OSC_{ind}$ – overhead site costs index,
- $OSC$ – overhead site costs observed in reality,
Train a Neural Network

This paper presents the use of supervised learning that is typically applied for a regression problem in which the value of the dependent variable has been observed and recorded in the learning sample, which allows to construct a prediction model (on a new testing sample). Training a RBF network is divided into two separate stages. During the first stage centres and divergences of radial neurons are determined; subsequently, in the second stage, the output layer is optimized [6].

The set of training data was divided into:
- the learning subset L – 60% of cases
- the validation subset V – 20% of cases,
- the testing subset T – 20% of cases.

The selected records of the database of teaching information, that is the values of describing variables \( x^{(p)}_1, \ldots, x^{(p)}_{10} \) and the values of the described variable \( y^{(p)} \), are presented in Table 2.

Table 2. The chosen records of the database of teaching information with the values of describing and variables.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( x^{(p)}_1 )</th>
<th>( x^{(p)}_2 )</th>
<th>( x^{(p)}_3 )</th>
<th>( x^{(p)}_4 )</th>
<th>( x^{(p)}_5 )</th>
<th>( x^{(p)}_6 )</th>
<th>( x^{(p)}_7 )</th>
<th>( x^{(p)}_8 )</th>
<th>( x^{(p)}_9 )</th>
<th>( x^{(p)}_{10} )</th>
<th>( y^{(p)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.90</td>
<td>0.90</td>
<td>0.5</td>
<td>0.5</td>
<td>0.05</td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
<td>0.10</td>
<td>0.1</td>
<td>0.1</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Source: own study

Network training began with involving 2 neurons in the hidden layer. Their number successively increased up to 8. For the purposes of training process subsets L, V, T were sampled 20 times. In total, 140 networks were trained, of which 8 networks that generated the most stable training results in every group were chosen.

Evaluation of the Model

The main aim was to construct a network that would be able to generalize and to operate well on a set of data that was excluded from training. Model evaluation concerned the efficiency of regression expressed as a number of correctly identified cases for the learning, validation and testing sets.

From the analyzed networks those were selected which yielded the most stable learning results. The quality of the network work was established on the basis of the specified values of learning and teaching errors: root mean square error, RMSE (3) and mean average percentage error MAPE (4). Such networks were searched for which the dispersion errors was as small as possible to indicate the networks the activity of which would not be laden with the type of drawing sets L, V and T.

\[
MAPE = \frac{1}{N} \sum_{p} \left( \frac{y^{(p)} - \hat{y}^{(p)}}{y^{(p)}} \right) \cdot 100\% \tag{3}
\]

\[
RMSE = \sqrt{\frac{1}{N} \sum_{p} \left( \frac{y^{(p)} - \hat{y}^{(p)}}{y^{(p)}} \right)^2} \tag{4}
\]

where:
- \( y^{(p)} \) – known values of variable \( Y \) for \( p \)-th sample (expected values of artificial network output),
\( \hat{y}^{(p)} \) – predicted values of variable \( Y \) for \( p \)-th sample (values predicted by the artificial network),

\( p \) – index of learning sample (for \( p \in L \)) or validating sample (for \( p \in V \)) or testing sample (for \( p \in T \)),

\( N \) – cardinality of the learning subset - \( L \) or validating subset - \( V \) or testing subset - \( T \).

The choice of the best network depended on lowest \( \text{RMSE} \) and \( \text{MAPE} \) errors obtained in the training process and the lowest dispersion of the \( \text{RMSE} \) errors between the learning, validation and testing subsets. \( \text{RMSE} \) errors are depicted in the figure 1. \( \text{RMSE} \) and \( \text{MAPE} \) errors are set together in the table 3.

![Figure 1. RMSE errors according to the RBF network structure. Source: own study.](image)

Table 3. Comparison of errors (\( \text{RMSE} \) and \( \text{MAPE} \)) for networks learning, validation and testing.

<table>
<thead>
<tr>
<th>Network type</th>
<th>RMSE L</th>
<th>RMSE V</th>
<th>RMSE T</th>
<th>MAPE L</th>
<th>MAPE V</th>
<th>MAPE T</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF 10-2-1</td>
<td>0,0654</td>
<td>0,0636</td>
<td>0,0544</td>
<td>62,7%</td>
<td>43,5%</td>
<td>43,7%</td>
</tr>
<tr>
<td>RBF 10-3-1</td>
<td>0,0601</td>
<td>0,0693</td>
<td>0,0520</td>
<td>52,8%</td>
<td>50,7%</td>
<td>32,1%</td>
</tr>
<tr>
<td>RBF 10-4-1</td>
<td>0,0363</td>
<td>0,0430</td>
<td>0,0332</td>
<td>27,0%</td>
<td>24,8%</td>
<td>33,0%</td>
</tr>
<tr>
<td>RBF 10-5-1</td>
<td>0,0607</td>
<td>0,0424</td>
<td>0,0349</td>
<td>49,7%</td>
<td>29,9%</td>
<td>31,0%</td>
</tr>
<tr>
<td>RBF 10-6-1</td>
<td>0,0466</td>
<td>0,0669</td>
<td>0,0533</td>
<td>41,2%</td>
<td>37,0%</td>
<td>42,7%</td>
</tr>
<tr>
<td>RBF 10-7-1</td>
<td>0,0606</td>
<td>0,0524</td>
<td>0,0444</td>
<td>53,7%</td>
<td>49,0%</td>
<td>43,2%</td>
</tr>
<tr>
<td>RBF 10-8-1</td>
<td>0,0536</td>
<td>0,0546</td>
<td>0,0503</td>
<td>45,7%</td>
<td>33,6%</td>
<td>35,2%</td>
</tr>
</tbody>
</table>

Source: own study.

The best results of training, taking into consideration the assumed evaluation criteria, were obtained for the RBF 10-4-1 network which includes 4 neurons in its hidden layer. As it can be found in the figure 1 and the table 3 \( \text{RMSE} \) and \( \text{MAPE} \) errors calculated for the learning, validating and testing subsets for RBF 10-4-1 are low in comparison to other investigated networks. On the other hand the dispersion of the errors is significantly low.
Figure 2. Scatter plot of the testing “T” and validating “V” (on the left); and learning “L” (on the right) results for the chosen RBF 10-4-1 network. Source: own study.

In the figure 2 a scatter plot of the learning, testing and validating results for the chosen RBF 10-4-1 network is presented. Coordinates of the points represent expected output values $Y$ and predicted output values $\hat{Y}$. The composition of the points in the graph is satisfying in most of cases. Most of the points corresponding to the testing - T, validating - V (on the left side of the graph) and learning - L (on the right side of the graph) are located in the cone of error ±25%.

Summary

The paper presents a model based on an artificial neural network designed to support estimation of site overhead cost index. The modelling performed here involved training, validation and testing the neural networks of radial basis functions, assuming a varying number of neurons in the hidden layer. The evaluation of the model was satisfactory. The best network proved to be the RBF 10-4-1 which generated the lowest RMSE and MAPE errors in the process of training. The results obtained proved applicability of RBF network in the prediction of site overhead cost. In the further research the authors will consider application of other types of artificial neural networks in the problem.

References


