Using Principal Component Analysis with a Back-Propagation Neural Network to Predict Industrial Building Construction Duration

Sou-Sen Leu¹ and Chi-Min Liu²

Key words: construction duration, prediction, principal component analysis (PCA), artificial neural network (ANN).

ABSTRACT

Industrial businesses must respond efficiently to market demands; therefore, industrial construction must accurately predict the project duration at the pre-investment stage. In practice, project duration predictions rely on the experience of project managers. To provide impartial expertise and quantitative estimate the predicted duration of constructing an industrial building, an extensive history of industrial building cases were collected to form a database. Principal component analysis was applied to the database to identify key factors to serve as input data for a back-propagation neural network (BP-NN) that was used to estimate the project duration. Three prediction models were identified and developed separately based on the total cost for large, medium, and small construction projects. The derived BP-NN prediction models are applicable for estimating construction duration during the initial stages of a project.

I. INTRODUCTION

Because the industrial marketplace is subject to rapid change of new competition, an accurate and expedient forecast of the amount of time required to construct a building is critical because it enables business to remain competitive. For example, the building construction cost of 12-in fly ash brick is only 5% of the total project cost—the remaining 95% of the cost includes equipment, installation, test runs, operation, and other factors. In addition, monthly sales constitute approximately four times the cost of constructing a building, creating an even greater incentive to complete construction on schedule. Ultimately, time is the principal concern of an industrial construction project. This study proposes a methodology for predicting the duration of industrial building construction projects that involves using principal component analysis (PCA), a back-propagation neural network (BP-NN), and a database containing 50 years of records of petrochemical industrial construction in Taiwan. The research scope of this study was limited to predicting industrial building construction duration, and the time requirements of equipment purchases, installation, test runs, and operation were excluded from the analysis.

II. PREDICTION ON PROJECT DURATION

Duration prediction has been extensively studied in numerous fields including management science (Yang et al., 2003), security inspection (Ding et al., 2003), medical research (Kelly, 2002), trade analysis (Goulielmos and Siropoulou, 2006; Huang et al., 2010), natural events (Monton and Kierland, 2006), and supplier selection (Jaskowski et al., 2010; Lam et al., 2010). Prediction methods can be classified into two categories: bottom-up methods and top-down approach methods. Table 1 shows a comparison between these methods. Bottom-up methods involve considering orders, resources, and the duration of each task in a construction project. To apply bottom-up methods, a skilled engineer’s experience and attention to detail regarding the design are required for an accurate schedule prediction. Bromilow (1969) indicated that only 12.5% of cases are completed on schedule, 40% are completed late, and 47.5% are completed before scheduled. Factors of uncertainty that can affect construction duration include the engineer’s experience, a contractor’s skill level, weather, economic conditions, price changes, and project alterations; additionally, a detailed design requires a substantial amount of time to prepare. Despite these uncertainties, accurately estimating construction duration is still crucial during the early stage of a project in professional practice.

Top-down approaches start from the case study of construction projects, decompose the relative factors, and build a reliable model based on information in a history database. These approaches can be used to estimate construction duration directly by integrating known project information into an artificial intelligence algorithm. Statistical and heuristic methods have been widely applied in bottom-up approaches. Hojat Adeli (2001) thoroughly reviewed artificial neural network (ANN) applications in civil engineering, structural engineering, and...
Combining PCA and ANN to forecasting models has been studied in several fields (Jan, 2003; Ran et al., 2004; Wang et al., 2009; Ma et al., 2011), but not to industrial construction. The current study adopted a top-down approach by combining PCA with ANN to estimate the duration of constructing industrial buildings. To develop a practical model, a database containing 50 years of history data was used. Various factors were analyzed, such as location, weather, price variation, the number of design changes, and contractor skill level. The proposed method enables directly estimating construction duration at the early stage of an industrial construction project. Although the proposed method cannot entirely replace the detailed planning involved in estimating a construction period, the results can serve as a critical reference for signing contracts and managing operational strategies.

### Table 1 Comparison between duration prediction methods

<table>
<thead>
<tr>
<th></th>
<th>Bottom-up Methods</th>
<th>Top-down Methods</th>
<th>Statistical Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Schelling Method</strong></td>
<td>Schelling methods used in Program Evaluation and Review Technique (PERT), MS Project, P3, Costchart, Critical Path Method, Probability Network Evaluation Technique (PNET) etc.</td>
<td>Role analysis (considering probability distribution and cumulative probability) with uncertainty</td>
<td>Linear regression [Montan and Earnest (2019)]</td>
</tr>
<tr>
<td><strong>Database</strong></td>
<td>DELPHI method [Expert opinion collection] [Lin (2002)]</td>
<td>Case-based reasoning [CBO] [Yao and Yang (1999)]</td>
<td></td>
</tr>
<tr>
<td><strong>Optimization techniques such as TASM, Search [Zhang (2002)], Genetic Algorithms [Lin (2003)], etc.</strong></td>
<td>Expert opinion [Yao (2003)]</td>
<td>Construction Database [Lin (2005)]</td>
<td></td>
</tr>
<tr>
<td><strong>Computation</strong></td>
<td>Forecast model [Zuo and Sim (2004)]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. 1. Based on duration of each task in a construction project...
2. 2. Recorded by the finished time of design and the acquisition of equipment...
3. 3. The task order is changed at the construction location...
4. 4. The probability obtained from simulation must have active data to provide solid software for decision making...
5. 5. External conditions cannot be considered such as weather, political factors, material prices, contractor's skill level...

**III. METHODOLOGY**

Figure 1 presents a flowchart depicting the methodology employed in this study. First, a set of target cases and factor selection criteria were selected, which are explained in the subsequent section. In PCA, direct oblimin rotation was used to determine the critical principle components of the selected factors. The obtained principle components and construction duration records from the history database were used to train the BP-NN model. The objective function denotes the difference between the true value and estimated result. The stopping criteria of the algorithm were the number of iterations and mean square error (MSE). The construction duration is presented as the construction cost per day (NT$ /day). This general representation is applicable to construction projects of various scales.

**1. Database**

The database comprised more than 20,000 cases of industrial building construction. The selection criteria were that the construction duration was longer than 6 months and no missing data. After filtering the construction data according to these requirements, 1,538 cases were identified. Table 2 shows the construction types and contractor classification of these cases.

Although numerous factors may affect construction duration, this study first classified a set of major categories, and then selected the corresponding representative factors from each category. All of the selected factors were quantified to facilitate conducting a scientific analysis and developing a forecasting model.

Recent studies (Lin, 2005) have indicated that critical factors include constructability, workspace acquisition, learning curve, weather, supervision efficiency, building type, contract systems, management effectiveness, district environment, and financial issues. The present study classified the factors into four categories: case type, participant, location, and time. Each category contains additional descriptive factors that could be directly obtained from the studied database. By contrast, factors requiring further calculation or were obtained from another da-
Table 2 Basic information of the selected cases

<table>
<thead>
<tr>
<th>Construction Company</th>
<th>Total</th>
<th>Level A</th>
<th>Level B</th>
<th>Level C</th>
<th>Level D</th>
<th>Level E</th>
<th>Level F</th>
<th>Level G</th>
<th>Level H</th>
<th>Level I</th>
<th>Level J</th>
<th>Level K</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA_ Decoration works style</td>
<td>264</td>
<td>112</td>
<td>40</td>
<td>13</td>
<td>34</td>
<td>7</td>
<td>11</td>
<td>0</td>
<td>17</td>
<td>34</td>
<td>3</td>
<td>12</td>
<td>69</td>
</tr>
<tr>
<td>NA_ Steel Structure works style</td>
<td>138</td>
<td>110</td>
<td>47</td>
<td>13</td>
<td>24</td>
<td>4</td>
<td>11</td>
<td>2</td>
<td>14</td>
<td>23</td>
<td>2</td>
<td>60</td>
<td>74</td>
</tr>
<tr>
<td>NA_ Reinforce Concrete structure works style</td>
<td>405</td>
<td>146</td>
<td>74</td>
<td>26</td>
<td>44</td>
<td>7</td>
<td>31</td>
<td>9</td>
<td>17</td>
<td>36</td>
<td>2</td>
<td>14</td>
<td>34</td>
</tr>
<tr>
<td>NA_ Miscellaneous works style</td>
<td>155</td>
<td>47</td>
<td>28</td>
<td>14</td>
<td>31</td>
<td>6</td>
<td>11</td>
<td>4</td>
<td>7</td>
<td>19</td>
<td>6</td>
<td>17</td>
<td>34</td>
</tr>
<tr>
<td>NA_ Electrical construction works style</td>
<td>126</td>
<td>47</td>
<td>24</td>
<td>8</td>
<td>11</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>2</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>NA_énergie works style</td>
<td>295</td>
<td>13</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>NA_ Engineering works style</td>
<td>153</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 3 Introduction of selected factors

<table>
<thead>
<tr>
<th>Direct Factors</th>
<th>Company Type</th>
<th>Case Type</th>
<th>Instruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA_ Decoration works style</td>
<td>NA_ Steel Structure works style</td>
<td>NA_ Reinforce Concrete structure works style</td>
<td>NA_ Miscellaneous works style</td>
</tr>
</tbody>
</table>

Notes:
1. Nine work types (WT) include AF_ Decoration works style, AH_ Steel Structure works style, AS_ Reinforce Concrete structure works style, AT_ Miscellaneous construction works style, AU_ backfilling and road pavement works style, HM_ Machinery and electrical construction works style, CF_ RC Foundation works style, CG_ Ground preparation works style, CB_ Harbor and preparation works style.
2. Seven contract types are: outsourcing-processing, outsourcing, turnkey, outsourcing appointment, outsourcing design, outsourcing design (type D), and procurement.
3. Contractor Level includes Level A general construction company, Level B 1st construction company, Level C 2nd construction company, Level D 3rd construction company, Level E material supplier, Level F general business company, Level G consulting company, Level H engineer incorporated company (Inc.), Level I other company (decoration, surveying, landscape gardening companies).
4. Locations include Taipei, Taichung, Tainan, Ilan, Linkou, Nantou, Taoyuan, Tzu Tzu, KaoHsiung, Keelung, Mailiao, Chiayi, Changhua, Shulin, and Guanyin.

2. Construction cost

Although construction costs are strongly related to construction duration (Figure 2), they are not considered a factor in this study because different construction cost levels involve distinct relationships between the factors and construction duration. To account for the influence of construction costs, the factors were classified into three categories based on the project scale (large-, medium-, and small-scale construction projects), and the training was performed separately for each model.

(1) Large-scale: Construction cost over NT$50 million, with a construction duration of more than 42 months (n = 183).
(2) Medium-scale: Construction cost between NT$10 million and NT$50 million dollars, with a construction duration of 12 to 42 months (n = 399).
(3) Small-scale: Construction cost below NT$10 million, with a construction duration of 6 to 18 months (n = 956).

3. Principal component analysis

PCA is a statistical method for converting potentially correlated variables of observation data into a set of linearly uncorrelated variables called principal components. The dimension of a principle component is equal to or less than the dimension of the original variable. The selected principle components can be used as input data for BP-NN model training. Because the original variables may be correlated, direct oblimin rotation is used for obtaining a non-orthogonal (oblique) solution, resulting in higher eigenvalues but diminished interpretability of the variables (Chen 2005). In the present study, PCA was conducted using SPSS Version 17. Figure 3 shows the output from PCA, wherein two principal components were obtained from analyzing five factors.

PCA was performed separately for each construction project scale.

A. Large-scale construction project: Three sets of principal components were selected.

(1) Two principal components, denoted as M11 and M12, were derived from the following five original variables: supervisor Su, seniority Se, number of stuffs NS, contractor level CL, and recent revenue RR. The two following components explain 69.155% of the variance:

\[ M_{11} = -0.227Su + 0.067Se + 0.474NS - 0.014CL + 0.460RR \]  

\[ M_{12} = -0.087Su + 0.594Se + 0.025NS - 0.591CL + 0.064RR \]

(2) Two principal components, denoted as T11 and T12, were derived from the following three original variables: start season SS, duration Du, and number of design changes Nc. The following two components explain 78.916% of the variance:

\[ T_{11} = 0.001SS + 0.603Du + 0.612Nc \]

\[ T_{12} = 0.978SS - 0.127Du + 0.126Nc \]
(3) Other variables include location $L_o$, work difficulty $W_D$, price index $P_I$, and work type $W_T$.

B. Medium-scale construction project: Three sets of principal components were selected.

(1) Two principal components, denoted as $O_{11}$ and $O_{12}$, were derived from the following three original variables: contract type $C_T$, work type $W_T$, and price index $P_I$. The following two components explained 57.563% of the variance:

$$O_{11} = 0.185C_T + 0.614W_T - 0.145P_I$$

$$O_{12} = 0.586C_T + 0.028W_T + 0.737P_I$$

(2) Two principal components, denoted as $M_{21}$ and $M_{22}$, were derived from the following three original variables: supervisor $S_u$, inspector $I_D$, number of stuffs $N_S$, and seniority $S_e$. The following two components explained 63.963% of the variance:

$$M_{21} = 0.064S_u - 0.069I_D + 0.598N_S + 0.607S_e$$

$$M_{22} = 0.684S_u + 0.606I_D + 0.006N_S + 0.008S_e$$

(3) Two principal components, denoted as $T_{21}$ and $T_{22}$, were derived from the following three original variables: start season $S_S$, duration $D_u$, and work difficulty $W_D$. The following two components explained 70.174% of the variance:

$$T_{21} = 0.763S_S - 0.064D_u + 0.593W_D$$

$$T_{22} = -0.237S_S + 0.851D_u + 0.428W_D$$

(4) Other variables include recent revenue $R_R$, capital $C_a$, number of design changes $N_C$, and location $L_o$.

C. Small-scale construction project: One set of principal components was selected.

(1) Two principal components, denoted as $T_{31}$ and $T_{32}$, were derived from the following four original variables: start season $S_S$, price index $P_I$, work difficulty $W_D$, and duration $D_u$. The following two components explained 88.869% of the variance.

$$T_{31} = -0.005S_S + 0.024P_I + 0.98W_D + 0.36D_u$$

$$T_{32} = 1.000S_S - 0.001P_I + 0.001D_u$$

(2) Other variables include recent revenue $R_R$, and number of stuffs $N_S$.

The results obtained from PCA (Table 4) indicated that several groups of variable sets can be classified according project scale (i.e., large, medium, and small) and factor type (i.e., participant, case type, time, and location).

### Table 4 Classification of principal components and others var. for BP-NN model training

<table>
<thead>
<tr>
<th>Scale of the construction</th>
<th>Principal component</th>
<th>Case type</th>
<th>Duration</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large scale</td>
<td>10 $M_{21}$ and $M_{22}$</td>
<td>$C_{large}$</td>
<td>$T_1$, $T_2$, and $T_3$</td>
<td>$C_{large}$</td>
</tr>
<tr>
<td>50 million or more dollars</td>
<td></td>
<td></td>
<td>$S_u$, $S_e$, $R_R$, $C_l$, $R_f$</td>
<td>$C_{large}$</td>
</tr>
<tr>
<td>(duration is greater than 3.5 years)</td>
<td></td>
<td></td>
<td>$D_u$, $W_D$</td>
<td>$C_{large}$</td>
</tr>
<tr>
<td>Medium scale</td>
<td>10 $M_{21}$ and $M_{22}$</td>
<td>$C_{medium}$</td>
<td>$T_1$, $T_2$, and $T_3$</td>
<td>$C_{medium}$</td>
</tr>
<tr>
<td>10 million and 50 million</td>
<td></td>
<td></td>
<td>$S_u$, $S_e$, $S_I$, $D_u$, $I_D$, $C_l$</td>
<td>$C_{medium}$</td>
</tr>
<tr>
<td>dollars (duration is between 1 and 3.5 years)</td>
<td></td>
<td></td>
<td>$S_u$, $S_e$, $D_u$, $I_D$, $C_l$</td>
<td>$C_{medium}$</td>
</tr>
<tr>
<td>Small scale</td>
<td>10 $R_R$</td>
<td>$C_{small}$</td>
<td>$T_1$</td>
<td>$C_{small}$</td>
</tr>
<tr>
<td>Less than 10 million dollars</td>
<td></td>
<td></td>
<td>$S_u$, $D_u$, $R_R$, $C_l$, $R_f$</td>
<td>$C_{small}$</td>
</tr>
<tr>
<td>(duration is between 6.5 and 1.5 years)</td>
<td></td>
<td></td>
<td>$S_u$, $D_u$, $R_R$, $C_l$, $R_f$</td>
<td>$C_{small}$</td>
</tr>
</tbody>
</table>

Notes:

Some factors have underline as NS・RR that not produced from PCA process but include in the BP-NN model. For the reason have two:

1. the purpose of PCA process is reduce the number of variables, but the NS and RR in the same attribute field just two variables, no need to do the process.
2. NS and RR are affect the duration indeed based on the domain knowledge, so we must be join these variables in the model to check the influence to the duration.

We set the price index as a variable to avoid waste time of calculation in the future (different years).
4. Back-Propagation Neural Network (BP-NN)

Figure 4 illustrates the calibration of the BP-NN model from using the obtained principal components as input data. The calibrated BP-NN model was used to predict the construction duration in units of construction cost per day (NT$/day). NeuroSolutions is used for the BP-NN model development. The structure of the ANN model features a single hidden layer based on the back-propagation approach, which is a supervised learning network. In the input layer, the number of neurons was equal to the number of principal components. The activity function adopted a summation function, which was a weighted summation of the neuron output from the preceding layer. The steepest descent method was used to determine the optimal solution, which is an optimal weighting matrix. The hyper tangent was selected as the transfer function.

The objective function is expressed in Eq. (13):

\[ \text{Min} \sum_{i} \left[ (Y_i - \sum_{j} w_{ij} \times X_j)^2 \right] \]  

where \( Y_i \) is the target value, which is the construction duration obtained from the database; \( w_{ij} \) is the weight; and \( X_j \) is the output from the neuron output from the preceding layer.

Two stopping criteria were the number of iterations (5,000 runs) and MSE (< 0.05), which is expressed in Eq. (14):

\[ \text{MSE} = \frac{\sum_{i=0}^{N} \sum_{j=0}^{P} (d_{ij} - y_{ij})^2}{NP} \]

where \( d_{ij} \) is the output which result from the model operation; \( y_{ij} \) denotes the known construction cost per day (depend var.); and \( N \) and \( P \) denote the number of independent variables.

The model training was performed separately according to the project scale. Table 5 shows the calibrated model parameters, which are the weights of the hidden layer. The dimension of the weight matrix varies with the construction project scale. The MSE of the large-, medium-, and small-scale projects are 0.02–0.05, 0.06–0.10, and 0.08–0.11 respectively, where lower MSE values indicate more accurate calibration of the BP-NN model.

![Diagram of BP-NN model](image)

Table 5 The calibrated weights for hidden layer (*10-1)

<table>
<thead>
<tr>
<th>Database of Industrial Building</th>
<th>Output</th>
<th>Principal Components Xn</th>
<th>N ( \times ) ( \sum_{n} w_{ij} \times X_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MSE (Comparison between Target Value and The Model Output)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>also ( \sum_{i} \left[ (Y_i - \sum_{j} w_{ij} \times X_j)^2 \right] )</td>
</tr>
</tbody>
</table>

**Result and Discussion**

Fig 5 shows a comparison output between the predicted (a dotted line) and real construction duration. Although the calibrated model shows only an intangible statistical relationship between the construction duration and principal components, this study attempted to reveal the physical meaning that may exist behind the black box model. In addition to the project scale, the relationship between construction duration and (1) participant, (2) location, (3) time, and (4) case type are addressed.

1. Participant
   - Large-scale projects:
     - The major factors of the principal components are number of stuffs NS and recent revenue RR for \( M_{11} \), and seniority Se and contractor level CL for \( M_{12} \). All of these factors indicated that the capability of the contractors has the strongest influence on the construction duration. Moreover, \( M_{11} \) and \( M_{12} \) also have large weights in the BP-NN model, further indicating the importance of this variable. Therefore, contractor capability must be considered as a constraint in large-scale construction projects.
   - Medium-scale projects:
     - \( M_{21} \) shows that contractor capability is a major influence in the number of stuffs NS and seniority Se. In addition to contractors level CL, the supervisor Su and inspector ID play crucial roles in medium-scale projects.
   - Small-scale projects:
     - Although no principal factors are generated, the number of design NC may be crucial for larger cases; however NC was non-significant for medium- and small-scale projects.

2. Time:
   - Starting season SS was the major factor for all large-, medium-, and small-scale projects. According to additional analysis for SS, the projects starting in summer have a negative impact on construction duration. This might be attributed to the typhoon season, which can prolong the construction period. In addition to SS, the number of design changes NC may be crucial for larger cases; however NC was non-significant for medium- and small-scale projects.

3. Location:
The location $Lo$ had a marked influence on large- and medium-scale projects. For construction projects in Taipei City, the cost per unit of time tends to be higher than in other areas because of restrictive regulations, higher risk of damaging the areas surrounding a construction site, and higher costs for labor. For projects located in suburban areas, such as in reclaimed land areas, the cost per unit of time is relatively lower, which could be attributed to less stringent regulations and easier mobility of construction equipment.

(4) Case type:

Regarding the work type $WT$ and work difficulty $WD$ variables, the $WT$ has significant impact on large- and medium-scale projects. The $WT$ may imply the complexity of a project. Ranked in descending order, the construction costs per unit of time are harbor engineering, steel works, normal constructions, and structural engineering. The $WD$ variable had a marked impact for all project scales, indicating that the work difficulty directly influences the construction cost per unit of time.

---

**REFERENCES**


Chen, S. H. (2005). Principal component analysis class notes of power point files. Feng Chia University Transportation Technology and Management, Taichung, Taiwan.


Zhang, L. (2002). Table search method is applied to the project schedule / cost tradeoff optimization of research. Master's thesis, Department of Construction Engineering, Chaoyang University of Technology, Taichung, Taiwan.