The following publication Lee, J. G., Lee, H. S., Park, M., & Seo, J. (2022). Early-stage cost estimation model for power generation project with limited historical data. Engineering, Construction and Architectural Management, 29(7), 2599-2614 is available at https://doi.org/10.1108/ECAM-04-2020-0261.

Early Stage Cost Estimation Model for Large-scale Project with Limited Historical Data

ABSTRACT

PURPOSE

Reliable conceptual cost estimation of large-scale construction projects is critical for successful project planning and execution. To address the limited data availability in conceptual cost estimation, this study proposes an enhanced ANN-based cost estimating model that incorporates artificial neural networks, ensemble modeling, and factor analysis approach.

DESIGN/METHODOLOGY/APPROACH

In the ANN-based conceptual cost estimating model, ensemble modeling component enhances training and thus improve its predictive accuracy and stability when project data quantity is low; and the factor analysis component finds the optimal input for an estimating model, rendering explanations of project data more descriptive.

FINDINGS

Based on the results of experiments, it can be concluded that ensemble modeling and FAMD (Factor Analysis of Mixed Data) are both conjointly capable of improving the accuracy of conceptual cost estimates. The ANN model version combining bootstrap aggregation and FAMD improved estimation accuracy and reliability despite these very low project sample sizes.

RESEARCH LIMITATIONS/IMPLICATIONS

The generalizability of the findings is hard to justify since it is difficult to collect cost data of construction projects comprehensively. But this difficulty means that our proposed approaches and findings can provide more accurate and stable conceptual cost forecasting in the early stages of project development.

ORIGINALITY/VALUE

From the perspective of this research, previous uses of past-project data can be deemed to have underutilized that information, and this study has highlighted that – even when limited in quantity – past-project data can and should be utilized effectively in the generation of conceptual cost estimates.

INTRODUCTION

Economic-feasibility studies of construction projects conducted during projects' planning phases are vitally important, as early decisions have more significant consequences than later ones (Ahiaga-Dagbui 2012). In large-scale infrastructure projects requiring huge capital investment, inappropriate budgeting can lead to incorrect project scoping and poor execution, whether in the form of poor quality or excessive cost overruns, and thus to economic loss for both clients and contractors (Kirkham, 2014). Yet, despite the importance of accurate conceptual cost estimation for budgeting, construction projects frequently suffer from large discrepancies between their estimated budgets and actual costs, due to uncertain or insufficient project information in their earliest stages.

In their attempts to provide reliable conceptual cost estimation, researchers have developed mathematical techniques including stochastic, parametric, and capacity/equipment-factored estimation methods, all of which rely on historical project-cost data. Recently, such techniques have been combined with machine learning approaches heralding a new era of data-driven conceptual cost estimation (Sonmez, 2011). The machine learning approaches have been shown to be more accurate than previous cost-estimation methods – e.g., regression models – due to its superior ability to analyze the relationships among cost variables (Kim et al. 2004_a; Wang and Gibson, 2010). The performance of data-driven cost-forecasting models – whether machine learning-assisted or otherwise – relies on a high quantity and quality of historical project cost data, to ensure an adequate reflection of the project's characteristics. The amount of available historical data for large-scale projects, however, is generally limited for typical machine learning-based cost estimation models. This is because any engineering, procurement and construction (EPC) projects such as power-generation plants have few precedents, and tend to be anomalous in terms of both its scope and delivery, unlike general construction projects such as housing and office projects. It is also difficult to find common quantitative variables across such distinctive large projects, due to the great variety in their scales and the types of equipment used. Therefore, the applicability of machine learning approaches to large-scale projects is challenging as the limited historical cost data could lead to inaccurate and unreliable cost estimation.

To address these issues, this study proposes an enhanced ANN-based conceptual cost-forecasting model that incorporates ensemble modeling and factor analysis to maximize the analytical power of the limited historical data for large-scale projects. Its ensemble-modeling addresses the abovementioned

57 data limitations by training the forecasting model using randomly sampled multiple data subsets and
58 aggregation of its collective learning, while the factor-analysis seeks to extract more descriptive features
59 from available cost variables to improve the model's predictive analytics. To evaluate the predictive
60 performance of the proposed model in the case of limited project data, we collected actual cost data
61 from 86 large power-plant projects and estimated the project costs by intentionally limiting the number
62 of available past cost data samples for training the proposed model. Based on the results of the case
63 study, the validity and applicability of the proposed approach are discussed with its limitations, and
64 future research directions that could aid its further improvement.
65
66 ## MACHINE-LEARNING APPROACHES TO CONCEPTUAL COST ESTIMATION
67 Conceptual costs of construction projects are foundational to vital business decision-making,
68 including economic feasibility studies, project-alternatives evaluations, and long-term project budgets.
69 To be effective, such estimation should be based on an anticipated scope of work, which is typically
70 accompanied by some key milestones that need to be achieved. During the earliest phase of project
71 development, only the scope and location of the project are generally defined, whereas the master
72 schedule, integrated project plan, escalation strategy, work breakdown structure, etc., are either mere
73 approximations or not yet ready at all (Christensen and Dysert, 2005). At the same time, the historical
74 project data with detailed information is difficult to be obtained due to a lack of project experience and
75 an understandable reluctance to share information between construction companies (Liu and Ling,
76 2005). Due to the limited information availability during the early stages of a project, estimators or
77 construction managers typically leverage their prior knowledge and experience, and the estimated cost
78 is likewise largely dependent on information about past projects. To maximize the practical value of
79 the available historical information, a considerable amount of recent construction-economics research
80 has been devoted to developing data-driven cost-estimation models (Elfaki et al. 2014). Table 1
81 summarizes major studies since the 2000s when the data-driven cost estimation approaches became a
82 prominent topic. Many approaches using statistical techniques have been deployed to analyze patterns
83 within project-cost databases. Regression analysis has been used to find relationships between project
84 variables and its costs (Lowe et al. 2006; Stoy et al. 2008; Mahamid, 2011). Such approaches, however,

85 are limited when dealing with non-linear and complex problems, since linear methods assume linear
86 relationships between cost variables, and thus do not guarantee adequate representation of cost variables'
87 actual relationships within complex project data (Flood and Kartam, 1994; Kohavi, 1998).

88 Recently, machine learning approaches have been used as an alternative to regression analysis as
89 they are better capable of solving non-linear and complex problems. Machine learning has been used to
90 analyze voluminous and complicated project data, and then using its knowledge from past project data
91 to make predictions on not-yet-seen data (Setyawati et al. 2002). As shown in Table 1, many types of
92 algorithms have been found to be applicable for conceptual cost estimation; including case-based
93 reasoning (CBR), support vector machines (SVM), and ANN. These algorithms has been widely applied
94 to forecast the conceptual costs of highway projects (Wilmot and Mei, 2005; Pewdum et al. 2009),
95 water and sewer installations (Alex et al. 2010; Dominic and Smith, 2012) and building projects (Lowe
96 et al. 2006; Arafa and Alqedra, 2011), and in each of these cases produced better forecasts than either
97 parametric estimation models or linear regression. Several studies have also made direct performance
98 comparisons among multiple machine learning algorithms: with Emsley et al. (2002), Sonmez (2004),
99 and Wang and Gibson (2010) all testing the accuracy of linear regression against ANN, and Kim et al.
100 (2004_a) comparing the accuracy of linear regression, ANN, and CBR. While some of these researchers
101 concluded that there was no significant difference (Emsley et al. 2002; Sonmez, 2004), others argued
102 that certain machine learning algorithms such as ANN are more accurate than others, at least for
103 particular project data used (Kim et al. 2004_a; Wang and Gibson, 2010). Even though machine learning
104 approaches have shown better performance than other traditional ones for construction cost forecasting,
105 the accuracy and reliability of the cost forecasting models would heavily rely on the cost-relevant
106 information availability such as 1) the number of historical project data samples and 2) the number of
107 cost variables that can be obtained from the historical data (Gardner et al. 2016).
108

109 Table 1. Prior Literature on Data-driven Conceptual Cost Estimation

Literature Source	Project Characteristics			Number of Variables			Forecast Method
The first author (published)	Type	Country	P. No.	Total	NV	CV	
Kim, G.H. (2004_b)	residential	S. Korea	530	10	6	4	ANN

Wilmot, C.G. (2005)	highway	USA	1723	11	11	0	ANN
Lowe, D. J. (2006)	building	UK	286	6	6	0	MRA
Petroutsatou, C. (2006)	road tunnel	Greece	33	9	9	0	MRA
An, S. (2007)	residential	S. Korea	580	10	5	5	SVM
Stoy, C. (2008)	residential	Germany	70	6	6	0	RA
Alex, D.P. (2010)	water facility	Canada	804	14	7	7	ANN
Elkassas, E. M. (2009)	industrial	Egypt	115	15	11	4	ANN
Pewdum, W. (2009)	highway	Thailand	51	8	6	2	ANN
Koo, C. (2010)	residential	S. Korea	101	11	7	4	CBR
Arafa, M. (2011).	building	Israel	71	8	7	1	ANN
Ji, S. (2011)	military barrack	S. Korea	129	18	10	8	CBR
Mahamid, I. (2011)	public road	Palestine	131	10	10	0	MRA
Dominic, D. A. (2012)	water facility	UK	98	11	6	5	ANN
Jin, R. (2012)	residential	S. Korea	99	11	11	0	CBR
Choi, S. (2013)	public road	S. Korea	207	17	11	6	CBR
El-Sawalhi, N. I. (2014)	building	Palestinian	169	11	3	8	ANN
M. Gunduz (2015)	HEPP	Turkey	54	12	11	1	ANN
Dursun, O. (2016).	building	Germany	657	16	10	6	ANN
Hyari (2016)	public project	Jordan	224	4	0	4	ANN
Hashemi, S.T (2017)	power plant	Iran	39	9	3	6	ANN

Note. P. No = Number of previous projects on which data was obtained; NV = numerical variables; CV = categorical variables; HEPP; hydroelectric power plant; ANN = artificial neural network; SVM = support vector machine; CBR = case-based reasoning; MRA = multiple regression analysis; RA = regression analysis.

Limited Information Availability (1): Insufficient Project Samples

The data-driven conceptual cost-estimation approaches that have been introduced so far can all be described as empirical research using historical project data (Fellows and Liu, 2009). A conceptual cost-estimation model is developed from collected construction project data, according to the purpose and scope of the proposed forecasting. Such research assumes that similar projects have similar characteristics and outcomes, and solves problems by recognizing similarities between the past projects and new projects. Table 1 describes the number of historical project samples that were used in the previous studies on data-driven cost prediction. Even though it is difficult to determine the required number of historical data samples for the data-driven cost estimation, previous studies have usually relied on more than 100 historical cost data samples except for some studies. Hagan et al. (2014) argued that the data-driven cost prediction models need comprehensive data since data-driven methods are not potent addressing extrapolation, i.e., the predictive power can be weak for situations outside the training dataset. As such, enough data collection is required to ensure the ability of that data to reflect the target project's characteristics adequately and to ensure the forecasting accuracy of the algorithm. The

availability of the historical project data for large-scale projects such as power plants or infrastructure projects, however, tends to be limited as they involve a tremendous amount of capital investment and long project duration (Christensen and Dysert, 2005). For instances, previous studies to address large-scale projects such as the hydroelectric power plant (Gunduz and Sahin, 2015) and four types of power plant projects (Hashemi and Kaur, 2017), used only 54 and 39 project samples, respectively. Also, the studies for civil infrastructure projects, such as roads and tunnels, have relied on a relatively smaller number of historical data (i.e., 51 samples for highway projects (Pewdum et al. 2009) and 33 samples for tunnel projects (Petroutsatou et al. 2006)). Although abovementioned studies show the possibilities of ANN-based model with small datasets for several types of construction projects, the application of ANN-based models on small data samples may lead to the reliability issues such as lack of generalization or overfitting (Bishop, 2006).

To overcome the insufficient project data issues, multiple regression analysis (MRA) is known as an effective method as MRA requires less amount of data samples compared to machine-learning approaches (Green, 1991). Still, MRA is not able to capture complex relationships within cost-relevant variables due to its linear assumptions (Elmousalami, 2020). In the field of machine learning for other applications such as pilot tests for manufacturing systems (Tsai and Li, 2008), ensemble modeling were applied to address the limited training data issues. Ensemble modeling learn from multiple datasets via repetitive random sampling of observations from the population (e.g., bootstrap aggregation or adaptive boosting) or combine different types of algorithms simultaneously (known as stacking), aiming to improve the algorithm's stability and performance (Breiman, 1996; Efron and Tibshirani, 1994). When training data is insufficient, the bootstrap aggregation method has been utilized to improve the predictive performance of ANNs through repeating the process of resampling data from the sparse training data (Efron & Tibshirani, 1994; Tsai and Li, 2008). In the construction domain, however, few studies have investigated how to deal with the limited data samples for machine learning-based cost prediction models. Sonmez (2011) utilized bootstrap aggregation to quantify the uncertainty level of the estimated cost item rather than improving estimation accuracy. The stacking methods have been used to improve the accuracy of forecasting the numbers of project disputes (Chou and Lin, 2012), the

155 unit price bids of highway project (Cao et al. 2018) and the level of cost overrun (Williams and Gong,
156 2014).
157
158 ***Limited Information Availability (2): Constraints on Variables***
159 The performance of data-driven cost-forecasting models depends not only on the quantity of data
160 but also on the characteristics of the input variables. Indeed, the smaller the available data sample, the
161 more essential it becomes to identify the critical features of the input data (Huang, 2015). Therefore,
162 identifying cost variables from collected project data is the first and the most critical stage. As shown
163 in Table 1, the data-driven conceptual cost-forecasting models developed in previous studies used more
164 quantitative than qualitative variables, while those models that were regression-based used no
165 qualitative variables at all. In case of Hyari (2016), the ANN model is developed with only four
166 qualitative variables of public construction projects in Jordan but showed a relatively higher error rate
167 compared to other studies (the average accuracy percentage of 28.2% was achieved during the test).
168 The previous studies indicate that the effective modeling of large EPC projects cost will require a higher
169 proportion of qualitative variables – such as project region, project type, contract terms, and delivery
170 method – than the modeling of general building projects does. And these qualitative variables are
171 especially essential in the early phase of project development (Matel et al. 2019) when the quantitative
172 variables remain unknown and are subject to change as the project progresses due to changes in client
173 requirements, design changes, or unexpected engineering problems. According to the study by Matel
174 (2019), these qualitative variables may include project scope, work type, contract type, clients'
175 characteristics, managers' competitiveness, pre-design quality and market type. These variables are
176 difficult to be fully collected due to the uncertainty at the early project phase. Moreover, while the
177 capacity and price of individual pieces of equipment remain important variables for the conceptual cost
178 estimation of EPC projects, the scopes and ranges of those variables are too diverse across projects to
179 be used as common variables among them (Christensen and Dysert, 2005). Even projects aimed at
180 creating the 'same' type of facilities vary enormously in their requirements and scope across time,

regions, clients, delivery methods, and financial aspects (Firoozabadi et al. 2013); and the quantity and quality of the collectable information may also vary depending on its source.

Under the constraints of the limited information available at the early project phase, previous research efforts have tried to how to select more informative variables to enhance the analytic power of the cost prediction models. For example, Gardner et al. (2016) tried to identify the appropriate number of input variables for reasonable data-driven cost estimates. They reported that, although choosing informative variables can lessen data collection efforts, it does not guarantee cost-forecasting models' accuracy, as the lower explanatory power of input data with fewer attributes tends to impact their predictive performance negatively. In this regards, various feature-engineering methods have been applied to select influential attributes based on qualitative and quantitative approaches (Günaydın and Doğan, 2004). The qualitative approaches include researchers' discretion, interviews, and surveys with experts (Attalla and Hegazy, 2003). Delphi method, fuzzy logic, and analytic hierarchy process (ElMousalami et al. 2018) were applied to select cost drivers from the perspective of experts. The cost driver identification process based on the questionnaire or interview has a limitation of generalization as it depends on the experience of the survey participants. The quantitative approaches are also applied to select cost variables from the collected project cost variables. Regression analysis (Stoy et al. 2008), correlation analysis (Ranasinghe, 2000), and factor analysis (Akintoye, 2000) were used to determine influential cost variables. The factor analysis, such as principal component analysis (PCA) is an algorithm to identify the explanatory components of collected data and thus find the hidden relationship among variables. Regression methods, correlation methods and PCA have their limitations when it comes to selecting critical variables for qualitative or categorical variables, which – as we have seen – are more important than quantitative ones for conceptual cost estimation. Thus, the feature engineering method for comprehensively processing both qualitative and quantitative variables is necessary to develop more reliable and accurate conceptual cost estimation models.

RESEARCH METHODOLOGY

This study proposes a novel machine learning-based cost-forecasting model for large-scale projects that have few precedents and sketchy preliminary project information. This model incorporates an ANN, ensemble modeling, and factor analysis to address the pre-mentioned limitations of previous data-driven methodologies for dealing with insufficient project information. Specifically, the ANN learns the nonlinear relationships within the collected project data; ensemble modeling helps enhance such learning when the quantity of project samples is low; and factor analysis identifies the most informative features within input variables to improve the model's forecasting performance. The proposed concepts have been implemented through two main functional steps: 1) data preprocessing, including factor analysis, and 2) forecasting-model development, including ANN and ensemble modeling (Figure 1).

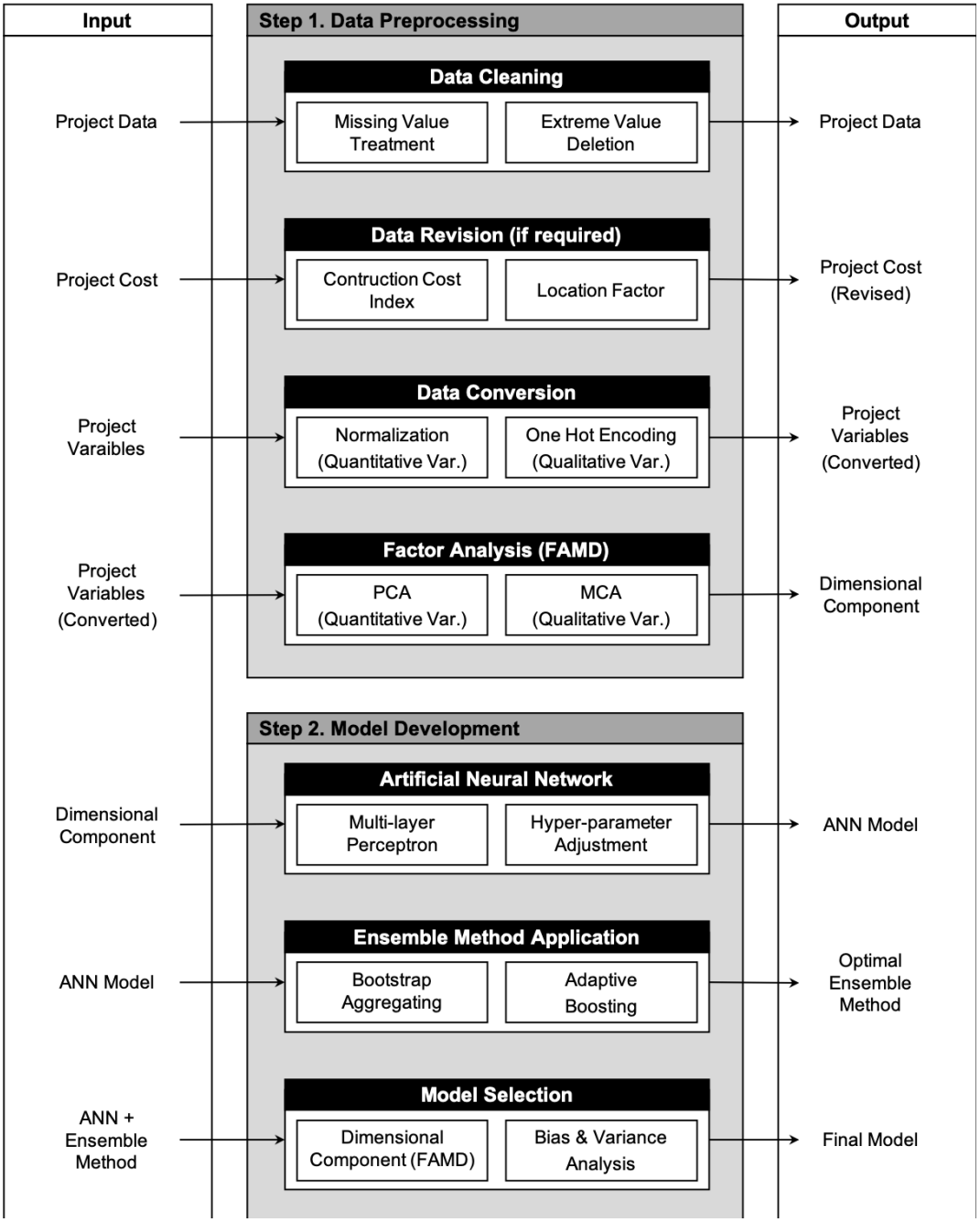

Figure 1. Proposed Methodologies for Model Development

Step 1. Data Preprocessing and Factor Analysis

As the quality of input data determines the performance of data-driven forecasting models, many researchers have highlighted the importance of data preprocessing, with Soibelman and Kim (2002), for example, noting that it accounts for approximately 60% of the total data-analysis work. It can include

data cleaning, conversion, and integration, usually performed first, involves removing missing and extreme values from the collected data (Yu, 2007). Next, all projects' costs should also be standardized using a cost-index method, to eliminate illusory analysis results by factors such as long-term changes in currency values. Various indexes can be used to adjust the cost value according to the difference of the time or region of the project information. Many public or private organizations including Association for the Advancement of Cost Engineering, Engineering News Record (ENR), or Turner Construction estimate and publish cost indexes of many types of the construction project regularly. After changes in prices have been accounted for, standardization and one-hot encoding are applied to quantitative and qualitative variables, respectively, to be processed in the ANN model. And lastly, with the aims of identifying the most descriptive features from the insufficient project data and transforming the project data into new dimensional components for the proposed forecasting model, FAMD is used to process the cost variables of early-stage cost estimation, which has a large proportion of qualitative variables. FAMD is a factorial method designed for handling data in which a group of variables is described both quantitatively and qualitatively (Saporta, 1990). FAMD covers the orthogonalization and feature extraction of both quantitative and qualitative variables, working as a form of PCA in the case of the former, and as multiple correspondence analysis (MCA) in the case of the latter. In mathematical terms, FAMD algorithm seeks to new dimensional components with input data include K quantitative variables k=1 to K, and Q qualitative variables q=1 to Q. The quantitative and qualitative variables are standardized during the analysis to balance the influence of each set of variables, and z is a quantitative variable that follows the process described in the following equation.

$$\gamma(z,k) = \textit{the correlation coefficient between var.}\, k \textit{ and } z$$

$$\eta^2(z,\, q) = \textit{the squared correlation ratio between var.}\, z \textit{ and } q$$

In the PCA of K, it looks for the function on I (a function on I assigns a value to each). The case in which initial variables and principal components are the most correlated to all K variables is expressed as:

$$\sum_k \gamma^2(z,k)$$

In MCA of Q, it looks for the function on I most related to all Q variables, using the following formula:

$$\sum_{q} \eta^2(z, q)$$

And in FAMD {K, Q}, it looks for the function on I that is most related to all K+Q variables, using:

$$\sum_{k} \gamma^2(z, k) + \sum_{q} \eta^2(z, q)$$

The original input variables contribute to the linear combinations of dimensional components with a new dimension and different weights. Thus, the dimensional components have explanatory power of original data, while avoiding a multicollinearity problem.

Step 2. Model Development: Neural Networks Ensemble Modeling

The proposed conceptual cost-forecasting model, developed through ANN modeling, includes an input layer corresponding to the project attributes, and an output layer consisting of one processing element, i.e., the project cost to be predicted. Multi-layer perceptron, a supervised-learning algorithm, was used to train the model using backpropagation with a rectifier activation function. To address the limited information availability in the early phase of a complex construction project and the difficulty of collecting a high quantity data on comparable projects, the ensemble techniques comprising bootstrap aggregation or adaptive boosting was then applied. Both these techniques start with the creation of N new training data sets via random sampling with replacement from the original dataset. Because every sample is returned to the dataset after sampling, a particular data point from the observed dataset could appear zero times, or more, in each bootstrap sample (Sonmez, 2011). In bootstrap aggregation, every element has the same probability of appearing in a new dataset due to sampling with replacement. Then the models are aggregated via voting in the case of categorical results, and by averaging the predictions in the case of numerical ones. The training stage of each model for bootstrap aggregation is built independently to combine weak learners to create a strong learner that can make accurate predictions. In boosting, on the other hand, the observations are weighted, and therefore some of them will take part

in the new sets more often than others. The boosting algorithm builds the new model sequentially, with each classifier being trained on a data subset, taking the previous classifiers' success into consideration. The weights are also redistributed after each training step; the weights of misclassified instances are increased, and the weights of correctly classified instances are decreased (Galar et al. 2011). This sequential training causes subsequent classifiers to focus more on misclassified data to improve its performance.

EXPERIMENTAL TESTS AND RESULTS

Experimental Settings and Data Collection

Two types of experiments were designed to demonstrate and validate the effectiveness and applicability of the proposed conceptual cost-forecasting model trained with limited project samples. Both tests incorporate the project data from 86 combined-cycle power-plant projects obtained from the GlobalData. The GlobalData is a London-based data analytics and consulting company covering diverse industries, and they collect and provide construction project database across 200 countries. The combined-cycle power plants generate and distribute electric power, and the project data with generating capacities of 500MW to 1000MW were collected. These 86 completed projects were in 29 countries, with 27 plants in North America, 20 in Central and South America, 16 in the Middle East and Africa, 14 in the Asia-Pacific region, and 11 in Europe. The developments of these 86 projects were started from 2005 to 2017, and the project values of these projects ranged from USD 390 million to USD 1500 million. To account for the inflation rate from one period to another, the Chemical Engineering Plant Cost Index (CEPCI) is used as a tool for adjusting plant construction costs. The CEPCI consists of a composite index assembled from a set of four sub-indexes: equipment (heat exchangers, tanks, pipes, valves, fittings, processing machinery, pumps, compressors); construction labor; buildings; and engineering/supervision. Most of these components correspond to Producer Price Indexes, updated and published monthly by the U.S. Department of Labor's Bureau of Labor Statistics. The collected information on each plant consists of three quantifiable variables – project capacity, project duration, and construction duration – and eight categorical ones: project stage, project operation

type, region, country, primary fuel, funding status, financing structure, and funding mode (Table 2). The collected variables contain all information deemed necessary to the Association for the Advancement of Cost Estimation's Class 5 estimates (with Class 4 estimates being partially covered).

Table 2. Descriptions of Project Data Attributes

Attribute	Description (Classified classes)
Project Capacity	Project capacity refers to the generating capacity of the power generation project, in units of MW (500MW to 1000MW).
Project Duration	Project duration refers to the length of the period which the project was developed, and this duration starts from the time in which the project was announced officially by the owner (2 years to 16.75 years).
Construction Duration	Construction duration refers to the length of the period which the construction was conducted, and this duration starts from the time in which the construction works commenced (0.5 years to 7 years)..
Project Stage	Project stages describe the current and exact status of the project data collected (Execution, EPC award, Tender, Pre-design, Design, and Construction complete).
Project Operation Type	Project operation type defines whether the project is developed as a single-phase/entity or as multiple-phase/entity (Parent, and Subproject).
Primary Fuel	Primary fuel is the energy source used primarily for power generation in nature and can be processed without any sort of energy conversion or transformation process (Gas, Oil).
Funding Status	Funding status is an attribute that describes how much funding the project is done(Fully funded, and Partially funded).
Funding Mode	Funding mode is an attribute that distinguishes how the project is delivered (Public, Private, Public-private joint venture).
Financing Structure	The funding structure is an attribute that describes how the funding of the project is made up of debt and equity (Dept, Equity, Dept and Equity)
Region	This attribute defines exactly where project development and construction are undertaken (5 regions).
Country	This attribute defines exactly where project development and construction are undertaken (29 countries).
Project Value	The project value is the total capital cost for the project, to be converted into US$ if sourced in other currencies (USD 390 million to USD 1500 million).

Note. MW=megawatt (one million watts); CIC= Construction Intelligence Center

The first experiment investigated the effects of applying the proposed ensemble modeling and FAMD to ANNs, by using *t*-testing to test the hypothesis that the accuracy of a cost-forecasting model

using ANNs can be improved by incorporating ensemble modeling and FAMD. The error rate for each model was calculated through a random-subsampling method, i.e., randomly dividing the entire dataset into a training set and a testing set each time. The error rate is calculated by models according to the numbers (i.e., from two to 10) of FAMD dimensional components that were input into the forecasting model. Then we compared the performance of the models in terms of accuracy and stability.

The second experiment tested the feasibility of the proposed method for conceptual cost-forecasting model training with limited project samples. In this case, three training datasets were created, one comprising 76 of the 86 projects, another comprising 38 of them, and the third, 25 of them, all selected at random, with ten randomly chosen projects' data being set aside for use as the test set. Based on its superior results during the first experiment, bootstrap aggregation was selected as the ensemble method. Because this approach's different weight initializations and random sampling of the training set can lead to different validation accuracies, five rounds of iterative analysis were conducted to ensure the validity of the results. For both experiments, the number of bootstrap aggregation samples was 500; and to estimate the discrepancy between the predicted value and the actual value for the test set, mean absolute percentage error (MAPE) was calculated, per the following formula.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_T}\right| \times 100$$

$$A_T = actual\ value \quad F_t = forecast\ value$$

Experiment 1 Results and Discussion

Table 3 shows the MAPEfor each attempt when the proposed methodologies (Bootstrap aggregation, Adaptive boosting, and FAMD) were used both individually and simultaneously. It can thus be seen that, while all three proposed techniques helped to improve prediction accuracy, the best result (10.8% MAPE) was obtained when bootstrap aggregation and FAMD were used jointly. Given the idiosyncrasies of every power-plant project and its high likelihood of cost deviation – covering between -20% and +30%, or -15% to +20%, depending on the estimation stage (Christensen and Dysert, 2005), the conceptual cost estimates results can be considered highly accurate. Especially given that the

available quantifiable variables numbered only three and only generation capacity was correlated closely with project cost, the proposed ANN incorporating bootstrap aggregation and FAMD was able to process limited quantities of project data with nonlinear relationships and apply them effectively to conceptual project-cost estimation. This also highlights that it is critical to find descriptive features in the training dataset if increasing the amount of training through ensemble modeling is to be effective.

Table 3. MAPE of Ensemble Methods and FAMD Application

Training Algorithm / Model (FAMD)	ANN only	ANN + Bootstrap Aggregation	ANN + Adaptive Boosting
Model without FAMD	16.0%	13.4%	16.3%
Model with FAMD	15.5%	10.8%	12.9%

Note. ANN=artificial neural network; FAMD=factor analysis mixed data algorithm

Table 4. Ensemble Modeling Techniques Comparison and *T*-test Results of Hypothesis Test

Training Algorithm / Category (unit)	Only ANN	ANN + Bootstrap Aggregating	ANN + Adaptive Boosting
Mean of MAPEs (%)	18.1%	14.2%	19.0%
Mean of residual (USD)	1827.7	420.1	1575.7
SD of residual (USD)	934.0	398.8	522.5
t-score		3.09	-0.69
p-value		0.0057	0.2051

Note. ANN=artificial neural network; MAPE=mean absolute percentage error, SD=standard deviation

To consider accuracy and stability when selecting a version of the developed forecasting model, it is required to analyze both residual and variance by input (i.e., the dimensional components in this

experiment). The "mean of MAPEs" in Tables 4 referred to an averaging of the MAPEMAPE by the numbers (i.e., from two to 10) of FAMD dimensional components that were input into the forecasting model. The mean of the residual is the average of the difference between the predicted value and the actual value, indicating the overall accuracy of the models, while the standard deviation (SD) of residual implying stability of models' predictive performance. In Table 4, both the mean values and standard deviation of residual were reduced when either of the ensemble methods was applied. The results further confirmed that the version of the proposed conceptual cost-forecasting model that incorporated bootstrap aggregation was able to achieve more accurate estimates than its adaptive-boosting counterpart. On the other hand, as can be seen from t-score and p-value of hypothesis testing, while bootstrap aggregation yielded a significant overall improvement in accuracy as compared to the accuracy of only ANN applied model, the positive effect of applying adaptive boosting was not statistically significant. The application of adaptive boosting exhibited lower performance of decreasing variance over the course of the random-sampling process that causes statistically not significant results from hypothesis testing. Thus, from the results of the first experiment, it can be concluded that the accuracy and stability of the forecasting model did not always improve due to the implementation of ensemble modeling, but rather, that such improvement was dependent upon on how the input values were processed during FAMD and sampled through ensemble techniques.

Experiment 2 Results and Discussion

The second experiment tested whether the proposed methodology could ensure stable performance despite a few project samples being available, and the results of the test are summarized in table 5 and figure 2. As shown in Table 5, when ANNs are used alone, their error rates are much higher with smaller datasets (i.e., 25 and 38 projects in this case) than with larger ones (i.e., 76 projects). Specifically, training using the smallest datasets (25 projects) resulted in an array of high error rates ranging from 41.80% to 67.03% (average; 53.13%). Applying bootstrap aggregation and FAMD singly both had positive effects on accuracy, though error rates were still higher when the amount of training data was smaller. For example, when only bootstrap aggregation was applied, error rates ranged up to 25.52%

with 38 projects (average; 23.76%), but as high as 41.53% with 25 projects (average; 30.05%). This suggests that ensemble modeling's positive effects are strictly limited when the quantity of training data falls below a certain threshold. Similarly, when only FAMD was applied, error rates were no more than 22.71% with 38 projects (average; 19.26%) but as high as 28.77% with 25 projects (average; 21.73%) – i.e., not as reliable as the values obtained by using all 76 projects' data, or even traditional conceptual cost-estimation practices.

Table 5. MAPE for Bootstrap Aggregation and FAMD Algorithm Application by Training Dataset

Applied Forecasting Methodologies	**Number of Projects in Training Dataset**		
	76	**38**	**25**
Default Neural Networks	30.49	55.28	53.13
ANN + Bootstrap Aggregation	21.63	23.76	30.05
ANN + FAMD	15.94	19.26	21.73
ANN + Bootstrap Aggregation + FAMD	11.55	12.90	12.06

Note 1. ANN=artificial neural network; FAMD=factor analysis mixed data

Note 2. For 38 and 25 projects dataset, the MAPE is an averaging of the MAPE from five-time test.

It is important to note, however, that when bootstrap aggregation and FAMD were applied simultaneously, the forecasting model performed similarly regardless of how much training data it was given. The results indicate that predictive performance and the effect of ensemble modeling are both likely to be better when the data input into the ANN-based model are processed using FAMD. Especially, this dual approach produced error rates ranging from 11.07% to 16.09% with 38 projects (average; 12.90%), and from 11.60% to 12.29% with 25 projects (average; 12.06%). The highest of these error rates pertained only to the fifth trial with the 38-project training dataset (16.09% of MAPE), and all four of the other tests found a similar level of performance comparable to that of the model trained on data from 76 projects (Figure 2). Another thing to note is that the models that simultaneously apply bootstrap aggregation and FAMD are also excellent in terms of prediction stability, as shown in

Figure 2. While the other models showed jagged error rate over five-time tests, the model with both bootstrap aggregation and FAMD maintained a similar level of error rate through five tests. Taken as a whole, the results of the present case study show that the proposed model incorporating an ANN, ensemble modeling, and FAMD can contribute to increasing both accuracy and reliability of conceptual cost forecasting when project samples are few.

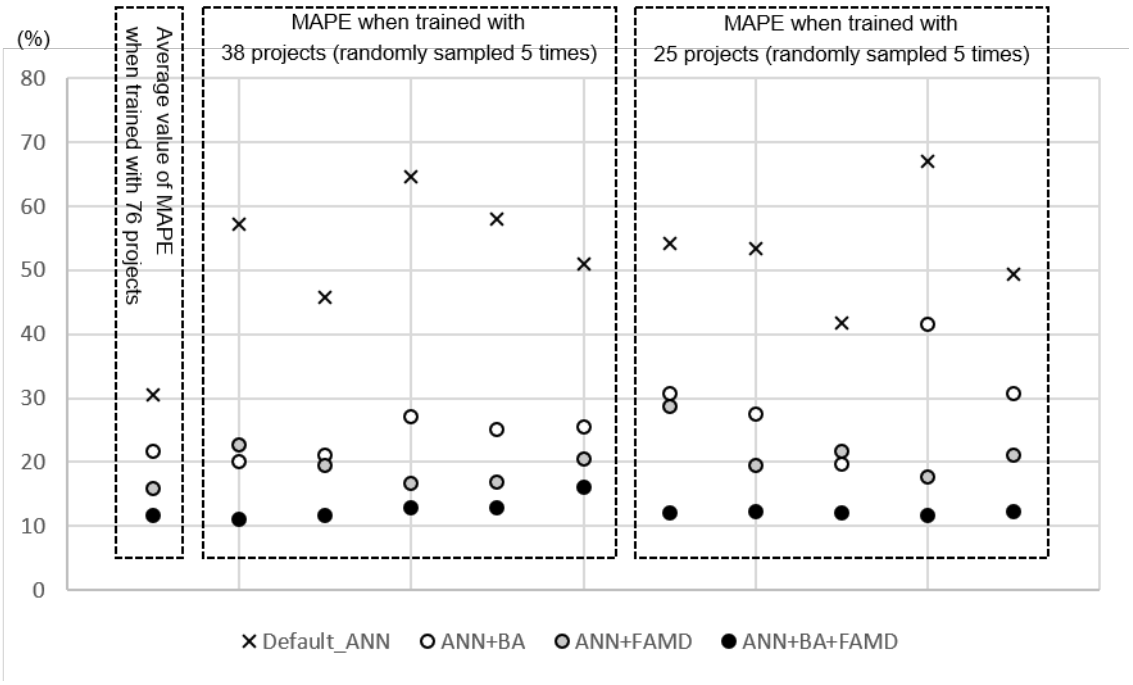

Figure 2. MAPE for Bootstrap Aggregation and FAMD Algorithm Application, by Training Dataset (76, 38, 25)

Note. ANN=artificial neural network; BA=bootstrap aggregation; FAMD= factor analysis mixed data; MAPE=mean absolute percentage error

In addition, to analyze the impact of each variable, a sensitivity analysis of input is conducted using one-factor-at-a-time (OAT) method. The OAT method investigates the changes in the output by removing one input variable while keeping others at the same values. The sensitivity analysis was performed in the same environment as in Experiment 1 (bootstrap aggregating were applied). Figure 3 shows the increased amount of MAPE when certain variables are removed for model training. The project capacity was found to be the factor that has the most significant influence on the predictive

accuracy. Even the result of the sensitivity analysis does not explain the effect of the variable on the project cost, since the neural networks are the black-box model, numerical variables had a more substantial impact on model accuracy than categorical variables.

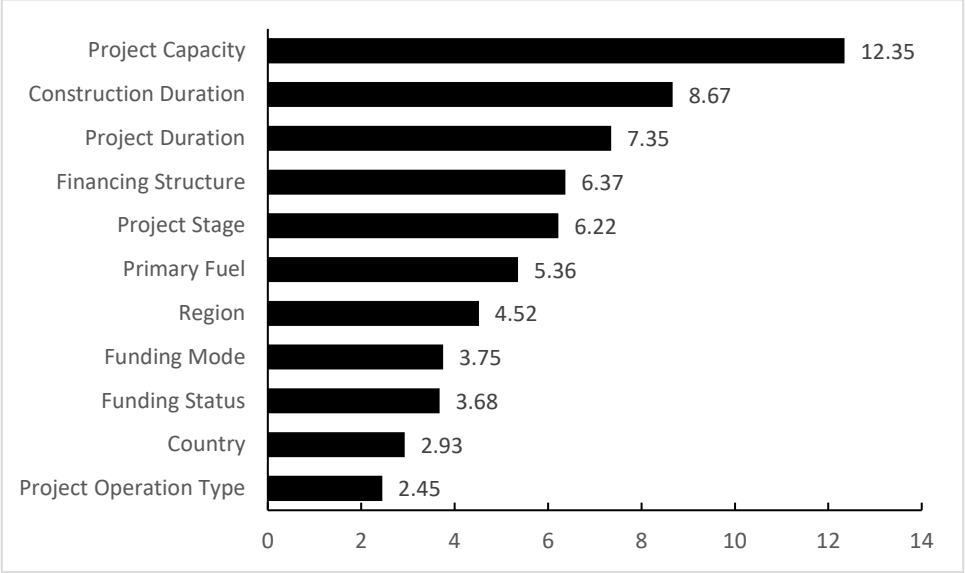

Figure 3. Result of Sensitivity Analysis

Note. Each value is an increased amount of MAPE when the model is trained excluding the specific variable (MAPE=mean absolute percentage error)

CONCLUSIONS

As forecasting the costs of large-scale construction projects with few precedents is notoriously difficult, the present study has proposed a conceptual cost-forecasting model incorporating an ANN, ensemble modeling, and a factor-analysis algorithm (FAMD). The ANN-based ensemble modeling helps to improve the model's prediction accuracy when the amount of project data is low; and FAMD finds the optimal inputs for the model. Two experiments using 86 combined-cycle power-plant projects data concluded that ensemble modeling and FAMD are both conjointly capable of improving the accuracy and stability of conceptual cost estimates. The performance of the proposed model was also tested on two randomly selected small training-data subsets, and the model version combining bootstrap aggregation and FAMD improved estimation accuracy and reliability despite these very low project sample sizes. It was also noteworthy that the effects of both bootstrap aggregation and adaptive boosting varied to increase accuracy and stability, depending on the number of input dimensional components a

forecasting-model version was given. These results indicate that the model-selection process should consider how each version of a forecasting model reflects the key features of the project data.

Since the data in the case study consisted only certain type of project, the generalizability of our findings is hard to gauge. More concrete conclusions, as well as better model performance, therefore await the acquisition of a more extensive dataset. More experiments involving a wider range of project types also need to be conducted if we are to generalize our findings across the construction industry. Additionally, it should be acknowledged that using orthogonalized dimensional components derived from FAMD as input parameters for an ANN-based model, as we did, has some disadvantages – because the ANN model is a 'black box' – our model does not facilitate understanding of the effects of individual variables on project costs. Therefore, the results of this research cannot answer questions about which or how many variables should be used to improve prediction performance, and further experiments should seek to establish the appropriate numbers and characteristics of input variables. Despite these limitations, the difficulty of applying prior data-driven conceptual cost-forecasting approaches to ever bigger and more complex construction projects means that our model-development approaches and findings can contribute to more reliable conceptual cost estimation in the early stages of project development.

DATA AVAILABILITY STATEMENT

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

ACKNOWLEDGMENT

This research study was supported by the Start-up Fund project (No. 1-ZE6Y) from the Hong Kong Polytechnic University, Hong Kong.

REFERENCES

1. Akintoye, A., 2000. Analysis of factors influencing project cost estimating practice. *Construction Management & Economics*, 18(1), pp.77-89.

2. Alex, D.P., Al Hussein, M., Bouferguene, A. and Fernando, S., 2010. Artificial neural network model for cost estimation: City of Edmonton's water and sewer installation services. *Journal of Construction Engineering and Management*, 136(7), pp.745-756.

3. An, S.H., Park, U.Y., Kang, K.I., Cho, M.Y. and Cho, H.H., 2007. Application of support vector machines in assessing conceptual cost estimates. *Journal of Computing in Civil Engineering*, 21(4), pp.259-264.

4. Arafa, M. and Alqedra, M., 2011. Early stage cost estimation of buildings construction projects using artificial neural networks. *Journal of Artificial Intelligence*, 4(1), pp.63-75.

5. Attalla, M. and Hegazy, T., 2003. Predicting cost deviation in reconstruction projects: Artificial neural networks versus regression. *Journal of construction engineering and management*, 129(4), pp.405-411.

6. Bishop, C.M., 2006. *Pattern recognition and machine learning*. springer.

7. Breiman, L., 1996. *Bagging predictors. Machine learning*, 24(2), pp.123-140.

8. Cao, Y., Ashuri, B. and Baek, M., 2018. Prediction of unit price bids of resurfacing highway projects through ensemble machine learning. *Journal of Computing in Civil Engineering*, 32(5), p.04018043.

9. Choi, S., Kim, D.Y., Han, S.H. and Kwak, Y.H., 2014. Conceptual cost-prediction model for public road planning via rough set theory and case-based reasoning. *Journal of Construction Engineering and Management*, 140(1), p.04013026.

10. Christensen, P. and Dysert, L.R., 2005. AACE International Recommended Practice No. 18R-97 Cost Estimate Classification System–As Applied in Engineering, Procurement, and Construction for the Process Industries (TCM Framework: 7.3–Cost Estimating and Budgeting). AACE.

485 11. Dominic, A.D.D. and Smith, S.D., 2014. Rethinking construction cost overruns: cognition,
486 learning and estimation. *Journal of financial management of property and construction*,
487 19(1), pp.38-54.

488 12. Dursun, O. and Stoy, C., 2016. Conceptual estimation of construction costs using the
489 multistep ahead approach. *Journal of Construction Engineering and Management*, 142(9),
490 p.04016038.

491 13. Efron, B. and Tibshirani, R.J., 1994. *An introduction to the bootstrap*. CRC press.

492 14. Elfaki, A.O., Alatawi, S. and Abushandi, E., 2014. Using intelligent techniques in
493 construction project cost estimation: 10-year survey. *Advances in Civil Engineering*, 2014.

494 15. Elkassas, E.M., Mohamed, H.H. and Massoud, H.H., 2009. The neural network model for
495 predicting the financing cost for construction projects. *International Journal of Project*
496 *Organisation and Management*, 1(3), pp.321-334.

497 16. El-Sawalhi, N.I. and Shehatto, O., 2014. A neural network model for building construction
498 projects cost estimating. *KICEM Journal of Construction Engineering and Project*
499 *Management*, 4(4).

500 17. ElMousalami, H.H., Elyamany, A.H. and Ibrahim, A.H., 2018. Predicting conceptual cost
501 for field canal improvement projects. *Journal of Construction Engineering and*
502 *Management*, 144(11), p.04018102.

503 18. Elmousalami, H.H., 2020. Artificial Intelligence and Parametric Construction Cost
504 Estimate Modeling: State-of-the-Art Review. *Journal of Construction Engineering and*
505 *Management*, 146(1), p.03119008.

506 19. Emsley, M.W., Lowe, D.J., Duff, A.R., Harding, A. and Hickson, A., 2002. Data
507 modelling and the application of a neural network approach to the prediction of total
508 construction costs. *Construction Management & Economics*, 20(6), pp.465-472.

20. Fellows, R.F. and Liu, A.M., 2015. *Research methods for construction*. John Wiley & Sons.

21. Firoozabadi, K.J., Rouhani, S. and Bagheri, N., 2013. Review of EPC projects cost estimation and minimum error technique introduction. *International Journal of Science and Engineering Investigations*, 2.

22. Flood, I. and Kartam, N., 1994. Neural networks in civil engineering. I: Principles and understanding. *Journal of computing in civil engineering*, 8(2), pp.131-148.

23. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. and Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), pp.463-484.

24. Gardner, B.J., Gransberg, D.D. and Jeong, H.D., 2016. Reducing data-collection efforts for conceptual cost estimating at a highway agency. *Journal of Construction Engineering and Management*, 142(11), p.04016057.

25. Green, S.B., 1991. How many subjects does it take to do a regression analysis. *Multivariate behavioral research*, 26(3), pp.499-510.

26. Günaydın, H.M. and Doğan, S.Z., 2004. A neural network approach for early cost estimation of structural systems of buildings. *International journal of project management*, 22(7), pp.595-602.

27. Gunduz, M. and Sahin, H.B., 2015. An early cost estimation model for hydroelectric power plant projects using neural networks and multiple regression analysis. *Journal of Civil Engineering and Management*, 21(4), pp.470-477.

28. Hagan, M., Demuth, H., Beale, M. and De Jesus, O., 2016. Neural Network Design 2nd. *Ed., Lexington, KY*.

29. Hashemi, S.T. and Kaur, H., 2017. A hybrid conceptual cost estimating model using ANN and GA for power plant projects. *Neural Computing and Applications*, pp.1-12.

30. Huang, J., Li, Y.F. and Xie, M., 2015. An empirical analysis of data preprocessing for machine learning-based software cost estimation. *Information and software Technology*, 67, pp.108-127.

31. Hyari, K.H., Al-Daraiseh, A. and El-Mashaleh, M., 2016. Conceptual cost estimation model for engineering services in public construction projects. *Journal of Management in Engineering*, 32(1), p.04015021.

32. Ji, S.H., Park, M. and Lee, H.S., 2012. Case adaptation method of case-based reasoning for construction cost estimation in Korea. *Journal of Construction Engineering and Management*, 138(1), pp.43-52.

33. Jin, R., Cho, K., Hyun, C. and Son, M., 2012. MRA-based revised CBR model for cost prediction in the early stage of construction projects. *Expert Systems with Applications*, 39(5), pp.5214-5222.

34. Kim, G.H., An, S.H. and Kang, K.I., 2004. Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and environment*, 39(10), pp.1235-1242.

35. Kim, G.H., Yoon, J.E., An, S.H., Cho, H.H. and Kang, K.I., 2004. Neural network model incorporating a genetic algorithm in estimating construction costs. *Building and Environment*, 39(11), pp.1333-1340.

36. Kirkham, R., 2014. *Ferry and brandon's cost planning of buildings*. John Wiley & Sons.

37. Koo, C., Hong, T., Hyun, C. and Koo, K., 2010. A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects. *Canadian Journal of Civil Engineering*, 37(5), pp.739-752.

38. Liu, M. and Ling, Y.Y., 2005. Modeling a contractor's markup estimation. *Journal of construction engineering and management*, 131(4), pp.391-399.

39. Lowe, D.J., Emsley, M.W. and Harding, A., 2006. Predicting construction cost using multiple regression techniques. *Journal of construction engineering and management*, 132(7), pp.750-758.

40. Mahamid, I., 2011. Early cost estimating for road construction projects using multiple regression techniques. *Construction Economics and Building*, 11(4), pp.87-101.

41. Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T. and Voordijk, H., 2019. An artificial neural network approach for cost estimation of engineering services. *International journal of construction management*, pp.1-14.

42. Petroutsatou, C., Lambropoulos, S. and Pantouvakis, J.P., 2006. Road tunnel early cost estimates using multiple regression analysis. Operational Research, 6(3), pp.311-322.

43. Pewdum, W., Rujirayanyong, T. and Sooksatra, V., 2009. Forecasting final budget and duration of highway construction projects. *Engineering, Construction and Architectural Management*.

44. Ranasinghe, M., 2000. Impact of correlation and induced correlation on the estimation of project cost of buildings. *Construction Management & Economics*, 18(4), pp.395-406.

45. Saporta, G., 1990. Simultaneous analysis of qualitative and quantitative data. In *Proceedings of the 35th Scientific Meeting of the Italian Statistical Society* (pp. 63-72).

46. Setyawati, B.R., Creese, R.C. and Sahirman, S., 2003. Neural networks for cost estimation (Part 2). *AACE International Transactions*, p.ES141.

47. Soibelman, L. and Kim, H., 2002. Data preparation process for construction knowledge generation through knowledge discovery in databases. *Journal of Computing in Civil Engineering*, 16(1), pp.39-48.

581 48. Sonmez, R., 2004. Conceptual cost estimation of building projects with regression analysis
582 and neural networks. *Canadian Journal of Civil Engineering*, 31(4), pp.677-683.

583 49. Sonmez, R., 2011. Range estimation of construction costs using neural networks with
584 bootstrap prediction intervals. *Expert systems with applications*, 38(8), pp.9913-9917.

585 50. Stoy, C., Pollalis, S. and Schalcher, H.R., 2008. Drivers for cost estimating in early design:
586 Case study of residential construction. *Journal of construction engineering and*
587 *management*, 134(1), pp.32-39.

588 51. Tsai, T.I. and Li, D.C., 2008. Utilize bootstrap in small data set learning for pilot run
589 modeling of manufacturing systems. *Expert Systems with Applications*, 35(3), pp.1293-
590 1300.

591 52. Wang, Y.R. and Gibson Jr, G.E., 2010. A study of preproject planning and project success
592 using ANNs and regression models. *Automation in Construction*, 19(3), pp.341-346.

593 53. Wilmot, C.G. and Mei, B., 2005. Neural network modeling of highway construction costs.
594 *Journal of construction engineering and management*, 131(7), pp.765-771.

595 54. Williams, T.P. and Gong, J., 2014. Predicting construction cost overruns using text mining,
596 numerical data and ensemble classifiers. *Automation in Construction*, 43, pp.23-29.

597 55. Yu, W.D., 2007. Hybrid soft computing approach for mining of complex construction
598 databases. *Journal of computing in civil engineering*, 21(5), pp.343-352.