Journal of Cost Analysis and Parametrics, 7:180–218, 2014
Copyright © ICEAA
ISSN: 1941-658X print / 2160-4746 online
DOI: 10.1080/1941658X.2014.984880

www.iceaaonline.com

A Process for the Development and Evaluation of Preliminary Construction Material Quantity Estimation Models Using Backward Elimination Regression and Neural Networks

BORJA GARCÍA DE SOTO[1], BRYAN T. ADEY[1], and DILUM FERNANDO[2]

[1]Institute of Construction and Infrastructure Management (IBI), Swiss Federal Institute of Technology (ETHZ), Zurich, Switzerland
[2]School of Civil Engineering, University of Queensland, Australia

During the early stages of a project, it is beneficial to have an accurate preliminary estimate of its cost. One way to make those estimates is by determining the amount of construction material quantities that are required and then multiplying the estimated construction material quantities by the corresponding unit cost. One advantage of making estimates in this way is that it allows for the segregation of quantities and costs. This way they can be updated separately as new information becomes available. They can also be tracked separately allowing decision makers to make better decisions about the project during its conceptual phase. There are several techniques that can be used to develop estimation models. The most common include regression analysis and artificial intelligence, such as neural networks. Work has been done, however, in a non-standardized way, leaving practitioners without guidance as to how to develop and evaluate models for their specific purposes. This can be seen in particular in the many different types of metrics used for the evaluation of models. The goal of the work presented in this article was to create a process to (1) develop models to be used to prepare preliminary estimates of construction material quantities taking into consideration the available data during the early stages of a project, and (2) evaluate the developed models using the Akaike information criterion. The proposed process is illustrated with an example in which data from 58 storage buildings was used to develop models to estimate the amount of concrete and reinforcement required using backward elimination regression analysis and neural network techniques. The developed models were then evaluated using a second-order correction Akaike information criterion (AIC$_c$) to select the most accurate model for each construction material quantity. The proposed process is useful for practitioners in need of developing robust estimation models in a consistent and systematic way, and the AIC$_c$ metric proved to be effective for selecting the most accurate models from a set.

Introduction

In the conceptual phase of a project, it is beneficial to have an accurate estimate of its cost. One way to make preliminary estimates is to determine the amount of construction material quantities (CMQs) to be used in the project (Bakhoum, Morcous, Taha, & El-Said, 1998; Chou, Peng, Persad, & O'Connor, 2006; Du & Bormann, 2014; Fragkakis, Lambropoulos, & Tsiambaos, 2011; Kim, Kim, & Kang, 2009; Oh, Park, & Kim, 2013; Singh, 1990, 1991;

Address correspondence to Borja García de Soto, Institute of Construction and Infrastructure Management (IBI), Swiss Federal Institute of Technology (ETHZ), Stefano-Franscini-Platz 5, Zurich 8093, Switzerland. E-mail: garcia.de.soto@ethz.ch

Son, Lee, Park, Han, & Ahn, 2013; Yeh, 1998) and then to multiply these estimates by the unit costs, which typically include fabrication and delivery, erection, installation, insurance, site indirect costs, supervision, and overhead and profit. One advantage of making estimates in this way, as compared to estimating cost directly, is that either the CMQs or their unit costs can be updated separately as new information becomes available. It also allows managers to make better decisions and keep a better track of the project (Chou et al., 2006; Yu, 2006) by controlling the changes in quantities and costs separately.

During the early stages of a project, CMQs are generally estimated by finding existing structures that are similar to the structures to be constructed and by adjusting the CMQs used in the former to estimate the CMQs to be used in the latter. The adjustments are done based on the differences in the values of key parameters (e.g., height), taking into consideration the experience of the estimator. Similar structures are, however, often difficult to find, or the required adjustments might become tedious due to the many possible values of key parameters, all which can have a significant effect on the design of a structure, hence the amount of CMQs required. Without the use of specific models to deal with the situations when no similar structures are available, the estimate of the CMQs can vary widely from estimator to estimator, and as a consequence, many of the estimates are not very accurate.

One improvement in the estimation of CMQs is the use of models that can recreate real cases. The main advantage of the use of models is the improvement in accuracy of CMQ estimates for projects. This leads, in turn, to better decisions as whether or not to proceed with construction, and if the decision is to proceed, to better decisions as to the type of structure that should be constructed (e.g., a storage building out of reinforced concrete or one out of a combination of steel and reinforced concrete), depending on specific project parameters and other conditions, such as economic ones. The main challenge when developing models is the availability of sufficient quantities of accurate data. As the number of independent variables (IVs) that have an effect on the quantity being modeled increases, so does the amount of data required to develop accurate models.

A sufficient amount of accurate data is difficult to obtain, either because it does not exist, or because those that have the data would not make it accessible. The former can happen, for example, when not enough structures of one type have been constructed or when enough structures of one type have been constructed but not enough data has been collected. The latter can happen in highly competitive industries.

Several studies have been conducted where the accuracy of models developed using different techniques using regression analysis (RA), or artificial intelligence, such as neural network (NN) and case-based reasoning (CBR), have been compared (Cho, Kim, Kim, & Kim, 2013; Kim, An, & Kang, 2004; Kim, Shin, Kim, & Shin, 2013; Lowe, Emsley, & Harding, 2006; Smith & Mason, 1997; Sonmez, 2004; Yeh, 1998). None of these, however, have included a systematic way to develop models to estimate CMQs using different techniques and to evaluate the developed models to determine the most accurate one from amongst them.

The objective of the work presented in this article was to fill this gap for models developed using RA and NN techniques. This was done by developing a process to both develop such models, taking into consideration the availability of data, and evaluating their performance or accuracy using an information criterion (commonly referred to Akaike information criterion, [AIC]; Akaike, 1974). The process is presented in this article and is illustrated through its use in developing models to estimate, in the conceptual phase of a project, the amount of concrete and reinforcement to be used to construct storage buildings, and evaluating the developed models to determine the most accurate one.

Background

Overview of Literature

Over the last few decades, there have been an increasing number of researchers who have focused on the development of models to estimate construction costs by first estimating CMQs and multiplying those estimates by unit costs (Table 1) and the evaluation of the most suitable models (Table 2) to be used.

A quick look at these tables shows that models have been developed to make construction cost estimates in a large number of areas, from school buildings to pressure vessels. This continued push by researchers to make ever better estimates can also be seen as an indication of the coming use of these models in practice, especially when it is taken into consideration that this work in numerous cases was funded by infrastructure management organizations.

It can, however, also be seen that the work has been done in a non-standardized way, something which is okay for research but leaves practitioners without guidance as to how to develop and evaluate models for their specific purposes. This can be seen in particular in the many different types of metrics used to evaluate models.

Model Types Included in Proposed Process

As RA and NN models are the most common types of models (Table 1 and Table 2), the proposed process has been developed for these types.

RA Models. RA is one of the most commonly used techniques in statistical modeling and has been used to make both direct (Bowen & Edwards, 1985; Khosrowshahi & Kaka, 1996; Kim et al., 2004; Lowe et al., 2006) and indirect construction cost estimates based on CMQ estimates (Bakhoum et al., 1998; Chou et al., 2006; Du & Bormann, 2014; Fragkakis et al., 2011; Kim et al., 2009; Oh et al., 2013; Singh, 1990, 1991; Son et al., 2013; Yeh, 1998). Equation 1 shows a generic linear regression model.

$$Y = \beta_o + \sum_{i=1}^{n} \beta_i X_i + \sum_{j=1}^{m} \beta_{n+j} X_{n+j} + \varepsilon \tag{1}$$

Where,

Y: Output from the regression equation
β_o: Constant term (y-intercept)
$\beta_1 \rightarrow \beta_n$: Regression coefficients for continuous variables
$\beta_{n+1} \rightarrow \beta_{n+m}$: Regression coefficients for categorical variables
$X_1 \rightarrow X_n$: Continuous IVs
$X_{n+1} \rightarrow X_{n+m}$: Categorical IVs
ε: Error term.

When linear relationships exist between the dependent variable (DV) and IVs, then linear RA should be used. Otherwise, attempts should be made to use nonlinear equations (Gerrard, Brass, & Peel, 1994; Yeh, 1998) or transform the data to improve its behavior and taking advantage of nonlinear forms (Chou et al., 2006). The general function used to generate the RA models for the transformed dataset (taking the natural log of the dataset) has the following form (Equation 2):

TABLE 1 Examples of research in which models were developed to make cost estimates based on CMQs

Source	Model	Estimate for	Output	No. of IVs
Singh (1990, 1991)	RA	High-rise commercial buildings	Quantities and cost of reinforced concrete beam and slab construction	6[1]
Yeh (1998)	Regression and NN	Steel and RC buildings	Total weight of steel in steel buildings and reinforcing in RC buildings	8[2] for steel and 10[3] for RC buildings
Bakhoum et al. (1998)	NN and RA	Pre-stressed concrete bridges	Quantity of concrete, reinforcement and pre-stressing in concrete approaches and the navigable span of bridges	6[4] for regression models and 5[5] for NN models
Chou et al. (2006)	RA	Highway repair projects	Quantities of different work categories from a typical work item breakdown structures (WBS; e.g., Earthwork and Landscape)	Depend on work category[6]
Kim et al. (2009)	RA	Pre-stressed concrete beam bridges	Standard work quantities (e.g., manufacturing PSC beam, rebar fabrication/placing)	3[7]
Fragkakis et al. (2011)	RA	Foundations of concrete bridges	Volume of concrete and weight of reinforcing steel	4[8]
Du & Bormann (2014)	CBR	Power plants	Quantities of 20 work items (e.g., volume of concrete pouring, finishing and curing)	12[9]
Oh et al. (2013)	RA	Substructure of steel box girder bridges	Quantities of materials for substructure components of steel box girder bridges to estimate cost.	7[10]

(*Continued*)

TABLE 1 (*Continued*)

Source	Model	Estimate for	Output	No. of IVs
Son et al. (2013)	RA	Mixed-use residential buildings	Quantity of ready-mixed-concrete, reinforcement and formwork of 4 components (i.e., foundation, basement, ground floor, and upper floor)	4 for foundation and 2 for basement, ground/upper floor[11]

[1] Grades of concrete, grid locations, number of stories, different structural schemes, grid sizes, section of beams;

[2] Number of stories, number of bays along and across frame, typical bay length along and across frame plane, seismic zone factor, live and dead load;

[3] Number of stories, total height of the building, number of bays along and across frame, typical bay length along and across frame plane, seismic zone factor, live and dead load, compressive strength of concrete;

[4] Maximum span length, superstructure type, structure system, superstructure construction method, contract type, and design type;

[5] Main navigable span length, superstructure type, structure system, construction method of superstructure, and contract type;

[6] E.g., for quantity of excavation: project length, project width, rehabilitation of existing road, percentage of trucks, vehicles per day, bridge widening or rehabilitation, interchange, new location non-freeway;

[7] Length of bridge, width of bridge, and the length of span;

[8] Height of pier, the width and length of the supported deck, and the type of the pier connection with the deck;

[9] Including megawatt reading, nature of the job, Engineering company, mechanical equipment configuration, total number of mechanical equipment, job site layout classification, measurement of center line (feet) of outside stacks in a multiple unit configuration, the vendor of steam turbine, installation of a project is at an existing facility, among others;

[10] Overall height of the bridge abutment, the number of piers, the overall height of the piers, the types and numbers of bridge bearings, the number of locations where steel pipe piles were used, and the length of the steel pipe piles;

[11] Area, perimeter and thickness of foundation; wall-to-floor ratio and floor area of basement; wall-to-floor ratio and floor area of ground floor; wall-to-floor ratio and floor area of typical upper floor.

$$\ln(Y) = \beta_o + \sum_{i=1}^{n} \beta_i \ln(X_i) + \sum_{j=1}^{m} \beta_{n+j} X_{n+j} + \varepsilon \tag{2}$$

However, interpretation of the transformed models can be tricky. Since these models are for natural logarithmic data, and the desired DV is not represented by a natural logarithm, the developed models need to be transformed back. This back transformation requires taking in account the nonlinear relationships between the DV and IVs by using linear regression. The back-transformed equation (Equation 3) is a particular type of nonlinear relationship, also known as a constant elasticity of multiplicative relationship (Albright, Winston, & Zappe, 2003).

$$Y = \left(X_1^{\beta_1} X_2^{\beta_2} \cdots X_n^{\beta_n}\right) e^{\left[\beta_o + \sum_{j=1}^{m} \beta_{n+j} X_{n+j} + \frac{SEE^2}{2}\right]} \tag{3}$$

TABLE 2 Examples of work used to evaluate models used to estimate construction costs

		Model type			Metrics used for model selection/accuracy					
Source	Estimate for	RA	NN	CBR	SVM*	MAPE	RMSE	MSE	R^2	ANOVA
Smith & Mason (1997)	Pressure vessels	1	1	–	–	1	1	–	–	–
Yeh (1998)	Steel and RC buildings	1	1	–	–	1	1	–	–	–
Sonmez (2004)	Continuing care community facilities	1	1	–	–	1	–	1	–	–
Kim et al. (2004)	Residential buildings	1	1	1	–	1	–	–	–	1
Lowe et al. (2006)	Buildings (unspecified types)	1	1	–	–	1	–	–	1	–
Kim et al. (2013)	School building	1	1	–	1	1	–	–	–	1
Cho et al. (2013)	Elementary schools	1	1	–	–	1	–	1	–	–

*Support Vector Machine (i.e., a robust classification and regression technique that maximizes the predictive accuracy of a model without over-fitting the training data. It is particularly suited to analyze data with a large number of independent variables; SPSS, 2010).

Models using this theory have already been applied for quantity-based preliminary cost estimates (Chou et al., 2006).

The linear models using the transformed data are unbiased (i.e., mean of the residuals is zero); however, once back-transformed, the mean of the residuals is no longer equal to zero. Therefore, a bias is introduced to the back-transformed model to bring the mean of the residuals to zero. The nature of this bias has been explored by several authors (e.g., Baskerville, 1972; Wood, 1986), and it is beyond the scope of this study. A correction factor (CF) addressing the transformation to address this bias must be added to the back-transformed model. When the back transformation is based on natural logarithms and the errors are normally distributed, the CF is $e^{SEE^2/2}$, where e is the exponential constant and SEE is the standard error of the estimate (Smith, 1993; Sprugel, 1983).

The RA models included in the proposed process are developed using the backward elimination technique (BET), a stepwise regression technique that uses statistical constraints to determine if a variable is kept or removed from the regression equation. Stepwise techniques have been criticized (Flom & Cassell, 2007) mostly because they are seen as an automated process in which analysts do not think the problem through thoroughly; however, they are justified as a tool to help in the identification of representative parameters (Abdul-Wahab, Bakheit, & Al-Alawi, 2005, Al-Alawi, Abdul-Wahab, & Bakheit, 2008; Chan & Park, 2005; Gray & Kinnear, 2012; Kim et al., 2004; Lowe et al., 2006).

The reason for choosing the BET over the forward selection one was based on preliminary findings from our work which showed similar findings from other researchers. For example, the study by Lowe et al. (2006) showed that the RA models developed using the BET performed better and used more significant variables than the ones using the forward selection technique. Koo, Hong, Hyun, and Koo (2010) also found that the best RA models were developed using the BET Field (2009) also indicated that in general, the BET is more preferred than the forward selection one. The main reasons for that are that in the forward elimination technique it is more likely that an IV that has a significant effect on the value of the DV is excluded from the RA model. This can happen, for example, if there are two correlated IVs that have a significant effect on the DV. In such a situation, it can happen that once the first IV is integrated into the regression equation, the integration of the second does not lead to a substantial improvement to the model. This does not happen when the BET is used because the process is reversed (i.e., all IVs are originally included and then sequentially removed; Field, 2009; Lowe et al., 2006).

There are software packages (e.g., MATLAB, 2012; SPSS, 2010) that can be used for the development of models using RA. The RA models for this study have been developed using the backward elimination function in SPSS (SPSS, 2010) with a default probability of F (p-value) for removal set at 0.1.

NN Models. The use of NN models in cost estimation was driven by the difficulty of determining the shape of the function in any parametric estimation using the RA technique, in particular, complex non-linear ones. NN models eliminate the need to define that function (Kim et al., 2004), and they are a good tool for the development of complex non-linear systems (Yeh, 1998). A NN model is constructed by creating connections between processing elements. The organization of the network, its activation function, and the weights of the connections in the network determine the output of the NN model, or in this case the estimated CMQs.

The most common NN models are fully connected feed-forward network with three layers (one hidden layer) and back-propagation supervised learning algorithm (Bishop, 1994, 1995; Hegazy & Ayed, 1998; Hegazy, Fazio, & Moselhi, 1994). A typical single-layer NN model contains an input layer consisting of *p* number of input variables plus a bias, a hidden layer consisting of *m* number of neurons plus a bias, and an output layer with the desired output (Figure 1).

The bias neurons are added to facilitate modeling, and they always have an input value of 1. While changes in the weights of non-biased neurons cause changes in the steepness of the activation function (Figure 2 *a*), the changes in the weight of biased-neurons allow the horizontal shiftiness of the activation function (Figure 2 *b*). In Figure 2, *w1, w2*, and *w3* are the weights between the input and the processing neurons, and *B1, B2*, and *B3* are the weights between the bias and the processing neurons. This horizontal shift from the origin to match the connections to the neurons is something critical for successful learning (Hegazy et al., 1994).

The NN learned relationship among sets of input–output data (training data) by adjusting network parameters (connection weights and activation functions). The number of

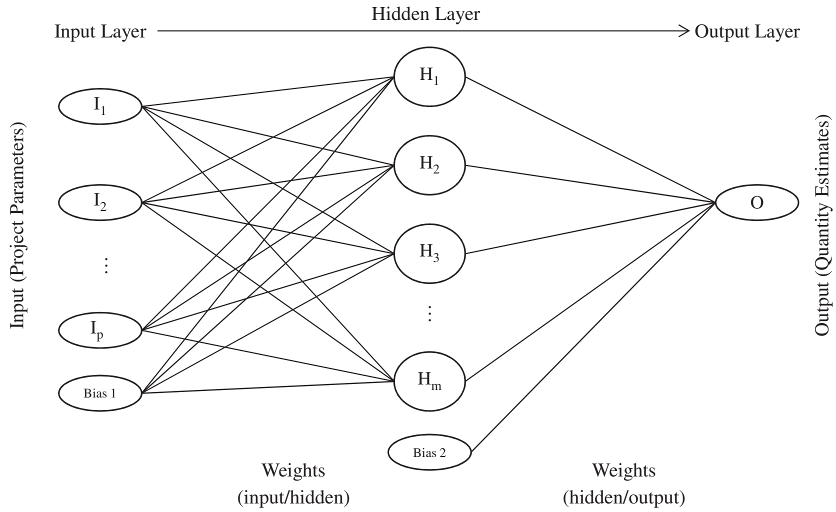
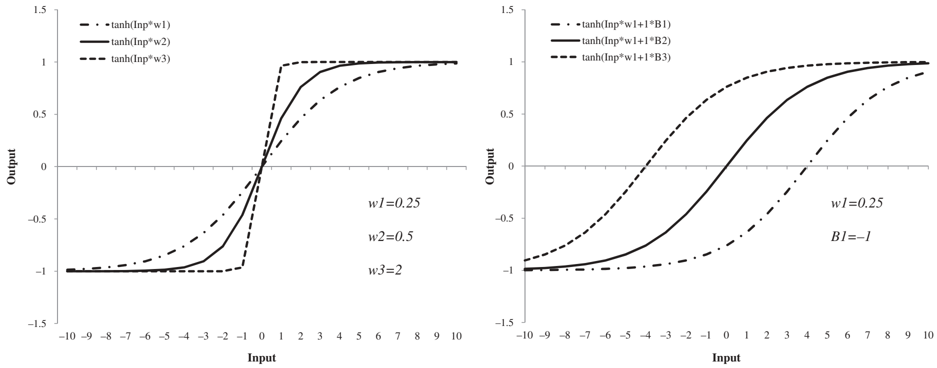


FIGURE 1 Typical NN configuration.



(a) without bias *(b) with bias*

FIGURE 2 Graphical representation of function behavior (Inp: Input).

hidden layers and number of neurons in each hidden layer can also affect the performance of the NN model. The basic algorithm used for NN training is the back propagation algorithm, through which the NN model with the lowest error is identified. This process can be done manually or automatically. The NN models used in the proposed process are developed as described below.

Input and output layers. The size of the input layer is determined from the IVs from the selected RA model. The output layer contains one neuron, that of the CMQ being estimated. The input layer is normalized by scaling it to match the range of the activation function (see section "Activation function"); hence the result of the output layer is scaled back. Equation 4 shows the general equation to normalize the data by scaling it to a given range.

$$\begin{aligned} &Norm(x) = Scale_{LB} + \frac{(X - X_{\min}) \times (Scale_{HB} - Scale_{LB})}{X_{\max} - X_{\min}} \\ &\{for\,[-1:1],\ if\,X_{\max} = X_{\min} \rightarrow Norm(x) = 0 \end{aligned} \tag{4}$$

Where,

Norm(x): normalized value of x (scaled between low and high bounds)
$Scale_{LB}$*:* low bound of scaling range (e.g., –1)
$Scale_{HB}$*:* high bound of scaling range (e.g., 1)
x: value to be normalized by scaling it to a selected range
X_{min}: minimum value for variable X
X_{max}: maximum value for variable X.

Number of hidden layers and processing neurons. The hidden layer consists of neurons, also referred to as processing neurons, which perform summation and function. One hidden layer is typically sufficient. NN models with one hidden layer provide reliable mapping between the input and the output, provided that sufficient connections are available (Hegazy & Ayed, 1998; Hegazy et al., 1994; Owusu-Ababio, 1998). NN models with two hidden layers or more are more prone to fall into a local minimum, and their estimating capabilities have been shown to be similar to NN models with one hidden layer (De Villiers & Barnard, 1993). As suggested by Hegazy et al. (1994), the number of neurons in the hidden layer is initially set as half of the number of input variables plus the output. The number of neurons is then determined by increasing and decreasing it until no improvements in the CMQs estimates are achieved during the training process (see section on Training of the NN model and determination of weights).

Activation function. The behavior of the neurons (i.e., the weights of the NN model) is determined by the activation function. Sigmoid-curve type functions are generally preferred for the development of NN models. Karlik and Olgac (2011) carried a comparative study on five conventional and monotonic activation functions (including the hyperbolic tangent function [*tanh*] as shown in Equation 5) and found that the hyperbolic tangent function performs better than other conventional activation functions used in their study. It has also been proved beneficial for the learning of the NN model to normalize the input by scaling it to match the range of the activation function (Hegazy et al., 1994). The output of the hyperbolic tangent function lies in the range –1 to +1 (Figure 3); hence, the input of the model needs to be normalized accordingly (e.g., using Equation 4).

$$\tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{5}$$

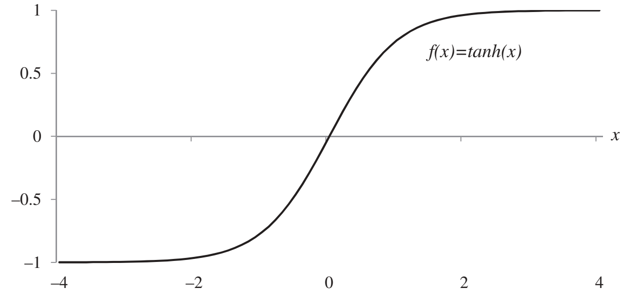


FIGURE 3 Graphical representation of the hyperbolic tangent (*tanh*) function.

Training of the NN model and determination of weights. To complete the construction of the NN model, the neurons among the different layers are connected, and those connections are weighted. The value of those weights is calculated during the training of the network by using a learning algorithm (e.g., back-propagation) that minimizes the objective function (e.g., sum-of-squares error) used by comparing the output of the network with the true values of the training set (supervised learning) using the information sent through the network (feed-forward). For the training of the NN model, the data are divided into a training set and a testing set. This helps to avoid over-fitting and to make generalization possible (Bishop, 1994, 1995). A random data split of 80%:20% was used, as recommended by Hegazy and Ayed (1998). The output of the neurons is forwarded to the next layer until the final output is calculated. The weights obtaining after training the network are recorded in two weight matrices, one between the input and the hidden layer ($\mathrm{W_A}$) and one between the hidden layer and the output layer ($\mathrm{W_B}$).

Following that process, the value of each processing neuron is calculated from the combination of the input and the connection weights (Equation 6).

$$H_j = \tanh\left(\sum_{i=1}^{p}(I_i \times W_{ij}) + 1.0 \times B_{1j}\right) \tag{6}$$

Where,

p: number of input parameters

H_j: output of neuron j from input to hidden layer (for $j = 1$ to the number of neurons m in the hidden layer)

I_i: scaled input i (for $i = 1$ to p)

W_{ij}: weight of neuron j corresponding to input I (for $j = 1$ to the number of neurons m in the hidden layer)

B_{1j}: weight of input bias of neuron j.

The value of each processing neuron is feed forward, becoming the input of the neuron in the output layer. The output of the NN model is calculated as shown in Equation 7. The output of the NN (O) is then scaled back to determine the estimated CMQ. Figure 4 shows the process between the input/hidden/output layers.

$$O = \tanh\left(\sum_{j=1}^{m}(H_j \times W_j) + 1.0 \times B_2\right) \tag{7}$$

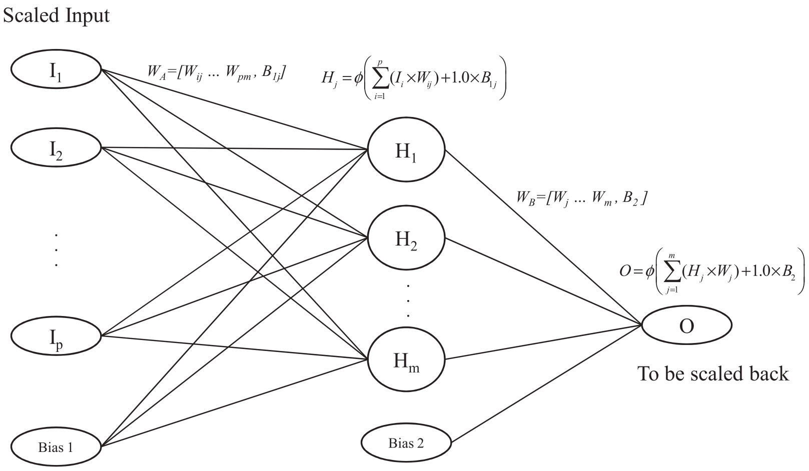

Scaled Input
I_1
I_2
I_p
Bias 1
$W_A=[W_{ij} \ldots W_{pm}, B_{1j}]$
$H_j=\phi\left(\sum_{i=1}^{p}(I_i \times W_{ij})+1.0\times B_{1j}\right)$
H_1
H_2
H_m
Bias 2
$W_B=[W_j \ldots W_m, B_2]$
$O=\phi\left(\sum_{j=1}^{m}(H_j \times W_j)+1.0\times B_2\right)$
O
To be scaled back

FIGURE 4 Graphical representation of the NN based on the selected RA model.

Where,

m: number of neurons
O: output of NN from hidden layer to output (scaled between –1 and 1)
H_j: output of neuron j (for $j = 1$ to the number of neurons m in the hidden layer)
W_j: weight of neuron j (for $j = 1$ to the number of neurons m in the hidden layer)
B_2: weight of output bias.

One of the disadvantages of using NN models is that they do not directly show how each IV (i.e., input) is related to the DV (i.e., output). It is, however, possible to use statistical software tools, such as SPSS (SPSS, 2010) or NeuroSolutions (NS, 2012), to perform sensitivity analyses to extract the cause and effect relationships between the inputs and outputs of the network, thereby determining the effect that each of the inputs has on each of the outputs (Koo et al., 2010). Other researchers (Olden & Jackson, 2002; Olden, Joy, & Death, 2004) use the connection weight approach and the Garson's algorithm to assess the importance of the connection weights and the contribution of the input variables to the output of the NN model. Often NN models are perceived as black-boxes (Kim et al., 2004), something that reduces the confidence in the results when compared to more traditional model types. There are, however, transparent ways to develop NN models, such as the one proposed by Hegazy et al. (1994), Hegazy and Ayed (1998), and Olden and Jackson (2002).

There is software (e.g., MATLAB, 2012; NS, 2012; SPSS, 2010) specialized for developing NN models. They allow certain flexibility and control by allowing the user to select among the different parameters that affect the network's performance (i.e., type of activation function, number of neurons in the hidden layer, type of objective function). Also, NN models can be done simulated using Generalized Reduced Gradient (GRG) nonlinear optimization to determine network weights using the Solver[1] add-in in MS Excel as explained by Hegazy and Ayed (1998). However, the simulation using the spreadsheet has limitations regarding the number of data points than can be used to generate the NN. For example, Solver, without special add-ins, is limited to 200 variables cells and 100 constraints; therefore, when more than 100 training data are used, the problem is too large for Solver to handle, and other software should be used to determine the weights of the NN. To avoid the data limitation of the Solver function, while giving control of the network functionality, the NN models for this study have been created using the manual configuration in SPSS.

Assessment of Model Performance

The concept of "model performance" is typically used very loosely and in general terms. For this study, the model performance was related to the accuracy of the model, (i.e., the error between the estimated and the actual quantities). There are many types of metrics to assess the performance of a model, and there is not a single one that provides an unambiguous indicator for model accuracy (Armstrong, 1985). In the literature reviewed, several metrics are used (e.g., Mean Absolute Percentage Error [MAPE], Root Mean Square Error [RMSE], Mean Square Error [MSE], coefficient of determination [R^2]), and in all cases, more than one metric was employed, the MAPE being the most popular one, used in all the studies (Table 2). These metrics are briefly covered in more detail in the following sections. The notation used in the equations below is the following:

CMQ_{e}: estimated value of the CMQ
CMQ_{a}: actual value of the CMQ
n: sample size.

MAPE. The mean absolute percentage error (MAPE) (Equation 8) is based on percentage errors. It is the most popular metric used to assess the performance of a model; however, despite its popularity, some researchers (Armstrong & Collopy 1992; Foss, Stensrud, Kitchenham, & Myrtveit, 2003; Makridakis, 1993) have pointed out some flaws about MAPE and indicated that it is a bias metric, as it puts a heavier penalty on equal errors when the estimate is greater than the actual, hence favoring low estimates.

$$\mathrm{MAPE}\,(\%) = \frac{100}{n} \times \sum \left(\frac{|CMQ_e - CMQ_a|}{CMQ_a} \right) \tag{8}$$

A study by Foss et al. (2003) showed that when using MAPE for choosing between two models, one would select the worst model. They found that the MAPE of the true model was higher (i.e., worst) than the MAPE of a model that consistently underestimated, yielding the selection of models that underestimate over the true model. To correct this flaw, a modified MAPE, referred to as adjusted or symmetric MAPE, was proposed (Armstrong, 1985), in which the difference between the actual and estimated amount are divided by the average of the actual and the estimated amount; however, for large errors, this correction makes the modified MAPE asymmetric (Goodwin & Lawton, 1999), treating large positive and negative errors very differently, so when large errors are expected, they recommended against using it.

ANOVA. Some researchers (Kim et al., 2004; Kim et al., 2013) have used an analysis of variance (ANOVA) to test the null hypothesis that the MAPEs of the different models are the same (H_0: $\mu_1 = \mu_2 = \ldots = \mu_n$; i.e., there is not a statistically significant difference among the MAPEs of the different models being compared).

MSE and RMSE. The Mean Square Error (MSE) (Equation 9) and the Root Mean Square Error (RMSE) (Equation 10) are scale-dependent metrics. The MSE was the preferred metric for comparing estimation models for a long time because of its computational convenience and theoretical relevance to statistics (Armstrong, 1985). In general, the RMSE is preferred to the MSE, as it is on the same units as the estimations (Hyndman & Koehler, 2006). The RMSE is also known as the standard error of the estimate (SEE) and typically calculated by most statistical packages when used to perform a simple RA[2]. However, empirical research (Armstrong & Collopy, 1992) has shown that the MSE and RMSE are unreliable, give more weight to larger errors than smaller errors, and are very sensitive to outliers, and hence are inappropriate for the evaluation of estimate accuracy and model selection (Armstrong & Collopy, 1992; Collopy & Armstrong, 2000).

$$\text{MSE} = \frac{\sum (CMQ_e - CMQ_a)^2}{n} \tag{9}$$

$$\text{RMSE} = \sqrt{MSE} \tag{10}$$

R^2. The coefficient of determination (R^2; Equation 11) is a metric of how well the regression equation fits the sample data. It is a very useful metric in RA but has not quite found its place in estimation (Makridakis & Hibon, 1995). R^2 is not a good metric to use for evaluating models because its value increases as more variables are included in the model. This means that the largest R^2 is simply obtained by including all IVs in the model, even though this is not necessarily the model that yields the most accurate estimates (Armstrong, 1985; Triola, 2001).

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{11}$$

Where,

$\sum (y_i - \hat{y}_i)^2$: unexplained variation or the sum of squared errors (SSE)
$\sum (y_i - \bar{y}\,)^2$: total variation or total sum of squares
y_i: actual values
$\hat{y}_i$: estimated y-value
$\bar{y}$: mean of the actual values.

In the case R^2-related metrics were to be used, the adjusted R^2 would be preferred. The adjusted R^2 penalizes a model based on the number of IVs used and the sample size (Field, 2009). The adjusted R^2 can be calculated in different ways, depending on the interest of the user. For example, the modified Wherry's formula (as cited in Armstrong, 1985), used in many standard software packages (e.g., Data Analysis in MS Excel, SPSS, R), is considered adequate to develop population expectations (Leach & Henson, 2003). However, the Lord's formula (Equation 12; Newman et al., 1979; Uhl & Eisenberg, 1970; Yin & Fan, 2001) should be used if estimation accuracy for future estimates is the main interest (Armstrong, 1985; Leach & Henson, 2003).

$$Adj\,R^2(L) = 1 - \left(\frac{n+k+1}{n-k-1}\right)\left(1 - R^2\right) \tag{12}$$

Where k is the number of IVs.

An Information Criterion

Although an AIC was not used in any of the studies shown in the literature review, it has been used by several researches (Blair, Lye, & Campbell, 1993; Burnham, Anderson, & Huyvaert, 2011; Castle, Qin, & Robert Reed, 2013; Jafarzadeh, Wilkinson, González, Ingham, & Amiri, 2013; Myung & Pitt, 1997; Panchal, Ganatra, Kosta, & Panchal, 2010; Posada & Buckley, 2004; Wagenmakers & Farrell, 2004;) in many fields (e.g., environmental science and ecology, economics, phycology, engineering) for model selection and can be used to overcome the disadvantages of the most common metrics addressed in the previous section.

The comparison of models using this metric is based on information theory. AIC is a measure of the amount of information lost when using a model as opposed to reality (Burnham et al., 2011). AIC can be used then to evaluate models and to select the one with the lowest AIC value. The model with the lowest AIC value is the one with the lower information loss, hence the one more likely to be the most accurate model from a set of models (Motulsky & Christopoulos, 2003).

AIC can be determined using the SSE (Equation 13) (Burnham et al., 2011). Although a second-order correction ($\mathrm{AIC_c}$) (Equation 14) should be used with small samples, (i.e., when the ratio of the sample size to the maximum K in a set is less or equal to 40), it is recommended to always use it, since as the sample size increases, $\mathrm{AIC_c} \approx \mathrm{AIC}$.

$$AIC = n \times \ln\left(\frac{SSE}{n}\right) + 2K \tag{13}$$

$$AIC_c = AIC + \left(\frac{2K\,(K+1)}{n-K-1}\right) \tag{14}$$

Where,

SSE: sum of squared errors
K: number of parameters in the model.

The value of the $\mathrm{AIC_c}$ of a given model by itself has no meaning. It becomes interesting when it is compared to the $\mathrm{AIC_c}$ of a series of models (Mazerolle, 2004). However, its interpretation might not be straightforward. To assist with that, the $\mathrm{AIC_c}$ can be used to determine the probability that the selected model is more likely to be the model with the least amount of information loss. Therefore, the probability (P) that one has chosen the correct model can be determined using Equation 15 (Motulsky & Christopoulos, 2003).

$$\mathrm{P} = \frac{e^{-0.5(\Delta AIC_c)}}{1 + e^{-0.5(\Delta AIC_c)}} \tag{15}$$

When comparing two models using Equation 15, say model A and model B, model A being the one with the lower $\mathrm{AIC_c}$, ($\Delta\mathrm{AIC_c} = \mathrm{AIC_{c\,B}} - \mathrm{AIC_{c\,A}}$), there is an x% probability that model B is the better model, and a (100–x)% probability that model A is the better model; in other words, model A is (100–x)/x times more likely to be the model that loses the least amount of information, hence the preferred model.

Process for Development and Evaluation of Estimation Models

Although many studies (e.g., Cho et al., 2013; Kim et al., 2004; Kim et al., 2013; Lowe et al., 2006; Smith & Mason, 1997; Sonmez, 2004; Yeh, 1998) have been conducted to develop estimation models using different techniques (e.g., RA, NN, CBR) and comparing them to determine which one performs better using different metrics (e.g., MAPE, RMSE, MSE, R^2), no systematic way has been proposed to determine the most suitable model for the estimation of CMQs.

A process (Figure 5) is proposed to be used for the development of estimation models, in this case using RA and NN techniques, as the two most prominent model types used to

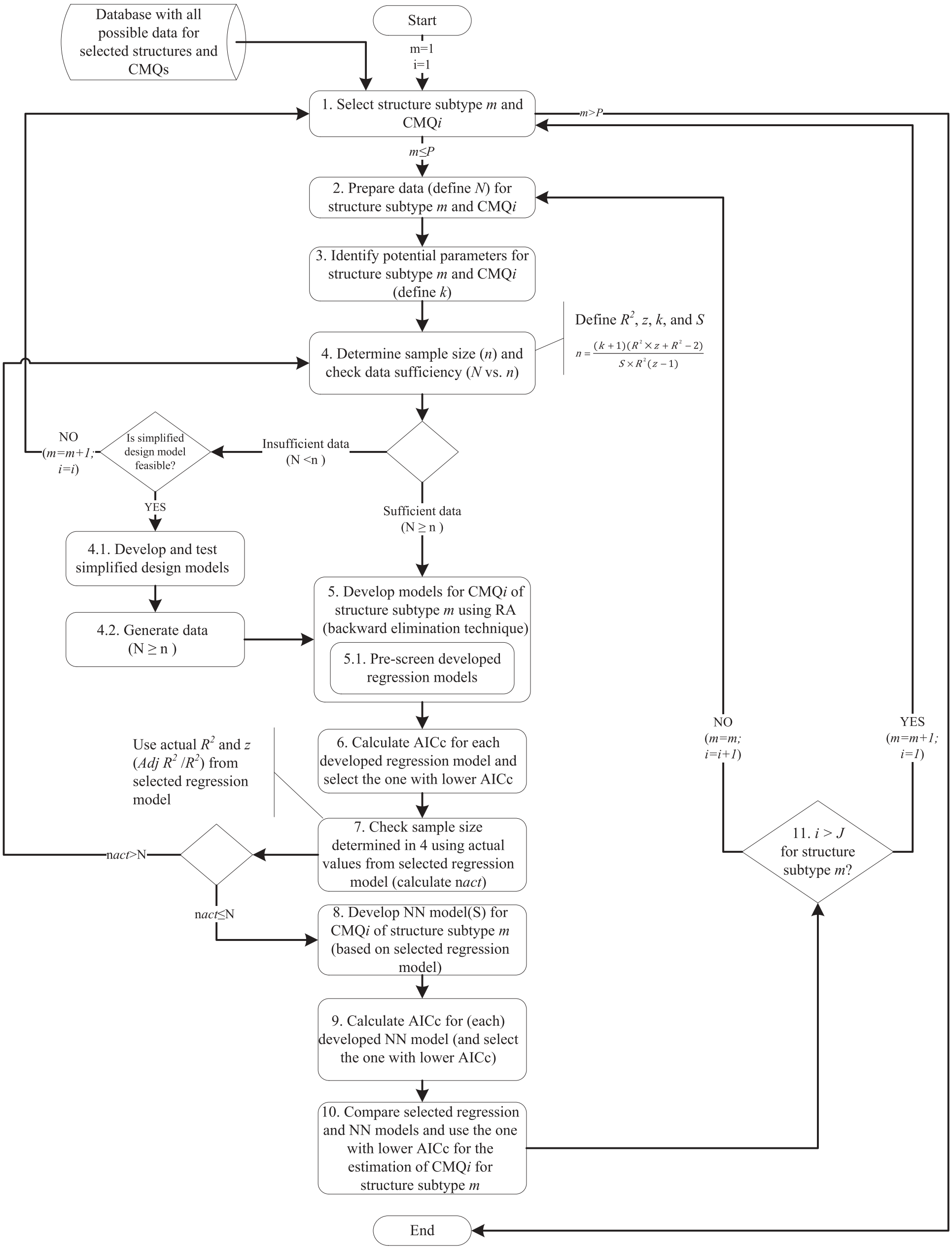

Database with all possible data for selected structures and CMQs
Start
m=1
i=1
1. Select structure subtype *m* and CMQ*i*
m>*P*
m≤*P*
2. Prepare data (define *N*) for structure subtype *m* and CMQ*i*
3. Identify potential parameters for structure subtype *m* and CMQ*i* (define *k*)
Define R^2, *z*, *k*, and *S*
$n = \frac{(k+1)(R^2 \times z + R^2 - 2)}{S \times R^2 (z-1)}$
4. Determine sample size (*n*) and check data sufficiency (*N* vs. *n*)
NO (*m*=*m*+*1*; *i*=*i*)
Is simplified design model feasible?
Insufficient data (N <n)
YES
Sufficient data (N ≥ n)
4.1. Develop and test simplified design models
4.2. Generate data (N ≥ n)
5. Develop models for CMQ*i* of structure subtype *m* using RA (backward elimination technique)
5.1. Pre-screen developed regression models
6. Calculate AICc for each developed regression model and select the one with lower AICc
Use actual R^2 and *z* (*Adj* R^2 /R^2) from selected regression model
7. Check sample size determined in 4 using actual values from selected regression model (calculate n*act*)
n*act*>N
n*act*≤N
8. Develop NN model(S) for CMQ*i* of structure subtype *m* (based on selected regression model)
9. Calculate AICc for (each) developed NN model (and select the one with lower AICc)
10. Compare selected regression and NN models and use the one with lower AICc for the estimation of CMQ*i* for structure subtype *m*
NO (*m*=*m*; *i*=*i*+*1*)
YES (*m*=*m*+*1*; *i*=*1*)
11. *i* > *J* for structure subtype *m*?
End

FIGURE 5 Process for the development and selection of models.

estimate CMQs and construction costs. For the evaluation of the models, $\mathrm{AIC_c}$ was used in lieu of the popular MAPE. Models with the lowest $\mathrm{AIC_c}$ were selected.

It is assumed that prior to starting this process, all the available data (e.g., dimensions, capacities, special characteristics, specific site characteristics, design parameters) have been gathered, and the structures, with the corresponding CMQs, have been carefully analyzed and selected by the estimator for their significance in the project for which models to estimate CMQs are to be developed and used in future projects.

From this gathering of data, all the possible structures for which models are to be developed (e.g., storage building type 1, storage building type 2, tall framed building 1, tall frame building 2) are identified. These structures can be grouped by type (e.g., storage buildings, tall framed building), which can be beneficial when developing the CMQ estimation models. There would be a total of *P* structure subtypes, referred to just as structures in the proposed process. Each identified structure contains *J* number of CMQs, which may be different from structure to structure (e.g., for storage building type 1, the CMQs might be concrete and reinforcement, while for storage building type 2, the CMQs might be concrete, reinforcement, and structural steel).

The proposed process includes both the situations where there are sufficient data and where there are not sufficient data. In the latter, data are to be generated using robust design models (Singh, 1990, 1991). To address this, a sub-process for data generation by developing simplified design models through Monte Carlo (MC) simulation is included.

Step 1: Select Structure m and CMQi

The structures and CMQs for which CMQ estimates are developed and evaluated depend on the specific project and is something to be determined by the estimator before starting this process. This process was done for the different selected structures (from structure *m* to *P*), such as storage buildings, framed buildings, and so on, and their corresponding CMQs (from CMQ *i* to *J*), such as concrete, reinforcing steel, structural steel, and so on. From this step, the required structures and corresponding CMQs should be clearly identified.

Step 2: Prepare Data

The collected data have to be prepared for further analysis. This step can consist of several components, including normalizing the data. The main purpose of normalizing the data is to make it homogeneous and consistent, hence checking it for consistency and adjusting it to eliminate any bias. There are guidelines available as to how data are collected and how it should be treated for further analysis (e.g., 2008 NASA Cost Estimating Handbook, 2008 International Society of Parametric Analysts Parametric Estimating Handbook, 2009 GAO Cost Estimating and Assessment Guide). Once the complete data for the selected structure and CMQ (*N*) have been prepared, the process can continue.

Step 3: Identify Potential Parameters

The identification of the potential parameters should be based on the information available to the estimator at the time of preparing the preliminary estimates and should use expert knowledge or information from previous research. The potential parameters should be selected to avoid redundancy; otherwise, the developed models might have problems with collinearity.

These parameters are further evaluated during the development of the models (see section "Step 4.1: Create Simplified Design Models") and finalized during the selection of the RA model (see section "Step 6: Calculate the AIC_c for RA Models and Select the One With Lower AIC_c"). In this step, all the potential parameters (*k*) to be used in the estimation of CMQ*i* for structure *m* are identified.

Step 4: Determine Sample Size Required for Model Development

A determination needs to be made to make sure that the gathered data (N) is sufficient to develop reliable models from a statistical point of view (Dupont & Plummer, 1998; Green, 1991; Kelley & Maxwell, 2003) so that N is greater than or equal to the minimum amount of data required for model development (n). Although there is not a consensus about the determination of n, it is agreed that the number of variables that can be used is limited by the sample size (Field, 2009). Therefore, it is important to ensure that there is enough data for the variables to be used and the reliability of the model.

Multiple rules of thumb to determine n have been developed (Bartlett, Kotrlik, & Higgins, 2001; Green, 1991; Mead, 1990); however, in many cases, their use could yield wrong determinations of the appropriate sample size (Brooks & Barcikowski, 2012; Field, 2009). It is proposed that the minimum sample size required be determined using Equation 16. This equation is derived by solving for the sample size n in the adjusted coefficient of determination equation by Lord (Equation 12) (Newman, 1979; Uhl & Eisenberg, 1970; Yin & Fan, 2001). It accounts for the shrinkage from R^2 to $Adj\ R^2$ (Equation 17) and the percentage of data to be used for training (i.e., developing) the model.

$$n = \frac{(k+1)(R^2 \times z + R^2 - 2)}{S \times R^2(z-1)} \tag{16}$$

$$Adj\,R^2 \geq z \times R^2 \tag{17}$$

Where,

n: minimum sample size (accounting for training and testing of model)
k: maximum number of IVs expected in the model
R^2: minimum desired coefficient of determination, in decimal form (i.e., amount of the total variation explained by the IVs)
z: allowable shrinkage from R^2 to $Adj\ R^2$, in decimal form (i.e., $Adj\ R^2/R^2$ or the reduction from R^2 expected to be captured in the $Adj\ R^2$)
S: percentage of data to be used for training the model, in decimal form.

Once n has been calculated, it is to be compared to N to determine if the existing data are enough or if more data should be generated (i.e., development of simplified design models). It is important to point out that when using this equation, preliminary values for the expected R^2 and allowable shrinkage (z) should be checked by recalculating the minimum sample size (n_{actual}) using the actual R^2, z, and number of IVs (k) from the selected model (see section “Step 6: Calculate the $\mathrm{AIC_c}$ for RA Models and Select the One With Lower $\mathrm{AIC_c}$”).

Step 4.1: Create Simplified Design Models. When data are not sufficient ($n > N$), consideration should be made to decide if additional work should be made carried out to collect more data or, if more data are not available, if it should be generated using appropriate design models. The development of such design models might require a fair amount of time and effort. For this reason, a decision should be made at this point whether or not the added work in development of simplified deign models is worthwhile. If it is not, the structure for which not sufficient data are available should be dropped from the list of significance structures for the project.

Simplified design models can be developed in several ways. All of them require an experienced structural engineer who can prepare the models in a way that they could be later automated to generate data that can be used for the development of models to estimate CMQs. The simplified design models should be validated using the existing data when possible. The data generated using simplified design models (see section "*Step 4.2:* Generate Data") should be enhanced, and eventually the estimation models should be replaced with models using real data when it becomes sufficient, as the data from simplified design models does not capture all aspects of the real constructions.

Step 4.2: Generate Data. Once the simplified design model has been developed and validated, it can be used to generate CMQ data (e.g., Figure 6). To do that, the possible values of the input parameters are determined based the available information, expert opinion, or previous research. The design calculations for the design model are done using the different load combinations used to calculate the required size of each of the elements in the structure. Once the sizes of the elements have been determined so that all code requirements are

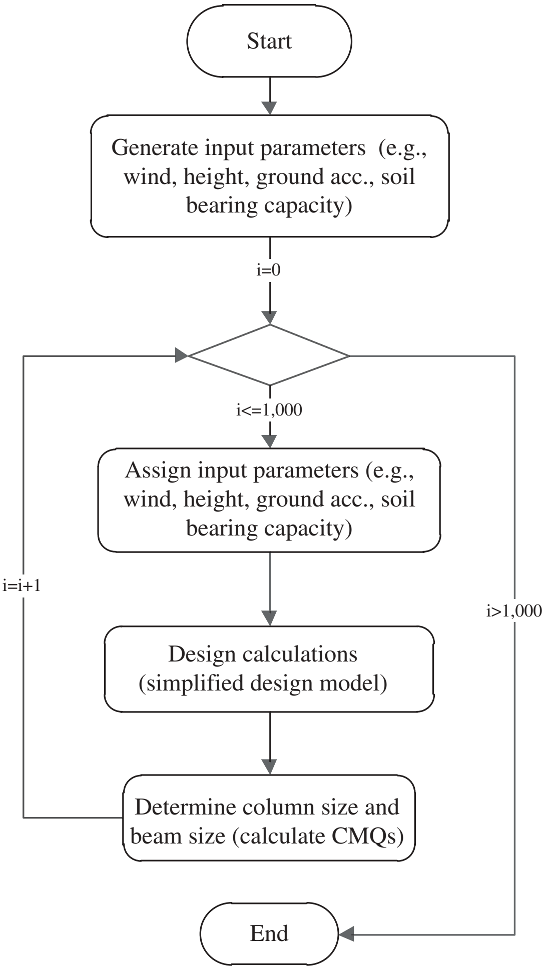

Start
Generate input parameters (e.g., wind, height, ground acc., soil bearing capacity)
i=0
i<=1,000
Assign input parameters (e.g., wind, height, ground acc., soil bearing capacity)
i=i+1
i>1,000
Design calculations (simplified design model)
Determine column size and beam size (calculate CMQs)
End

FIGURE 6 Sample process for data generation using simplified design models.

met, then CMQs for each structural element is determined (e.g., columns, beams, slabs). The input/output data set can be used to develop the estimation models.

Step 5: Develop RA Models

Using all the potential parameters identified in the section "Step 3: Identify Potential Parameters," develop RA models using the BET.

Step 5.1: Pre-Screen Developed RA Models. In addition to the standard checks made on the developed RA models (e.g., assumptions related to RA, such as significance of variables, collinearity, no correlation between IVs and residuals, residuals are homoscedastic, uncorrelated, and normally distributed; Field, 2009), it is recommended to check that the described relationships, seen through the coefficients of the variables, behave as expected. Sometimes, especially if the data used for model development has a lot of noise, the models may show relationships that are nonsensical (e.g., the larger a building, the less concrete it requires). Constrained RA models should be developed if models have nonsensical relationships. The constraint consists on controlling the sign of the variable that is causing the nonsensical relationship while running the regression.

Step 6: Calculate the AIC$_c$ for RA Models and Select the One with Lower AIC$_c$

Once the set of RA models derived from the section "Step 5: Develop RA Models" have been screened (see section "Step 5.1: Pre-Screen Developed RA Models") the AIC$_c$ for the remaining models is calculated using Equation 14 (see section "An Information Criterion"). The variable of K should include also the constant and error terms (i.e., number of IVs plus 2). The RA model with the lowest AIC$_c$ is selected as the RA model with the lowest amount of information loss (i.e., the RA model that is more likely to be correct for the structure and CMQ being evaluated). Equation 15 (see section "An Information Criterion") is used to determine the relative probability of the models being the model with the least amount of information loss, when compared to the model with the lower AICc.

Step 7: Check Sample Size Determined in the section "Step 4: Determine Sample Size Required for Model Development"

Using the RA model with the lowest AIC$_c$, the assumptions made to determine the minimum amount of data needed (see section "Step 4: Determine Sample Size Required for Model Development") shall be checked by recalculating the minimum sample size (now referred to as n_{actual}) using Equation 16 with the same variable for S and the actual values for R^2, $Adj\ R^2$ and the number of IVs (k) from the selected RA model. The n_{actual} is then compared with the data used to develop the model (N).

If the sample size is sufficient ($n_{actual} \leq N$) then the selected RA model is confirmed and the process can continue. If the sample size is not sufficient ($n_{actual} > N$) then the models developed in the "Step 5: Develop RA Models" section should be discarded and the values assumed in the "Step 4: Determine Sample Size Required for Model Development" section should be revised and the process restarted from there.

Step 8: Develop NN Model

The NN model is developed based on the selected RA model. The NN models considered in this process are the feed-forward NNs with three layers (one hidden layer) and back-propagation learning algorithm, using the SSE as the objective function. The characteristics of the NN models developed in this process are:

1. Input neurons = IVs (dummy coding not required)
2. Input scaled between –1 and +1
3. Hyperbolic tangent activation function between input/hidden and hidden output layers
4. One hidden layer (with number of neurons determined heuristically)
5. Objective function to optimize weights = sum-of-squares error
6. Use data split during the learning process
7. One output

To avoid over fitting and ensure generalization of the NN model, several NN models should be developed, each with a different number of neurons in the hidden layer.

Step 9: Calculate the AIC$_c$ for (Each) Developed NN Model (and Select the One with Lower AIC$_c$)

For the NN models derived from the "Step 8: Develop NN Model" section, the AIC$_c$ is then calculated using Equation 14 (see section "An Information Criterion"). When Equation 14 was used to select among different NN models, the variable K shall be modified to be the total number of weights in the NN model. The NN model with the lowest AIC$_c$ is selected as the NN model with the lowest amount of information loss (i.e., the NN model that is more likely to be correct for the structure and CMQ being evaluated).

For the selected NN model, or when only one NN model has been developed in the "Step 8: Develop NN Model" section, the AIC$_c$ should be calculated by using K as the number of IVs, including the bias neuron. This AIC$_c$ should be used when comparing it to the AIC$_c$ from the RA model in the next step.

Step 10: Compare Selected RA and NN Models and Use the One with Lower AIC$_c$

The evaluation of the RA and NN models is based on the AIC$_c$.

The model with the lowest AIC$_c$ value is the one with the least amount of information loss, or the model more likely to be the correct model. The selected model was used to estimate the CMQ of the structure being evaluated.

Next CMQ or Structure

Once the process is completed for a given CMQ, the next CMQ for the structure evaluated was used. If all the CMQs for a given structure have been completed then, the next structure with its corresponding CMQs is introduced in the process.

The process ends when models for all the CMQs for all the structures have been developed and evaluated. At the end of the process, the estimator has developed and selected the models for the estimation of CMQs for the structures identified for the project.

Example

The proposed process for the development and evaluation of RA and NN models, as described in the "Process for Development and Evaluation of Estimation Models" section and illustrated in Figure 5, is implemented in this example assuming that an estimator is going over this process. Models for the estimation of concrete and reinforcing steel (reinforcement) required in the construction of storage buildings are developed. A total of 58 storage buildings from eight plants located around the word were used. The storage buildings were classified based on the density of the material they stored, as it would affect the configuration and characteristics of these structures (Table 3).

In the interest of space, both CMQs for which models are to be developed (concrete and reinforcement) have been presented and discussed simultaneously. The estimator using the proposed process, however, would be doing one CMQ at the time.

Step 1: Select Structure m and CMQi

For this example, models to estimate the amount of concrete and reinforcement required in the construction of storage buildings (of the types shown in Table 3) are developed.

Step 2: Prepare Data (Define N)

Data from a total of 58 storage buildings was collected. The information is related to the general site characteristics for the different plants (e.g., design wind speed, soil bearing capacity, spectral response acceleration; Figure 7) and specific structure characteristics (e.g., storage material, dimensions, capacity, type, CMQs; Figure 8), all of which would be readily available during the early stages of a project. The information related to the CMQs (i.e., concrete and reinforcement) was obtained from the final bill of quantities reported by the contractors, and total concrete and reinforcement quantities were used (Figure 9). The data went through several checks (e.g., investigation of outliers, uniformity of units) to ensure it was suitable for further analysis. The reinforcement area/concrete area ratios (As/Ag) for all the structure used ranged between 1.72 and 2.68% with an average value of 2.13% (SD = 0.23 percentage points). All within the code specified maximum values of 1-3% (based on ACI codes; Figure 10).

Data transformation was considered to improve the relationship between the different variables. Data transformation by taking the natural logarithm of all continuous variable values was performed. When RA models developed using the raw datasets (i.e., without transformation) were compared to RA models developed using the transformed (LN) datasets, an improvement in the normality of the data and the coefficient of determination (R^2) was observed. Similarly, the Pearson correlation coefficient improved in all cases.

TABLE 3 Summary of storage capacities for the different storage building types

Storage building type	Number
Type A	18
Type B	24
Type C	16

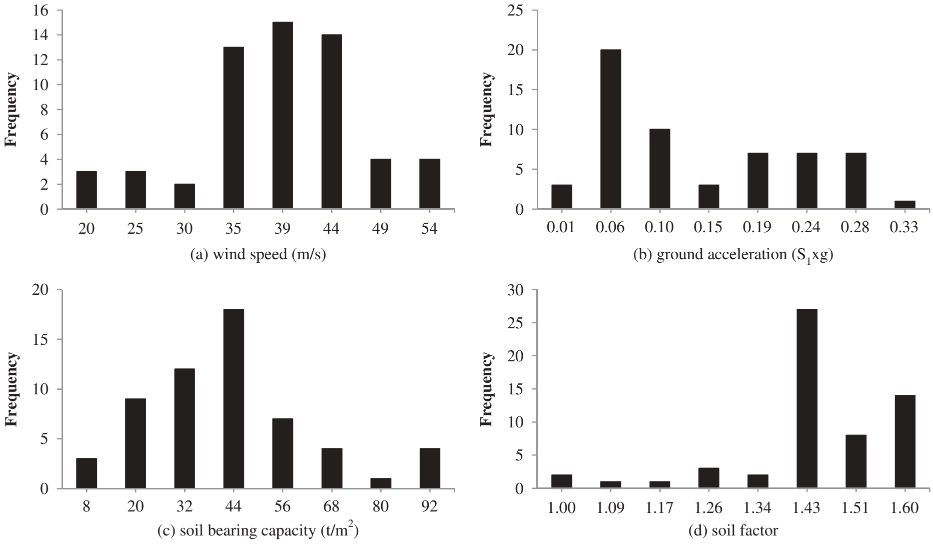


FIGURE 7 Histograms of variables for general site characteristics.

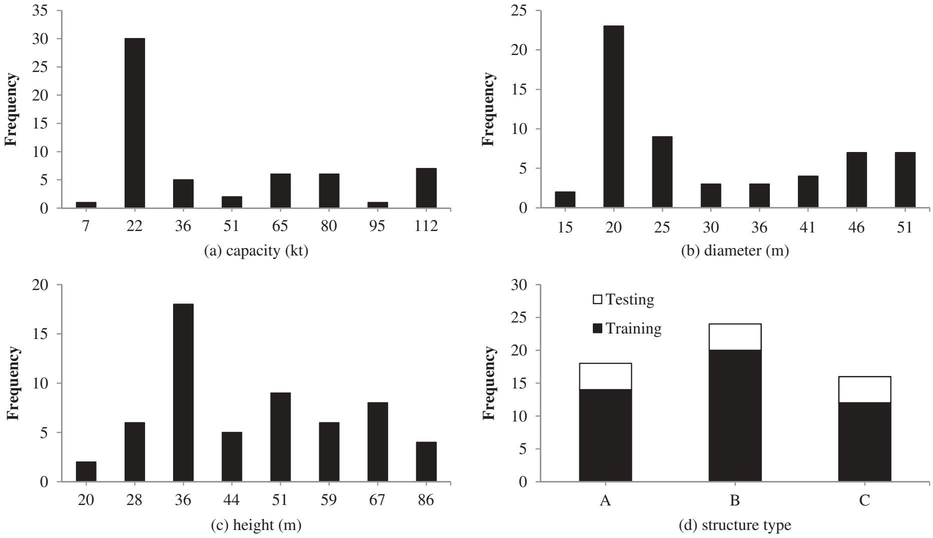


FIGURE 8 Histograms of variables for specific structure characteristics.

Step 3: Identify Potential Parameters

From the data collected, the proposed parameters were identified. Seven IVs were selected taking into consideration the limited information available to the estimator when preparing the preliminary estimates. The maximum, minimum, and average values for the variables related to storage buildings are summarized in Table 4.

There are three categorical variables, one for each structure type (Type A, Type B, and Type C). When using these variables in the regression models, the variables were recoded (Storage Type I, Storage Type II, and Storage Type III) as 0 and 1, with 1 indicating that the associated observation has the given categorical value. Any two of the three recoded

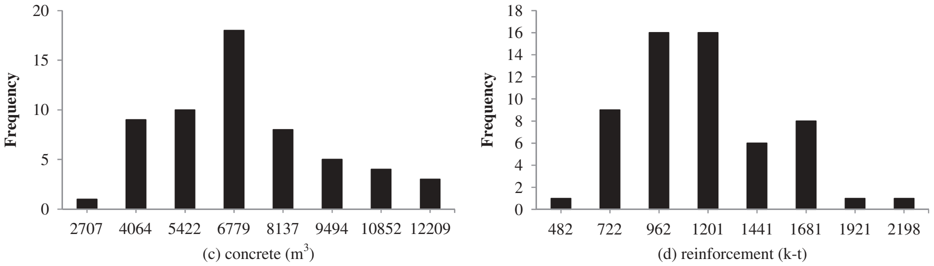


FIGURE 9 Histograms for CMQs.

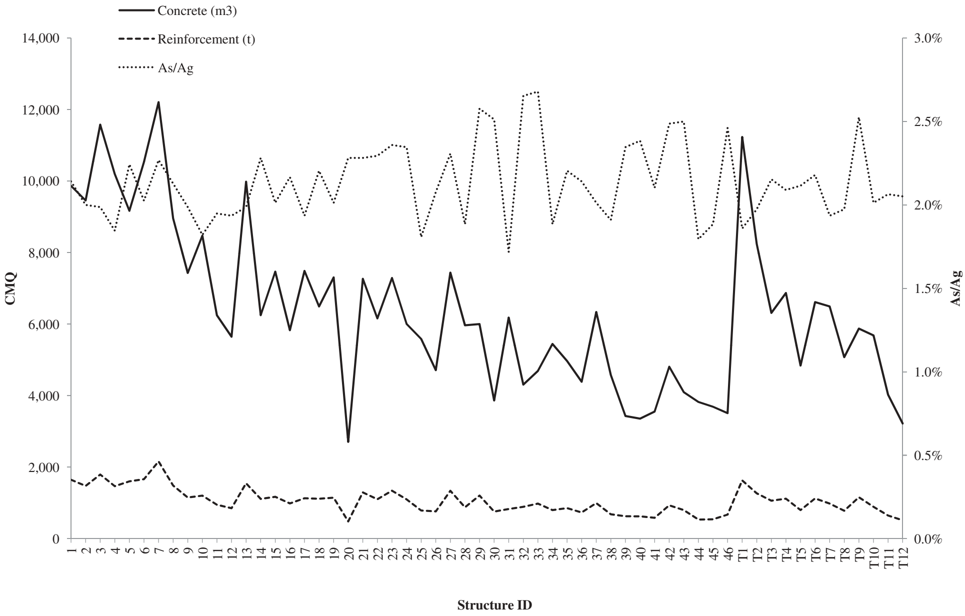


FIGURE 10 CMQ for concrete and reinforcement and A_s/A_g.

variables should be included in the model. In this case, Type III was excluded and used as the reference variable.

Step 4: Determine Sample Size (n) and Check Data Sufficiency

Equation 16 was used to determine the sample size required for the development of models. In this case, the maximum number of IVs (k) that could be used in the models is nine (out of the three categorical variables, Type III was excluded from the model and used as the reference variable). The expectation is that the variables used should explain not less than 90% of the variation in the model ($R^2 \geq 0.9$), and the reduction of R^2 that should be captured in the *Adj* R^2 is at least 90% (i.e., *Adj* $R^2 \geq 0.81$; $z = 0.9$). A typical 4/1, or 80%:20% random data split for training and testing the models, respectively, was used ($S = 0.8$). Using this information in Equation 16, the following is obtained:

$$n = \frac{(9+1)\,(0.9 \times 0.9 + 0.9 - 2)}{0.8 \times 0.9(0.9 - 1)} = 40.3 \qquad (18)$$

TABLE 4 List of variables and description of data for storage buildings (as-built data)

	ID	Code	Description	Max	Min	Average	Type	Units
Input	1	Cap_t	Storage capacity	112,000	7,000	36,000	Continuous	Tons
	2	Diam_m	Interior diameter	51	15	29	Continuous	Meters
	3	Total_Height_m	Height from top of foundation to top of concrete structure	86	24	56	Continuous	Meters
	4	Wind_Speed_mps	Design wind speed in accordance with the Eurocode 1, EN 1991 1-4 (2010; wind design)	54	20	36	Continuous	Meters/second
	5	S1_g	Spectral response acceleration for 1.0 sec. period (2% probability of exceedance in 50 years)	0.33	0.01	0.10	Continuous	Gravitational acceleration expressed in decimal form as a function of g ($S1_x g$)
	6	Soil_BC_tpm2	Soil bearing capacity	92	7.5	37	Continuous	Tons/m^2
	7	Soil_Factor	Soil factor/soil coefficient in accordance with the Eurocode 8, EN 1998 1-6 (2006)	1.6	1.0	1.4	Continuous	n/a
	8	Type_I	Storage Type A	0	1	n/a	Categorical	n/a
	9	Type_II	Storage Type B	0	1	n/a	Categorical	n/a
	10	Type_III	Storage Type C (used as reference)	n/a	n/a	n/a	Categorical	n/a
Output	11	Concrete	Total concrete	12,209	2,707	6,293	Continuous	m^3
	12	Reinforcement	Total reinforcement	2,198	487	1,133	Continuous	Tons

From Equation 18 it is determined that the total minimum sample size required for model development is 40.3 (say 41), from which 33 would be used for the development of models and the rest for testing them. Since there are 58 storage buildings available, $n < N$ and the process can continue with the development of models without the need of developing simplified design models to generate additional data.

The 58 structures were combined and assigned a random generated ID (unique between 1 and 58). That process was done for a random number of times, after which the random generated IDs were locked and sorted from smallest to largest. Once that was done, four structures from the top, middle, and bottom of the randomly sorted dataset were selected (i.e., structures with the following random generated IDs: 1–4, 28–31, 55–58). That is, 12 structures (approximately 20%) were randomly selected and used to test the models (structure IDs T1 through T12 in Figure 10) and 46 structures (approximately 80%) were randomly selected and used for model development (i.e., training; structure IDs 1 through 46 in Figure 10). The assumptions made here (e.g., k, R^2, and z) to determine the sample size requirements must be checked once the RA model has been selected (see section "Step 7: Check Sample Size Determined in the Section 'Step 4: Determine Sample Size (n) and Check Data Sufficiency'").

Step 4.1: Create Simplified Design Models. As $n < N$ (see section "Step 4: Determine Sample Size (n) and Check Data Sufficiency"), this step is not applicable in this example.

Step 4.2: Generate Data. As $n < N$ (see section "Step 4: Determine Sample Size (n) and Check Data Sufficiency"), this step is not applicable in this example.

Step 5: Develop RA Models

The RA models were developed using transformed data (Equation 2). Using the BET, a total of nine models were developed for both concrete and reinforcement quantities, four for concrete and five for reinforcement. The unstandardized coefficients for the different models are summarized in Table 5. All the variables were within the specified significance level ($\alpha = 0.05$) and acceptable variance inflation factors (i.e., multicollinearity was not an issue; O'Brian, 2007).

From Table 5, one can see that the backward regression technique eliminates the same continuous variables at each step. This indicates that for the concrete and reinforcement models, the variables used have a similar effect. In general, as the amount of concrete increases, the amount of reinforcement is expected to increase as well. This was anticipated based on the $\mathrm{A_s/A_g}$ values (see section "Step 2: Prepare Data (Define N)") and the relationship between concrete and reinforcement for the structures used (Figure 11).

For the subsequent steps, the back-transformed models were used (Equation 3). All the statistics used and reported (e.g., R^2, z, $\mathrm{AIC_c}$) are then corresponding to the back-transformed models.

Step 5.1: Pre-Screen Developed RA Models. The developed models are checked to ensure that the described relationships, seen through the coefficients of the variables, behave as expected. For example, the unstandardized coefficient for wind speed (LN_Wind) in models CO-RA1, CO-RA2, and CO-RA3 is negative (–0.028), indicating that a 1% increase in the wind speed would decrease the CMQ of concrete by 0.028%. Hence, models where the coefficient for wind speed is negative should be constrained (i.e., by controlling the sign of the coefficients) or not considered. In this case, only CO-RA4 should be considered. However, for illustration purposes, in this example, all the models were used for the required calculations (i.e., $\mathrm{AIC_c}$) in subsequent steps.

TABLE 5 Summary of RA models for concrete and reinforcement quantities

Concrete (CO) Model*		Unstandardized Coeff. B	Std. Error	Reinforcement (RE) Model		Unstandardized Coeff. B	Std. Error
CO-RA1	(Constant)	1.904	0.4232	RE-RA1	(Constant)	0.6501	0.5998
	LN_Cap	0.4269	0.04634		LN_Cap	0.4260	0.06545
	LN_Diam	0.7852	0.1071		LN_Diam	0.7640	0.1513
	LN_Height	0.09801	0.06114		LN_Height	0.04843	0.08624
	LN_Wind	−0.02842	0.05223		LN_Wind	0.1230	0.07451
	LN_S1	0.5956	0.1905		LN_S1	0.8072	0.2692
	LN_SoilBC	0.001424	0.0284		LN_SoilBC	−0.001135	0.0465
	LN_SoilFact	−0.03263	0.1696		LN_SoilFact	0.06902	0.2397
	Type I	−0.1770	0.03992		Type I	−0.1090	0.05582
	Type II	−0.8846	0.1223		Type II	−0.9112	0.1739
CO-RA2	(Constant)	1.912	0.3650	RE-RA2	(Constant)	0.6452	0.5167
	LN_Cap	0.4277	0.04503		LN_Cap	0.4263	0.06464
	LN_Diam	0.7842	0.1034		LN_Diam	0.7642	0.1455
	LN_Height	0.09831	0.06013		LN_Height	0.04813	0.08533
	LN_Wind	−0.02863	0.05135		LN_Wind	0.1237	0.07271
	LN_S1	0.5937	0.1760		LN_S1	0.8096	0.2495
	LN_SoilFact	−0.03481	0.1577		LN_SoilFact	0.07064	0.2233
	Type I	−0.1773	0.03814		Type I	−0.1095	0.05483
	Type II	−0.8837	0.1207		Type II	−0.9124	0.1698

(*Continued*)

TABLE 5 (*Continued*)

Concrete (CO) Model*		Unstandardized Coeff. B	Std. Error	Reinforcement (RE) Model		Unstandardized Coeff. B	Std. Error
CO-RA3	(Constant)	1.893	0.3507	RE-RA3	(Constant)	0.6838	0.4966
	LN_Cap	0.4298	0.04462		LN_Cap	0.4218	0.06244
	LN_Diam	0.7812	0.1026		LN_Diam	0.7754	0.1428
	LN_Height	0.09631	0.05823		LN_Height	0.05363	0.08351
	LN_Wind	−0.02842	0.05012		LN_Wind	0.1241	0.07132
	LN_S1	0.5825	0.1673		LN_S1	0.8323	0.2378
	Type I	−0.1761	0.03701		Type I	−0.1114	0.05361
	Type II	−0.8831	0.1185		Type II	−0.9125	0.1677
CO-RA4	(Constant)	1.823	0.3234	RE-RA4	(Constant)	0.6661	0.4929
	LN_Cap	0.4264	0.04334		LN_Cap	0.4437	0.05122
	LN_Diam	0.7735	0.09812		LN_Diam	0.7865	0.1393
	LN_Height	0.1021	0.05714		LN_Height	0.1323	0.06944
	LN_S1	0.5732	0.1650		LN_S1	0.8777	0.2245
	Type I	−0.1761	0.03773		Type I	−0.1222	0.04953
	Type II	−0.8681	0.1140		Type II	−0.9884	0.1172
				RE-RA5	(Constant)	0.5398	0.4845
					LN_Cap	0.3952	0.06525
					LN_Diam	0.6470	0.1435
					LN_Height	0.1357	0.08124
					LN_S1	0.6956	0.2466
					Type II	−0.6673	0.1472

*CO-RA# means model for Concrete (CO) using the backward elimination regression technique (RA). The # indicates the number of the model developed using that technique. RE stands for reinforcement. The number of significant figures has been taken as 4 to be consistent to the number of significant figures from the tables in the "Validation" section.

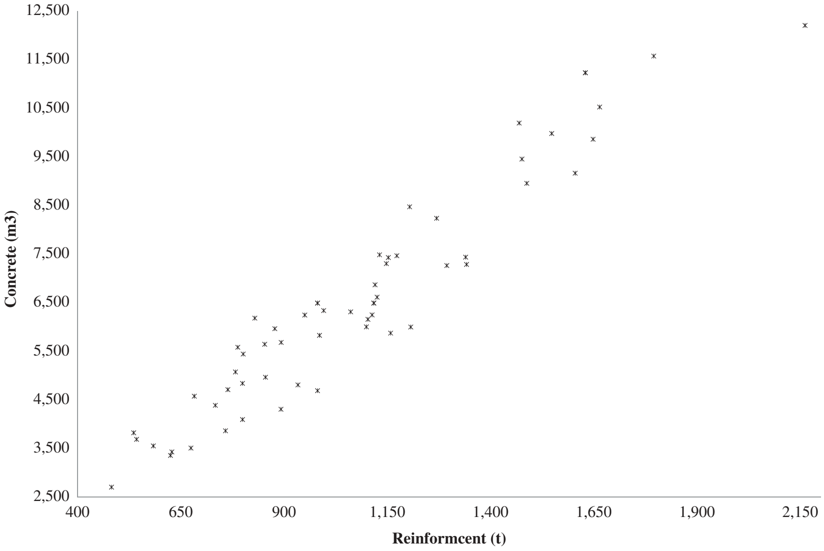


FIGURE 11 Relationship between concrete and reinforcement for the structures used.

Step 6: Calculate the AIC$_c$ for RA Models and Select the One with Lower AIC$_c$

Table 6 shows the AIC$_c$ for the different models, as well as the actual *k, R^2*, and *z* values for the back-transformed datasets. From the set of models analyzed, the one with the lowest AIC$_c$ are CO-RA4 (592.86) and RE-RA5 (461.33) for concrete and reinforcement, respectively.

As stated in the section "An Information Criterion," the AIC$_c$ value by itself has no meaning. It becomes useful when it is compared to the AIC$_c$ of other models in a set (ΔAIC$_c$). To assist with the interpretation of the AIC$_c$, the probability that the selected model is more likely to be the model with the least amount of information loss is determined.

To determine ΔAIC$_c$, the minimum AIC$_c$ is kept fixed when determining the probability of all the models in Equation 15. For example, when comparing the CO-RA2 model with the CO-RA4 model (ΔAIC$_c$ =601.7–592.9), one can see that there is a 1.213% probability that the CO-RA2 model is the correct model and a 98.79% probability that the CO-RA4 model is the correct model; in other words, the CO-RA4 model is 81 times more likely to be the correct model (i.e., the one with the least amount of information loss) than the CO-RA2 model.

Step 7: Check Sample Size Determined in the Section "Step 4: Determine Sample Size (n) and Check Data Sufficiency"

Plugging in the actual values for *k, R^2*, and *z* (Table 6) for the selected models (i.e., for CO-RA4 [k = 6, R^2 = 0.9531, z = 0.9822] and for RE-RA5 [k = 5, R^2 = 0.8863, z = 0.9613]) into Equation 16, the n_{actual} is obtained. The value for the percentage of data to be used for training the model is kept the same (S = 0.8).

The n_{actual} for the concrete and the reinforcement models is 57, hence $n_{actual} \leq N$. Since for both of the selected models the sample size requirements are met, the selected RA models are confirmed. The process can continue.

TABLE 6 Summary of AIC$_c$, k, R^2 and z for RA models for concrete and reinforcement

Model ID	AIC_c	k	R^2	z	Model ID	AIC_c	k	R^2	z
CO-RA1	605.3	9	0.9501	0.9713	RE-RA1	471.6	9	0.8923	0.9332
CO-RA2	601.7	8	0.9501	0.9752	RE-RA2	468.1	8	0.8924	0.9414
CO-RA3	598.0	7	0.9512	0.9782	RE-RA3	465.4	7	0.8912	0.9485
CO-RA4	592.9	6	0.9531	0.9822	RE-RA4	464.1	6	0.8862	0.9544
					RE-RA5	461.3	5	0.8863	0.9613

Step 8: Develop NN Model

Using the selected RA model to determine the input variables, the NN models for concrete and reinforcement were developed with the characteristics determined in the "Develop NN Model" section. The number of neurons in the hidden layer of the NN model was determined empirically. For the estimation of concrete and reinforcement quantities, the NN models with the lowest generalization error and AIC$_c$ had 3 neurons in the hidden layer. Table 7 shows the weights between the input and the hidden layer and the hidden layer and the output for the NN models for concrete and reinforcement.

Step 9: Calculate the AIC$_c$ for (Each) Developed NN Model (and Select the One with Lower AIC$_c$)

The AIC$_c$ for the NN model is calculated using Equation 14 with K equal to the number of inputs and the bias neuron (6). The AIC$_c$ for the concrete and reinforcement NN models are 613 and 474, respectively.

Compare Best RA and NN Models and Use the One with Lower AIC$_c$

The AIC$_c$ for the RA and NN models is summarized in Table 8. The AIC$_c$ for the RA models is lower than the NN models; hence, the models selected for the estimation of concrete and reinforcement for storage buildings are CO-RA4 and RE-RA5, respectively.

Next CMQ or Structure

As indicated at the beginning of this example, both CMQs for which models are to be developed (concrete and reinforcement) have been presented and discussed simultaneously; therefore, for this example, this step is not applicable.

The estimator using the proposed process, however, would be doing one CMQ at the time.

Validation

Using the proposed process to develop and evaluate RA and NN models led to the selection of the CO-RA4 and RE-RA5 models for the estimation of the amount of concrete and reinforcement for storage buildings, respectively.

The validation of the selected models was done using the 12 data points not utilized for model development. The testing data were fed into the developed RA and NN models and the accuracy ratio (Equation 19) (Armstrong, 1985) was determined. The accuracy ratio is unit free, simple to calculate, and easy to understand. The closer the accuracy ratio is to 1, the more accurate the model is in terms of its deviation between the estimated and actual quantity (Q).

$$\text{Accuracy Ratio} = \frac{\sum Q}{n} \tag{19}$$

Where,

$$Q = \begin{cases} \dfrac{CMQ_a}{CMQ_e} \text{ if } CMQ_a > CMQ_e \\ \dfrac{CMQ_e}{CMQ_a} \text{ if } CMQ_e > CMQ_a \end{cases}$$

TABLE 7 Summary of estimated weights for NN model using 5 inputs and 3 neurons for concrete and reinforcement

		Concrete				Reinforcement			
		Input/Hidden Layer (W_A)			Output Layer	Input/Hidden Layer (W_A)			Output Layer
Input		H_1	H_2	H_3	Total Concrete (m^3)	H_1	H_2	H_3	Total Reinforcement (t)
Input Layer	I_1: Cap_t	0.377	−0.535	−0.385	—	−0.538	0.371	−0.146	—
	$I_{2:}$ Diam_m	0.654	−0.526	0.149	—	−0.119	0.337	−0.476	—
	$I_{3:}$ Height_m	0.192	0.281	−0.675	—	−0.412	0.142	−0.351	—
	$I_{4:}$ S1	−0.297	−0.376	−0.417	—	−0.001	0.051	−0.409	—
	I_5: Type_ID	0.402	−0.098	0.318	—	0.093	0.140	0.034	—
	Bias 1	0.138	0.733	0.032	—	0.073	−0.144	−0.093	—
Hidden/Output Layer (W_B)	H_1	—	—	—	0.721	—	—	—	−0.543
	H_2	—	—	—	−0.533	—	—	—	0.447
	H_3	—	—	—	−0.677	—	—	—	−0.492
	Bias 2	—	—	—	0.246	—	—	—	0.007

TABLE 8 Summary of $\mathrm{ACI_c}$ for RA and NN models for concrete and reinforcement

Model ID	$\mathrm{AIC_c}$
CO-RA4	592.9
CO-NN	613.0
RE-RA5	461.3
RE-NN	474.2

TABLE 9 Accuracy ratio for developed RA and NN models using testing data

Model ID	Accuracy Ratio	Model ID	Accuracy Ratio
CO-RA1	1.039	RE-RA1	1.079
CO-RA2	1.039	RE-RA2	1.079
CO-RA3	1.037	RE-RA3	1.082
CO-RA4	1.036	RE-RA4	1.086
CO-NN	1.111	RE-RA5	1.075
		RE-NN	1.132

TABLE 10 MAPEs for RA and NN models developed (training set)

Model ID	MAPE (%)	Model ID	MAPE (%)
CO-RA1	5.495	RE-RA1	9.081
CO-RA2	5.489	RE-RA2	9.078
CO-RA3	5.504	RE-RA3	9.058
CO-RA4	5.497	RE-RA4	9.041
CO-NN	8.057	RE-RA5	9.823
		RE-NN	11.44

In addition to the accuracy ratio, and due to its popularity, the MAPE was used to determine the model accuracy (Equation 20) and as the selection metric for the developed models. This allowed the $\mathrm{AIC_c}$ to be further checked as the metric used in the proposed process to select the most accurate model and complement the validation process.

The results of the validation using the accuracy ratio are summarized in Table 9.The models with the accuracy ratio closest to one for the estimation of concrete and reinforcement are CO-RA4 and RE-RA5 with an accuracy of 1.036 and 1.075, respectively. These models correspond to the models selected during the proposed process (see section "Compare Best RA and NN Models and Use the One With Lower AICc"). This indicates that the evaluation of models using $\mathrm{AIC_c}$ is adequate to select the most accurate models for the estimation of CMQs in future projects.

The MAPE for the different developed models using the training set was calculated using Equation 8. The results are summarized in Table 10.

Using MAPE as the evaluation metric to select the most accurate models would lead to choosing CO-RA2 for concrete and RE-RA4 with the lower MAPE in each set with a value of 5.489% and 9.041%, respectively. The MAPE and the model accuracy for all the models were determined using the testing data. The results are summarized in Table 11. The closer the model accuracy is to 100%, the better.

TABLE 11 MAPE and model accuracy for developed RA and NN models using testing data

Model ID	MAPE (%)	Model Accuracy (%)	Model ID	MAPE (%)	Model Accuracy (%)
CO-RA1	3.785	96.22	RE-RA1	7.776	92.22
CO-RA2	3.795	96.21	RE-RA2	7.775	92.23
CO-RA3	3.591	94.41	RE-RA3	7.998	92.00
CO-RA4	3.475	96.53	RE-RA4	8.420	91.58
CO-NN	9.307	90.70	RE-RA5	7.437	92.56
			RE-NN	11.80	88.21

$$\text{Model Accuracy}\,(\%) = \max(100 - \mathit{MAPE},\ 0) \tag{20}$$

The model accuracy for the CO-RA2 and RE-RA4 models is 96.21% and 91.58%, respectively; however, there are other models that are more accurate. The models with the highest model accuracy for the estimation of concrete and reinforcement are CO-RA4 and RE-RA5, with 96.53% and 92.56%, respectively. These models correspond to the models with the lowest AIC_c (see section "Compare Best RA and NN Models and Use the One With Lower AICc"). This indicates that MAPE would not yield the most accurate model when used as metric to evaluate different models.

Further investigation of the preferred models using MAPE showed that those models had more underestimated quantities when compared to the models using AIC_c; for example, for the CO models, the preferred model using MAPE underestimated 50% of the cases, while the preferred model using AIC_c underestimated 40% of the cases, indicating that MAPE would favor models that, on average, tend to underestimate. Similar findings were made by other researchers (Armstrong & Collopy, 1992; Foss et al., 2003; Makridakis, 1993).

Discussion

The proposed process can be used by practitioners to develop estimation models in a consistent and systematic way. All the key steps, from data collection and analysis to model evaluation, are considered. Special importance was given to the issue of sample size required for model development while allowing for data split to test or validate the developed models. However, a priori assumptions need to be made to determine the sample size (e.g., *k, R^2*, and *z*), which is then checked against the available data. The selection of those variables might not be intuitive at first, but with experience, practitioners would gain an improved understanding of those variables and, from the outcome of the developed models, use the ones that best fit their needs.

The selection metric used (AIC_c) is not as popular as other typically used metrics when comparing models or determining model performance (e.g., MAPE, RMSE); however, it has been successfully applied in other fields and shown to be more efficient than MAPE for the selection of models to estimate the CMQs in future projects. Nevertheless, the calculation of the AIC_c values are straightforward, and the information required to determine the AIC_c values from the developed models is readily available (i.e., most statistical software packages used to develop the model types used in the proposed methodology provide those values by default, so no extra effort is expected to determine the AIC_c values of the developed models). In addition, when comparing models using $AIC_{c,}$ the probability that one model is better than the others can be easily determined.

For the test cases, the developed models perform well and have high generalization capabilities. For the 12 random structures used for testing the models, the overall MAPEs from the regression models (Table 10; 5.469% (CO) and 9.216% (RE)) are slightly better than from the training dataset (Table 11; 3.662% (CO) and 7.881% (RE)). This can be attributed to the amount of data available for training and testing the models.

Conclusion

The proposed process allows for a systematic way to develop and evaluate models using RA and NN techniques, taking into consideration the available data and evaluating the developed models using an information criterion.

The process was illustrated by its use in developing models to estimate, during the conceptual phase of a project, the amount of concrete and reinforcement required in the construction of storage buildings, and in determining the most accurate model from those developed. The required sample size for model development in the example was calculated, and it was determined that the data available was sufficient to develop robust models. From the 58 existing structures, 46 were randomly selected and used for model development; the rest were used to validate the results from the process. All the potential variables, including the storage building type, were employed when using the backward elimination regression technique, which led to the development of four models for the estimation of concrete quantities and five models for the estimation of reinforcement quantities.

The results from the back-transformed regression models were evaluated using a second-order correction AIC (AIC_c). From this evaluation, the model with the lowest AIC_c was chosen, as it indicated that it was the model with the least amount of information lost (i.e., the most accurate model). Model CO-RA4 and RE-RA5 for concrete and reinforcement, respectively, were selected and used as basis for the development of feed-forward network with three layers (one hidden layer) and back-propagation supervised learning algorithm. The evaluation between the RA and NN models using AIC_c showed that the RA models performed better than the NN models, indicating that the selected functional form for the RA, in combination with the BET, was adequate. Similar findings have been made where models using RA showed better performance, in terms of accuracy, variability, model creation, and model examination, than NN models (Smith & Mason, 1997). In addition, the AIC_c metric proved to be effective for selecting the most accurate models from a set and more consistent than the MAPE.

Ongoing work related to this process and the estimation of CMQs includes the development of NN models using the same input sets from each RA model developed using the backward elimination regression technique (not just the selected one) and the implementation of the proposed process to develop a methodology using CBR for the estimation of CMQs during the conceptual phase of a project.

ORCID

Borja García de Soto http://orcid.org/0000-0002-9613-8105
Dilum Fernando http://orcid.org/0000-0001-7481-7935

Notes

1. Solver is part of a suite of commands, sometimes called what-if analysis tools, in MS Excel. Solver uses a variety of methods, from linear programming and nonlinear optimization to genetic and evolutionary algorithms, to find solutions.
2. For multiple regression, the denominator of Equation 9 would be n–k–1, where k is the number of IVs in the regression model.

References

Abdul-Wahab, S. A., Bakheit, C. S., & Al-Alawi, S. M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, *20*(10), 1263–1271.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Al-Alawi, S. M., Abdul-Wahab, S. A., & Bakheit, C. S. (2008). Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone. *Environmental Modelling & Software*, *23*(4), 396–403.

Albright, S. C., Winston, W. L., & Zappe, C. J. (2003). *Data analysis and decision making*. Pacific Grove, CA: Brooks/Cole, Thomson Learning, Inc.

Armstrong, J. S. (1985). *Long-range forecasting: from crystal ball to computer*. New York, NY: Wiley.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, *8*(1), 69–80.

Bakhoum, M., Morcous, G., Taha, M., & El-Said, M. (1998). Estimation of quantities and cost of prestressed concrete bridges over the Nile in Egypt. *Journal of Egyptian Society of Engineers/Civil*, *37*(4), 17–32.

Bartlett, J. E., Kotrlik, J. W., & Higgins, C. C. (2001). Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, *19*(1), 43–50

Baskerville, G. L. (1972). Use of logarithmic regression in the estimation of plant biomass. *Canadian Journal of Forest Research*, *2*(1), 49–53.

Bishop, C. M. (1994). Neural networks and their applications. *Review of Scientific Instruments*, *65*(6), 1803–1832.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.

Blair, A. N., Lye, L. M., & Campbell, W. J. (1993). Forecasting construction cost escalation. *Canadian Journal of Civil Engineering*, *20*(4), 602–612.

Bowen, P. A., & Edwards, P. J. (1985). Cost modelling and price forecasting: Practice and theory in perspective. *Construction Management and Economics*, *3*(3), 199–215.

Brooks, G. P., & Barcikowski, R. S. (2012). The PEAR method for sample sizes in multiple linear regression. *Multiple Linear Regression Viewpoints*, *38*(2), 1–16.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). MAPE model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

Castle, J. L., Qin, X., & Robert Reed, W. (2013). Using model selection algorithms to obtain reliable coefficient estimates. *Journal of Economic Surveys, 27*(2), 269–296.

Chan, S. L., & Park, M. (2005). Project cost estimation using principal component regression. *Construction Management and Economics*, *23*(3), 295–304.

Cho, H.-G., Kim, K.-G., Kim, J.-Y., & Kim, G.-H. (2013). A comparison of construction cost estimation using multiple regression analysis and neural network in elementary school projects. *Journal of the Korea Institute of Building Construction*, *13*(1), 66–74.

Chou, J. S., Peng, M., Persad, K. R., & O'Connor, J. T. (2006). Quantity-based approach to preliminary cost estimates for highway projects. *Transportation Research Record: Journal of the Transportation Research Board*, *1946*(1), 22–30.

Collopy, F., & Armstrong, J. S. (2000). *Another error measure for selection of the best forecasting method: The unbiased absolute percentage error*. Retrieved May 12, 2012, from http://www.forecastingprinciples.com/paperpdf/armstrong-unbiasedAPE.pdf

De Villiers, J., & Barnard, E. (1993). Backpropagation neural nets with one and two hidden layers. *Neural Networks, IEEE Transactions on*, *4*(1), 136–141.

Du, J., & Bormann, J. (2014). Improved similarity measure in case-based reasoning with global sensitivity analysis: An example of construction quantity estimating. *Journal of Computing in Civil Engineering, 28*(6), 04014020.

Dupont, W. D., & Plummer, W. D., Jr. (1998). Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials*, *19*(6), 589–601.

Field, A. (2009). *Discovering statistics using SPSS*. London, UK: Sage Publications Limited.

Flom, P. L., & Cassell, D. L. (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. Paper presented at NorthEast SAS Users Group Inc 20th Annual Conference: 11–14th November 2007, Baltimore, Maryland.

Foss, T., Stensrud, E., Kitchenham, B., & Myrtveit, I. (2003). A simulation study of the model evaluation criterion MMRE. *Software Engineering, IEEE Transactions on*, *29*(11), 985–995.

Fragkakis, N., Lambropoulos, S., & Tsiambaos, G. (2011). Parametric model for conceptual cost estimation of concrete bridge foundations. *Journal of Infrastructure Systems*, *17*(2), 66–74.

Gerrard, A. M., Brass, J., & Peel, D. (1994). Estimating vessel costs via neural networks. In *Proceedings of the 13th International Cost Engineering Congress, London, October 9–12*.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*(4), 405–408.

Gray, C. D., & Kinnear, P. R. (2012). *IBM SPSS statistics 19 made simple*. New York, NY: Psychology Press.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, *26*(3), 499–510.

Hegazy, T., & Ayed, A. (1998). Neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management*, *124*(3), 210–218.

Hegazy, T., Fazio, P., & Moselhi, O. (1994). Developing practical neural network applications using back-propagation. *Computer-Aided Civil and Infrastructure Engineering*, *9*(2), 145–159.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

International Society of Parametric Analysts. (2008). *Parametric estimating handbook* (4th ed.). Vienna, VA. Retrieved November 19, 2014, from http://www.galorath.com/images/uploads/ISPA_PEH_4th_ed_Final.pdf.

Jafarzadeh, R., Wilkinson, S., González, V., Ingham, J., & Amiri, G. (2014). Predicting seismic retrofit construction cost for buildings with framed structures using multilinear regression analysis. *Journal of Construction Engineering and Management*, *140*(3), 04013062.

Karlik, B., & Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of Neural Networks. *International Journal of Artificial Intelligence Expert Systems*, *1*(4), 111–122.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological methods*, *8*(3), 305.

Khosrowshahi, F., & Kaka, A. P. (1996). Estimation of project total cost and duration for housing projects in the UK. *Building and Environment*, *31*(4), 375–383.

Kim, G. H., An, S. H., & 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), 1235–1242.

Kim, G. H., Shin, J. M., Kim, S., & Shin, Y. (2013). Comparison of school building construction costs estimation methods using regression analysis, neural networks, and support vector machine. *Journal of Building Construction and Planning Research*, *1*, 1–7.

Kim, K. J., Kim, K., & Kang, C. S. (2009). Approximate cost estimating model for PSC Beam bridge based on quantity of standard work. *KSCE Journal of Civil Engineering*, *13*(6), 377–388.

Koo, C., Hong, T., Hyun, C., & 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), 739–752.

Leach, L. F., & Henson, R. K. (2003). The use and impact of adjusted R2 effects in published regression research. Paper presented at the Annual Meeting of the Southwest Educational Research Association, San Antonio, TX.

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of Construction Engineering and Management*, *132*(7), 750–758.

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, *9*, 527–529.

Makridakis, S., & Hibon, M. (1995). Evaluating accuracy (or error) measures (Working Paper No. 1995/18/TM). *The National Institute of Statistics and Applied Economics (INSEA)*. Retrieved January 12, 2013, from http://www.insead.edu/facultyresearch/research/details_papers.cfm?id=2265

Matlab. (2012). MATLAB 7.14 R2012a [computer software]. Natick, MA: The MathWorks Inc.

Mazerolle, M. J. (2004). Making sense out of Akaike's Information Criterion (MAPE): Its use and interpretation in model selection and inference from ecological data. (Doctoral dissertation). Université Laval, Faculté de Foresterie et de Géomatique, Quebec City, Québec, Canada. Retrieved October 16, 2013, from http://theses.ulaval.ca/archimede/fichiers/21842/apa.html

Mead, R. (1990). *The design of experiments: Statistical principles for practical applications*. Cambridge, UK: Cambridge University Press.

Motulsky, H. J., & Christopoulos, A. (2003). *Fitting models to biological data using linear and nonlinear regression: A practical guide to curve fitting*. San Diego, CA: GraphPad Software Inc.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

National Aeronautics and Space Administration. (2008). *2008 NASA Cost Estimating Handbook (CEH)*. Washington, DC. Retrieved November 19, 2014, from http://www.nasa.gov/pdf/263676main_2008-NASA-Cost-Handbook-FINAL_v6.pdf.

Newman, I. (1979). A Monte Carlo evaluation of estimated parameters of five shrinkage estimate formuli. *Multiple Linear Regression Viewpoints*, *9*(5), 57–74.

NS. (2012). NeuroSolutions (Version 6.2) [computer software]. Gainesville, FL: NeuroDimensions, Inc.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690.

Oh, C. D., Park, C., & Kim, K. J. (2013). An approximate cost estimation model based on standard quantities of steel box girder bridge substructure. *KSCE Journal of Civil Engineering*, *17*(5), 877–885.

Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": A randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, *154*(1), 135–150.

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling*, *178*(3), 389–397.

Owusu-Ababio, S. (1998). Effect of neural network topology on flexible pavement cracking prediction. *Computer-Aided Civil and Infrastructure Engineering*, *13*(5), 349–355.

Panchal, G., Ganatra, A., Kosta, Y. P., & Panchal, D. (2010). Searching most efficient neural network architecture using Akaike's Information Criterion (AIC). *International Journal of Computer Applications*, *1*(5), 41–44.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, *53*(5), 793–808.

Singh, S. (1990). Cost model for reinforced concrete beam and slab structures in buildings. *Journal of Construction Engineering and Management*, *116*(1), 54–67.

Singh, S. (1991). Cost estimation of prestressed concrete beam and reinforced concrete slab construction. *Construction Management and Economics*, *9*(2), 205–215.

Smith, A. E., & Mason, A. K. (1997). Cost estimation predictive modeling: Regression versus Neural Network. *The Engineering Economist*, *42*(2), 137–161.

Smith, R. J. (1993). Logarithmic transformation bias in allometry. *American Journal of Physical Anthropology*, *90*(2), 215–228.

Son, B. S., Lee, H. S., Park, M., Han, D. Y., & Ahn, J. (2013). Quantity based active schematic estimating (Q-BASE) model. *KSCE Journal of Civil Engineering*, *17*(1), 9–21.

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

Sprugel, D. G. (1983). Correcting for bias in log-transformed allometric equations. *Ecology*, *64*(1), 209–210.

SPSS. (2010). IBM SPSS Statistics for Windows (Version 19.0) [computer software]. Armonk, NY: IBM Corporation.

Triola, M. F. (2001). *Elementary statistics using Excel*. Boston, MA: Addison-Wesley Publishing Company.

Uhl, N., & Eisenberg, T. (1970). Predicting shrinkage in the multiple correlation coefficient. *Educational and Psychological Measurement*, *30*, 487–489.

United States Government Accountability Office. (2009). *GAO Cost Estimating and Assessment Guide (GAO-09-3SP)*. Washington, DC. Retrieved November 19, 2014, from http://www.gao.gov/assets/80/77175.pdf.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196.

Wood, A. G. (1986). A potential bias in log-transformed allometric equations. *Wader Study Group Bulletin*, *47*,17–19.

Yeh, I. C. (1998). Quantity estimating of building with logarithm-neuron networks. *Journal of Construction Engineering and Management*, *124*(5), 374–380.

Yin, P., & Fan, X. (2001). Estimating R 2 shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education*, *69*(2), 203–224.

Yu, W. D. (2006). PIREM: A new model for conceptual cost estimation. *Construction Management and Economics*, *24*(3), 259–270.

About the Authors

Borja García de Soto received the degree of Bachelor of Science in Civil Engineering from Florida International University (FIU) in 2000, the degree of Master of Science in Civil Engineering in the area of Structural Design from FIU in 2001 and the degree of Master of Science in Engineering in the area of Engineering and Project Management from the University of California at Berkeley (UC-Berkeley) in 2004. He is a registered Civil Engineer (licensed in California and Florida) with international experience in multiple aspects of project management, including risk management and control, delay analysis, forensic engineering, and project cost estimation. He worked as a Project Engineer in charge of the structural design of steel, wood, and reinforced concrete structures at FC Consulting Engineers, Inc. in Miami, Florida (2000-2002). In 2003 he was a Project Manager at Cyopsa-Sisocia, S.A. (Madrid) for the Northern Spain Region. After completing his MSc at UC-Berkeley in 2004, he joined JKA Inc., a construction consulting firm in the San Francisco Bay Area, California (2004-2010). As a Senior Consultant and licensed Professional Engineer at JKA he was responsible for a large variety of fast-paced projects covering multiple aspects of the construction industry. In July 2010 he founded the BGSL Consulting Group. Until July 2011 he led a large variety of fast-paced projects in the construction industry. In April 2011 he became part of the Infrastructure Management Group (IMG) in the Institute of Construction and Infrastructure Management (IBI) at the Swiss Federal Institute of Technology in Zürich (ETHZ).

Bryan T. Adey obtained a Bachelor's of Civil Engineering from Dalhousie University in 1995, a Master's of Science in Structural Engineering from the University of Alberta in 1997 and a Ph.D. in the area of Infrastructure Management from the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland in 2002. His Ph.D. focused on the integration of the consideration of natural hazards into bridge management systems. After completing his Ph.D. he was employed in the Division of Maintenance and Safety at the EPFL (2002-2003).

In 2003, Dr. Adey co-founded the consultancy Infrastructure Management Consultants Ltd. (IMC) and assumed the role as vice-president. From 2003-2009 he evaluated, developed and improved business models and business processes related to the maintenance and operation of infrastructure, developed and implemented infrastructure management systems, and analyzed infrastructure with respect to physical condition, risk and economics. This worked focused principally on road and rail infrastructure, but also included water distribution and water transportation networks. During this time he also continued research in the area of infrastructure management, in particular focusing on the methodologies to be used to estimate the total costs of infrastructure interventions, to evaluate the risk associated with road networks, and to assess risk reducing intervention strategies for road networks.

In 2010, Dr. Adey left IMC to become the head of the Infrastructure Management Group (IMG) in the Institute of Construction and Infrastructure Management (IBI) at the Swiss Federal Institute of Technology in Zürich (ETHZ). His vision for the IMG is to be a world leader in the provision of cutting edge frameworks, methodologies, models and tools to improve the management of infrastructure.

Dilum Fernando obtained a Bachelor's of Civil Engineering from Monash University in 2005. During the last semester of his bachelor studies, he joined Connell Wagner (pty) Ltd., as a part time design engineer. After completing his bachelor's degree in May 2005, he joined Connell Wagner (pty) Ltd. as a full time design Engineer and worked there until April 2006. After that Dr. Fernando started his PhD studies at The Hong Kong Polytechnic University in the area of Carbon Fibre Reinforced Polymer (CFRP) strengthening of metallic structures. His main research topics involved bond behavior between CFRP and steel, analytical and numerical modelling of debonding behavior of CFRP-steel bond joints and numerical modelling of the behavior of various CFRP strengthened metallic structures. After completing his PhD in 2010, he joined the Infrastructure Management Group (IMG) in the Institute of Construction and Infrastructure Management (IBI) at the Swiss Federal Institute of Technology in Zürich (ETHZ), as a Post-Doctoral fellow. The main research topics during his post-doctoral work at ETHZ involved, modelling the behavior of transportation infrastructure objects and networks under hazard situations, sustainable design of transportation infrastructure systems, which consider variations of design parameters, novel materials and techniques, and the development of administrative tools that facilitate the improved design and management of infrastructure assets. Dilum Fernando joined The University of Queensland in August 2013 as a lecturer in the School of Civil Engineering. His recent research topics include, structural rehabilitation, innovative applications of emerging materials in new structures, advanced numerical modeling, sustainable design and management of infrastructure assets.