Available online at www.sciencedirect.com

ScienceDirect

Journal of Taibah University for Science xxx (2016) xxx–xxx

Journal of Taibah University for Science

www.elsevier.com/locate/jtusci

Construction cost prediction model for conventional and sustainable college buildings in North America

Othman Subhi Alshamrani *

Department of Building Engineering, College of Architecture and Planning, University of Dammam, 31451 Al-Dammam, Saudi Arabia

Received 27 August 2015; received in revised form 24 January 2016; accepted 25 January 2016

Abstract

The literature lacks in initial cost prediction models for college buildings, especially comparing costs of sustainable and conventional buildings. A multi-regression model was developed for conceptual initial cost estimation of conventional and sustainable college buildings in North America. RS Means was used to estimate the national average of construction costs for 2014, which was subsequently utilized to develop the model. The model could predict the initial cost per square feet with two structure types made of steel and concrete. The other predictor variables were building area, number of floors and floor height. The model was developed in three major stages, such as preliminary diagnostics on data quality, model development and validation. The developed model was successfully tested and validated with real-time data.

Keywords: Initial cost; Sustainability; College building; Regression model; RS Means; Normal distribution; Residual analysis

1. Introduction

Practitioners and researchers have recognized the uncertainty of construction cost estimates and the need to improve the capability of cost prediction models [1]. Substantial efforts have been made to address this issue, and considerable conceptual cost prediction models are currently available in practice based on such techniques as probabilistic cost estimation, regression analysis, neural network (NN), fuzzy logic (FL), genetic algorithm (GA), and case-based reasoning (CBR). The relative merits and demerits of these techniques were analyzed by experts and are well-documented [2–4]. However, a review of updated literature related to the current study is presented in this report.

Regarding studies on regression analysis, Li et al. [5] proposed step-wise linear regression models for office buildings in Hong Kong, while a multivariate regression model named estimate score procedure was developed by Trost and Oberlender [6]. A similar study was reported by Lowe et al. [7], who developed linear regression models to predict the construction cost of buildings in the United Kingdom based on 286 sets of real data. Application of NN, FL and GA for construction cost prediction has attracted researchers and practitioners and literature in this area is abundant. Siqueira [8] applied NNs for cost estimation of low-rise prefabricated

* Tel.: +966 548008002; fax: +966 38578739.
E-mail address: osalshamrani@uod.edu.sa
Peer review under responsibility of Taibah University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jtusci.2016.01.004

structural steel buildings in Canada. The data were collected from 75 building projects over a 3-month period. A similar study was reported from Turkey [9], which used data from 30 projects to train and test the NN model developed for cost prediction of reinforced concrete structural systems of four- to eight-story residential buildings. Kim et al. [10] incorporated GA in their backpropagation network (BPN) model to improve the accuracy of construction cost estimation. Data for 530 residential buildings constructed in Korea between 1997 and 2000 were used for training and evaluation of the model. The web-based intelligent cost estimator (WICE) model developed by Lai and Lee [11] included features of WWW, neuro-fuzzy system, and data mining. The proposed model was claimed to provide not only a globally accessible and promptly responding means for cost estimation but also an effective and reliable tool for real-time decision-making. A subsequent study [12] proposed the evolutionary fuzzy neural inference model (EFNIM), which incorporated features of GA, FL and NNs. The EFNIM was subsequently combined with WWW and historical data to form evolutionary web-based conceptual cost estimators (EWCCE), which provided two types of estimators for conceptual construction cost. The artificial neural network (ANN)-based evolutionary fuzzy hybrid neural network (EFHNN) was developed by Cheng et al. [13], which was claimed to be effective for accurate cost estimation during the early stages of construction projects. Other recent NN-based models included those reported by Juszczyk [14], Bala et al. [1] and Aibinu et al. [15].

In the CBR model, new problems are solved by recognizing the similarity to a known problem and adapting solutions that were used to resolve the previous problems [16]. Many studies on model development based on CBR were reported. For instance, An et al. [16] proposed a CBR model using the analytic hierarchy process (AHP), which included experience in all processes of cost estimation. While few similar models were developed by Koo et al. [17], Hong et al. [18] and Park and Lee [19], an advanced CBR model with 101 cases of multi-family housing projects was reported by Koo et al. [20] integrating the advantages of CBR, multiple regression analysis (MRA), ANN, and the optimization process using GA. The model was developed by the Microsoft Excel-based visual basic application (VBA), which is user-friendly [20]. In a recent study, the CBR model was integrated with AHP for cost estimation of highway projects [21].

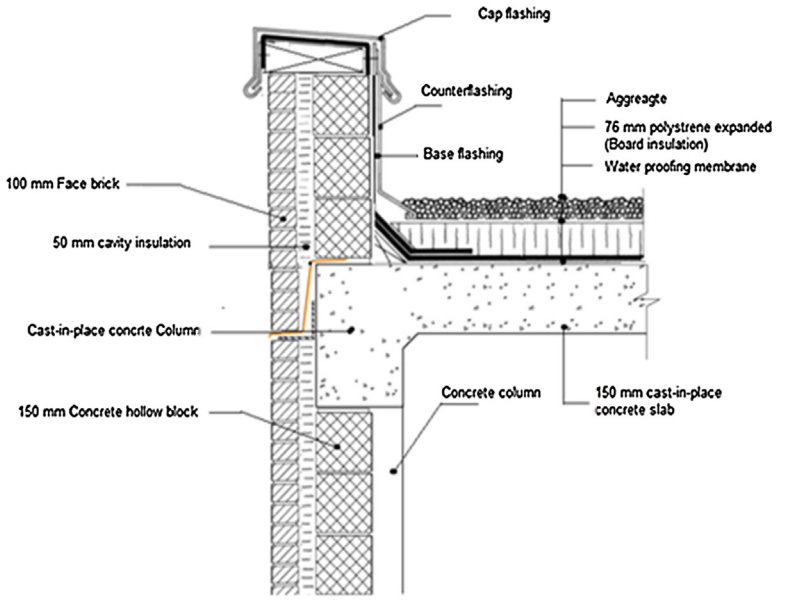
Attempts were also reported on developing new and hybrid prediction models. The online analytical processing (OLAP) environment introduced by Moon et al. [22], the principal item ratios estimating method (PIREM) proposed by Yu [23], and the bootstrap approach presented by Sonmez [24], are good examples for new approaches. Although green buildings are designed to reduce negative environmental impacts with enhanced functionality, initial cost is a matter of concern for the owners [25]. This issue was addressed by Ahn [25], who developed a multiple regression model to establish the relationship between initial cost and saving using life cycle cost (LCC) for implementing green building strategies into capital projects in the United States. However, as far as the author is aware, there are few published studies concerning the development of initial cost prediction models for college buildings, particularly in provision of cost comparison of sustainable and conventional buildings and selection of economically viable structure options. This paper presents a regression model for initial cost prediction for conventional and sustainable college buildings in North America. The initial cost is estimated for two structural alternatives (steel and concrete) with specific area, floor height and number of floors. RS Means was used to estimate the national average of construction costs for the year 2014.

2. Methodology

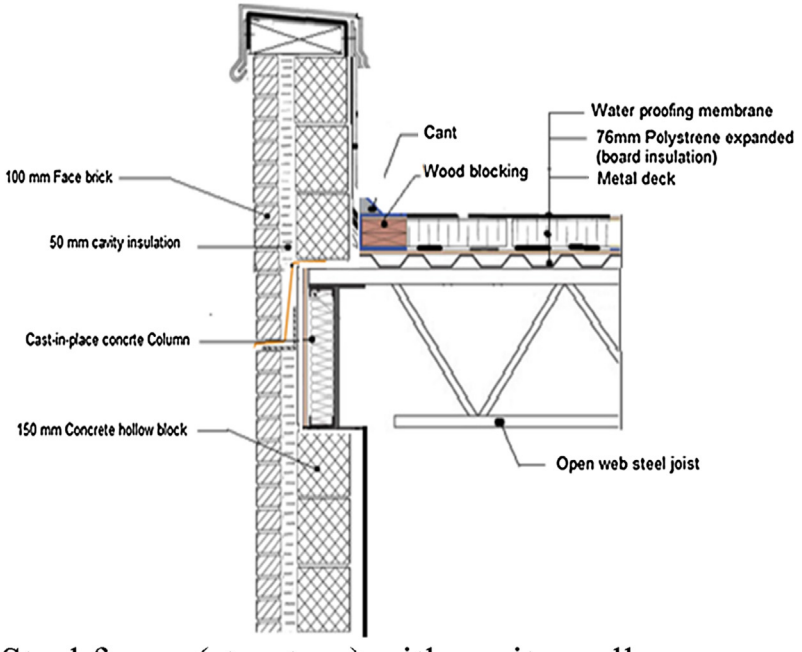
RS Means was used to estimate the construction costs of new green mid-rise college buildings by identifying significant parameters. Several input parameters were defined for initial cost calculation, such as building area (50,000–350,000 ft^2), floor height (12–18 ft), number of floors (1–3), and structure type (steel and concrete, as detailed in Fig. 1). All of the costs were estimated based on cost index of the third quarter of year 2014 and national average costs for North American cities. The breakdown cost components accounted for were substructure, super-structure, services, interiors, equipment and furnishings, and contractor and architecture fees (Fig. 2). Estimated breakdown component costs are used for calculation of subtotal costs which, together with contractor and architecture fees, ultimately determine total building costs.

The initial costs were computed by applying 80 different scenarios (Fig. 3) to build a correlation between input parameters and total base cost per square-feet. Each structure-envelope alternative was estimated for green mid-rise college buildings with specific area, floor height and number of floors. A total of 320 construction cost estimation scenarios were obtained from the combination of the complete range of input parameters for college buildings. The input parameters (independent variables) for initial costs were gathered from RS Means.

Please cite this article in press as: O.S. Alshamrani. Construction cost prediction model for conventional and sustainable college buildings in North America, J. Taibah Univ. Sci. (2016), http://dx.doi.org/10.1016/j.jtusci.2016.01.004



a. Concrete bearing wall (structure) with cavity wall



b. Steel frame (structure) with cavity wall

Fig. 1. Detailed sections for various tested alternatives.

Several of these variables were normalized, such as location and year of construction. Other parameters, such as structure-envelope type, floor height, number of floor, and building area, were variables significantly affecting initial costs. These factors were investigated to develop their correlation with the initial cost results (dependent factor). Computed initial costs from RS Means included 320 data points, out of which 250 points were used for model development and 70 points were used for model validation.

Please cite this article in press as: O.S. Alshamrani. Construction cost prediction model for conventional and sustainable college buildings in North America, J. Taibah Univ. Sci. (2016), http://dx.doi.org/10.1016/j.jtusci.2016.01.004

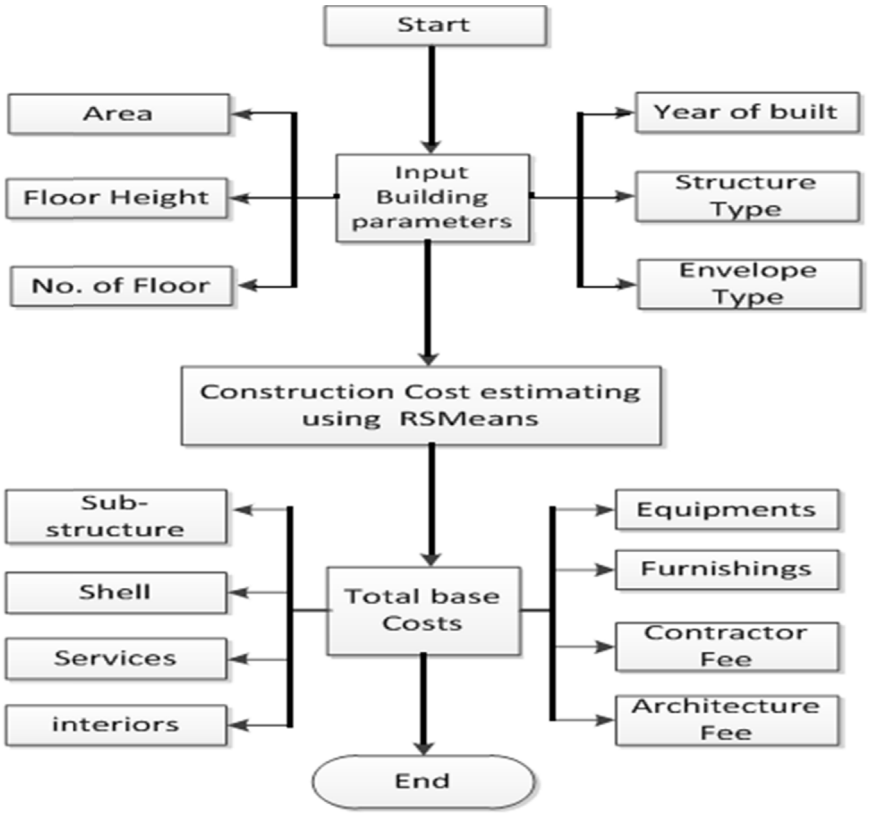


Fig. 2. Initial cost estimation process using RS Means.

2.1. Model development process

The regression model development methodology is illustrated in Fig. 4, which consists of three major stages such as preliminary diagnostics on data quality, the model development process and model validation; each step in these stages are described as follows:

2.1.1. Preliminary data diagnostics

The first step for preliminary data diagnostics was to detect and address any existing multicollinearities or possible interactions of predictor variables of the developed models. The matrix scatter plot for all predictor variables was simulated against the response factor to detect a correlation. Scatter plot representation was significant in detecting the linearity of data or any other correlation between predictors and response variables, as well as among predictor variables themselves. The next step was to perform best subset regression analysis. This test identifies the best possible combination of predictors with regards to the highest R^2 and R^2 (adjusted) values and the lowest error and variation

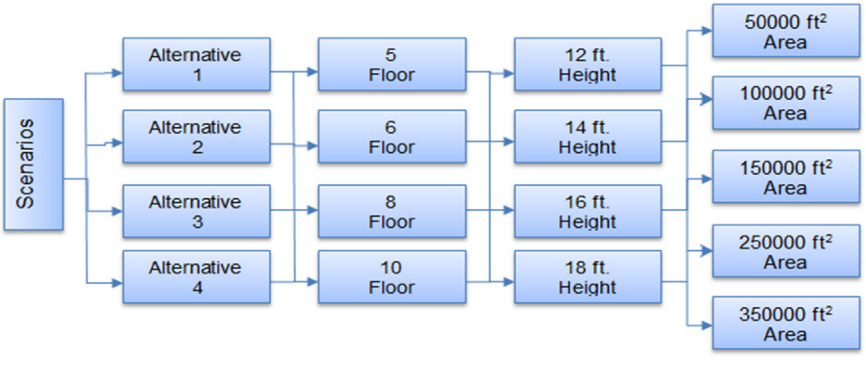


Fig. 3. Various scenarios tested for initial costs.

Please cite this article in press as: O.S. Alshamrani. Construction cost prediction model for conventional and sustainable college buildings in North America, J. Taibah Univ. Sci. (2016), http://dx.doi.org/10.1016/j.jtusci.2016.01.004

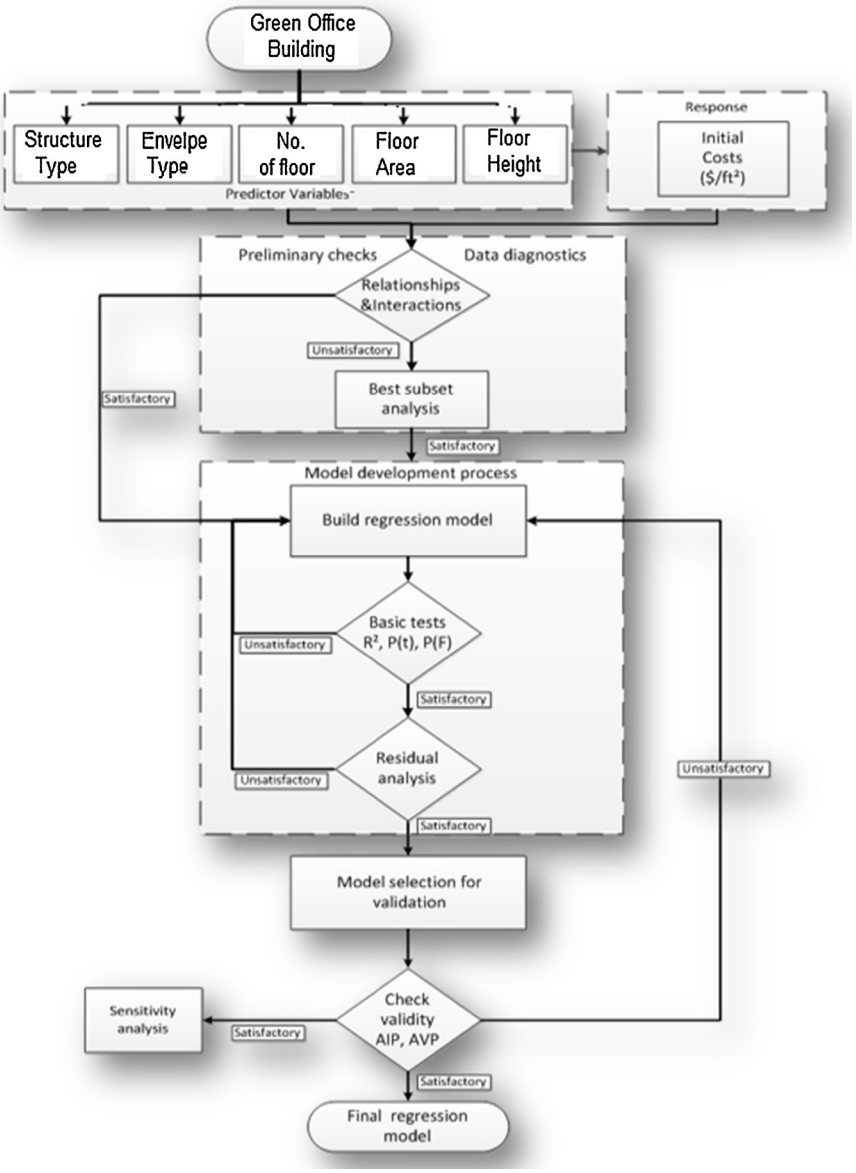


Fig. 4. Regression model development process.

values. Hence, the best-fit regression model that can be developed with the specified number of variables is determined. After detecting the correlation and identifying the best data subset, the regression model was developed using the best data set and RS Means. The computed data were stored in Microsoft Excel. The Minitab® statistical software package was employed for regression model development.

2.1.2. Model adequacy test

Preliminary tests of the regression model included a test of the coefficient of multiple determinations (R^2), regression test (F), and a test for each regression parameter coefficient "β_k". The second test was the regression relation test (F). To determine $p(F)$ for the whole model, a hypothesis test was applied. The assumption of the null hypothesis (H_0) was $\beta_0 = \beta_1 = \beta_{p-1} = 0$. The assumption for the alternative hypothesis (H_a) was that not all coefficients were equal to zero. If the p-value (statistical significance) is 0.00, the null hypothesis is rejected and the estimated model is significant at $\alpha = 0.05$ indicating that at least one coefficient in the developed regression model is not equal to zero. The third test was the t-test, which determines the validity of the regression

Please cite this article in press as: O.S. Alshamrani. Construction cost prediction model for conventional and sustainable college buildings in North America, J. Taibah Univ. Sci. (2016), http://dx.doi.org/10.1016/j.jtusci.2016.01.004

coefficient and is performed separately for β_0, β_1, …, β_{p-1}. In the case of β_0, the null hypothesis (H_0) of the *t*-test assumed that $\beta_0 = 0$, while the alternative hypothesis (H_a) assumed that $\beta_0 \neq 0$. After coefficients and bases were satisfactorily diagnosed, the residuals and their patterns were analyzed. The normality of error was checked to verify the linearity of correlation assumptions. Normal probability and frequency were represented in a graph of the developed model in order to perform residual analysis.

2.1.3. Model validation

The first step in model validation was to compare the actual observation with predicted values for the validation data for each developed model. This validation was performed using the excluded 70 data points, which were plotted to compare the prediction model with the observed data in hand. The mathematical validation method was performed using average validity and invalidity percentages. Average invalidity and validity percentage was computed using the following equations [26],

$$AIP = \frac{\sum_{i=1}^{n} \left| 1 - \left(\frac{E_i}{C_i} \right) \right|}{n}$$

and

$$AVP = 1 - AIP$$

where *AVP* is the average validity percentage, *AIP* is the average invalidity percentage, E_i is the predicted value, C_i is the actual value, and n is the number of observations. The AIP value varies from 0 to 1.

3. Results and discussion

3.1. The developed regression model

The regression model developed in the present study for predicting the initial cost of sustainable college buildings in North American cities is

$$IC = 171.3 + 0.666 * H + 4.498 * n_f - 0.000129 * A + 6.292 * S + 5.003 * Str$$

where *IC* – predicted initial cost in USD/ft^2, *H* – height of one floor (12–18 ft), *A* – area of building in ft^2, n_f – number of floors (1–3), *S* – sustainability index (1 for conventional and 2 for sustainable), Str – structure type (1 for concrete bearing wall steel and 2 for steel frame).

Table 1
Statistical diagnostic of the model.

Predictor	Coefficient	SE coef.	*T*	*P*	VIF
Constant	171.3	1.84	92.91	0	
Height (ft)	0.666	0.0989	6.73	0	1
No. of floor	4.498	0.272	16.52	0	1.02
Area (ft^2)	−0.000129	0.000004	−29.59	0	1.02
Sustainability	6.292	0.441	14.26	0	1.01
Structure	5.003	0.44	11.37	0	1

3.2. Correlation tests

Correlation tests were conducted to test the linearity of the data by detecting a possible correlation of obtained scatter plot matrix and correlation matrix with the transformed Y' variable. The results are presented in Fig. 5, which shows that the data are constant and distributed evenly across the graph without forming any pattern. The plots also indicate that each of the predictor variables is nearly linearly associated with the response variable. Hence, the plots are considered satisfactory.

3.3. Best subset analysis

The output of subset regression analysis generates various regression models in each line, as shown in Fig. 6. In the WC regression model, the highest values of R^2 and R^2 (adjusted) are recorded at 87.6% and 87.3%, respectively, while the lowest values of C_p and standard deviation (S) are recorded at 6.0 and 3.1396, respectively. The result of the best subset analysis proves that all predictors are significant and should be combined and included in the developed regression model. This combination of variables is proven to be the best case for the developed regression model.

3.4. Results of t-test and F-test

As summarized in Table 1, the *t*-test shows that the *p*-values for the estimated coefficients of all predictors are 0.000. Therefore, the null hypothesis is rejected and the alternative hypothesis is accepted. This indicates that the predictors are significantly correlated with the response variable, i.e., 'initial cost' at $\alpha = 0.1$. The *p*-value (statistical significance) in the analysis of variance is 0.000 as shown in Table 2. The null hypothesis is thus rejected. This finding demonstrates that the estimated model is significant at $\alpha = 0.05$. Consequently, at least one coefficient in the developed regression model is not equal to zero.

Please cite this article in press as: O.S. Alshamrani. Construction cost prediction model for conventional and sustainable college buildings in North America, J. Taibah Univ. Sci. (2016), http://dx.doi.org/10.1016/j.jtusci.2016.01.004

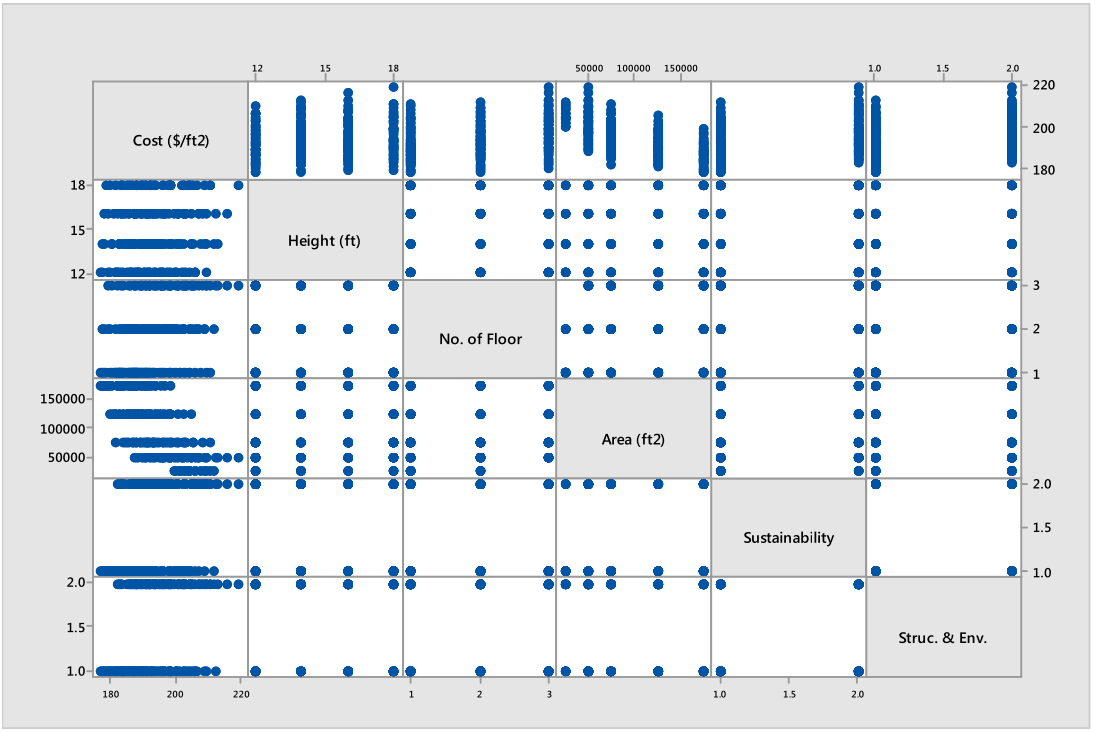


Fig. 5. Scatter matrix plot for regression model parameters.

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	Height (floor)	No. of Floor	Area (ft2)	Sustainability	Struc. & Env.
1	45.9	45.6	44.9	664.3	6.4942		X			
1	11.2	10.8	9.5	1217.8	8.3179			X		
2	63.2	62.8	62.0	390.1	5.3702	X	X			
2	60.1	59.7	59.0	439.0	5.5892		X	X		
3	76.6	76.3	75.6	177.5	4.2906	X	X	X		
3	71.8	71.4	70.6	254.4	4.7115	X	X			X
4	84.8	84.5	84.0	49.3	3.4717	X	X	X	X	
4	79.5	79.1	78.4	133.3	4.0265	X	X	X	X	
5	87.6	87.3	86.8	6.0	3.1396	X	X	X	X	X

Fig. 6. Best subset analysis result using Minitab.

3.5. Result of residual analysis

Figs. 7 and 8 represent the normal distribution plot and the histogram, respectively, from residual analysis. The normal probability plot (Fig. 7) indicates that the error terms are approximately normally distributed. The minor departures from normality observed are considered to be unusual possible outliers. Errors around a regression line

Please cite this article in press as: O.S. Alshamrani. Construction cost prediction model for conventional and sustainable college buildings in North America, J. Taibah Univ. Sci. (2016), http://dx.doi.org/10.1016/j.jtusci.2016.01.004

Table 2
Analysis of variance of the model.

Source	DF	SS	MS	*F*	*P*
Regression	5	13,795.0	2758.99	279.9	0
Height (ft)	1	446.7	446.74	45.32	0
No. of floor	1	2691.0	2690.96	273	0
Area (ft^2)	1	8629.5	8629.48	875.46	0
Sustainability	1	2005.3	2005.32	203.44	0
Struc. & env.	1	1274.6	1274.59	129.31	0

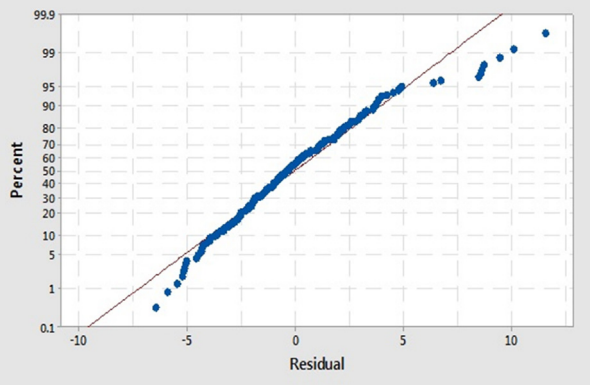


Fig. 7. Normal probability plot from residual analysis.

should be independent for each predicted value. Fig. 9 shows the residuals versus the order of data plot, which illustrates that the positive residuals are observed at inner bands of X values, and the outer bands largely consist of negative residuals. The R^2 values and other statistical parameters could be improved by eliminating these outliers. However, the model would not be the best representation of the real world data in hand. This result is satisfactory because minor departures from normality are acceptable [27].

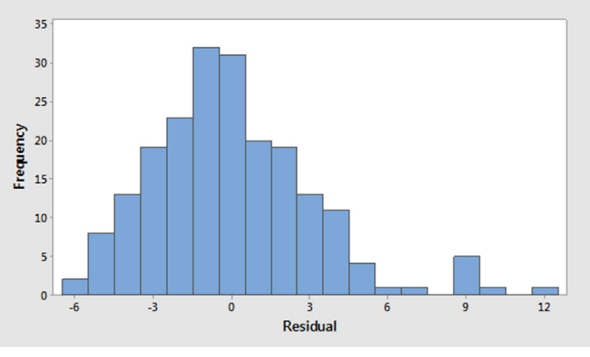


Fig. 8. Histogram from residual analysis.

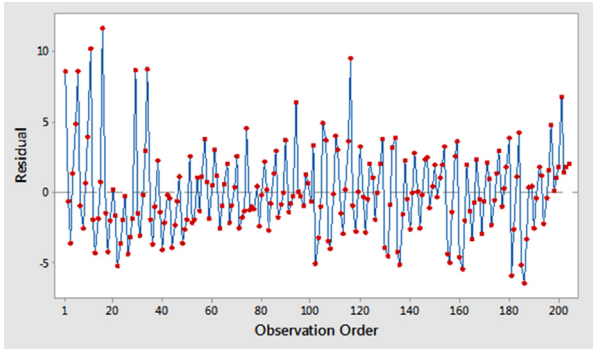


Fig. 9. Residual versus observation order plot.

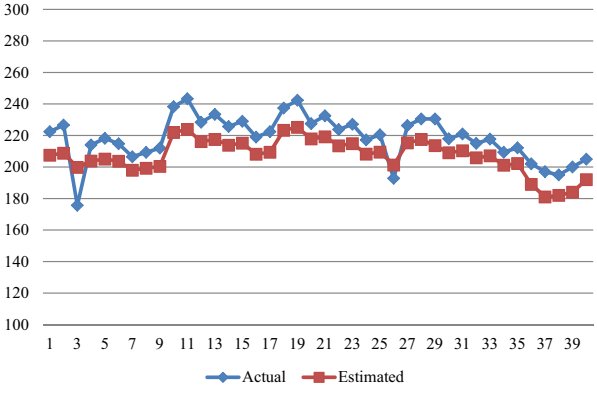


Fig. 10. Plot validation for the developed regression model.

3.6. Model validation

The developed regression model is validated by plot validation and mathematical validation. Fig. 10 presents the plot validation comparing the actual observation and predicted output. The predicted values are scattered around the actual values for the response variable. Hence, the result is considered satisfactory. The mathematical validation as presented below indicates that the predicted model is almost 94.3% accurate.

$$AIP = \frac{\sum_{i=1}^{n}\left|1-\left(\frac{E_i}{C_i}\right)\right|}{n} = AIP = \frac{1.98}{35} = 0.056,$$
$$AVP = 1 - AIP = 0.943$$

4. Conclusions

A user-friendly regression model has been developed to predict the initial cost of conventional, as well as sustainable, college buildings with a maximum of three floors in North America. RS Means was used to estimate

Please cite this article in press as: O.S. Alshamrani. Construction cost prediction model for conventional and sustainable college buildings in North America, J. Taibah Univ. Sci. (2016), http://dx.doi.org/10.1016/j.jtusci.2016.01.004

the national average of construction costs for the year 2014. The input parameters considered were building area, floor height, number of floors, and structure and envelope types. The model was validated by comparing the predictions with real data, as well as by using mathematical equations. The model validation demonstrated acceptable discrepancies, while the accuracy obtained by mathematical validation was 94.3%. The contributory aspect of this model compared to previous models is that besides predicting the initial cost of mid-rise college buildings, it enables universities to assess the economic viability of sustainable buildings over conventional buildings and compare the costs of concrete and steel structures.

References

[1] K. Bala, S.A. Mustani, B.S. Waziri, A computer-based cost prediction model for institutional building projects in Nigeria, J. Eng. Des. Technol. 12 (4) (2014) 518–529.
[2] R. Sonmez, Review of conceptual cost modeling techniques, AACE Int. Trans. (2005) EST.07.1–EST.07.4.
[3] G.-H. Kim, S.-H. An, K.-I. Kang, Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning, Build. Environ. 39 (10) (2004) 1235–1242.
[4] R. Sonmez, Conceptual cost estimation of building projects with regression analysis and neural networks, Can. J. Civil Eng. 31 (4) (2004) 677–683.
[5] H. Li, Q.P. Shen, P.E.D. Love, Cost modelling of office buildings in Hong Kong: an exploratory study, Facilities 23 (9/10) (2005) 438–452.
[6] S.M. Trost, G.D. Oberlender, Predicting accuracy of early cost estimates using factor analysis and multivariate regression, J. Constr. Eng. Manag. 129 (2) (2003) 198–204.
[7] D.J. Lowe, M.W. Emsley, A. Harding, Predicting construction cost using multiple regression techniques, J. Constr. Eng. Manag. 132 (July (7)) (2006) 750–758.
[8] I. Siqueira, Neural Network-based Cost Estimating (Master thesis), Concordia University, 1999.
[9] H.M. Günaydın, S.Z. Doğan, A neural network approach for early cost estimation of structural systems of buildings, Int. J. Proj. Manag. 22 (7) (2004) 595–602.
[10] G.-H. Kim, J.-E. Yoon, S.-H. An, H.-H. Cho, K.-I. Kang, Neural network model incorporating a genetic algorithm in estimating construction costs, Build. Environ. 39 (11) (2004) 1333–1340.
[11] W. Yu, C. Lai, W. Lee, A WICE approach to real-time construction cost estimation, Autom. Constr. 15 (1) (2006) 12–19.
[12] M.Y. Cheng, H.C. Tsai, W.S. Hsieh, Web-based conceptual cost estimates for construction projects using evolutionary fuzzy neural inference model, Autom. Constr. 18 (2) (2009) 164–172.
[13] M.-Y. Cheng, H.-C. Tsai, E. Sudjono, Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry, Expert Syst. Appl. 37 (6) (2010) 4224–4231.
[14] M. Juszczyk, The use of artificial neural networks for residential buildings conceptual cost estimation, in: AIP Conference Proceedings 1558, vol. 1302, no. May, 2013, pp. 1302–1306.
[15] A.A. Aibinu, D. Dassanayake, T. Chan, R. Thangaraj, Cost estimation for electric light and power elements during building design – a neural network approach, Eng. Constr. Archit. Manag. 22 (2) (2015) 190–213.
[16] S.-H. An, G.-H. Kim, K.-I. Kang, A case-based reasoning cost estimating model using experience by analytic hierarchy process, Build. Environ. 42 (7) (2007) 2573–2579.
[17] C. Koo, T. Hong, C. Hyun, S.H. Park, J. Seo, A study on the development of a cost model based on the owner's decision making at the early stages of a construction project, Int. J. Strateg. Prop. Manag. 14 (2) (2010) 121–137.
[18] T. Hong, C. Hyun, H. Moon, CBR-based cost prediction model-II of the design phase for multi-family housing projects, Expert Syst. Appl. 38 (3) (2011) 2797–2808.
[19] S.-H. Ji, M. Park, H.-S. Lee, Cost estimation model for building projects using case-based reasoning, Can. J. Civil Eng. 38 (May (5)) (2011) 570–581.
[20] C. Koo, T. Hong, C. Hyun, The development of a construction cost prediction model with improved prediction capacity using the advanced CBR approach, Expert Syst. Appl. 38 (7) (2011) 8597–8606.
[21] S. Kim, Hybrid forecasting system based on case-based reasoning and analytic hierarchy process for cost estimation, J. Civil Eng. Manag. 19 (1) (2013) 86–96.
[22] S.W. Moon, J.S. Kim, K.N. Kwon, Effectiveness of OLAP-based cost data management in construction cost estimate, Autom. Constr. 16 (3) (2007) 336–344.
[23] W. Yu, PIREM: a new model for conceptual cost estimation, Constr. Manag. Econ. 24 (3) (2006) 259–270.
[24] R. Sonmez, Parametric range estimating of building costs using regression models and bootstrap, J. Constr. Eng. Manag. 134 (December (12)) (2008) 1011–1016.
[25] Y.H. Ahn, The Development of Models to Identify Relationships Between First Costs of Green Building Strategies and Technologies and Life Cycle Cost for Public Green Facilities (PhD thesis), Virginia Polytechnic Institute and State University, 2010.
[26] T.M. Zayed, D.W. Halpin, Productivity and cost regression models for pile construction, J. Constr. Eng. Manag. 131 (7) (2005) 779–789.
[27] M. Kutner, C. Nachtsheim, J. Neter, W. Li, Applied Linear Statistical Models, 5th ed., McGraw-Hill/Irwin, 2004.

Please cite this article in press as: O.S. Alshamrani. Construction cost prediction model for conventional and sustainable college buildings in North America, J. Taibah Univ. Sci. (2016), http://dx.doi.org/10.1016/j.jtusci.2016.01.004