DATA PREPROCESSING METHOD FOR COST ESTIMATION OF BUILDING PROJECTS

Sae-Hyun Ji
Ph.D. Candidate, Dept. of Architecture, Seoul National Univ., Seoul, Korea
oldclock@snu.ac.kr

Moonseo Park
Professor, Ph.D., Dept. of Architecture, Seoul National Univ., Seoul, Korea
mspark@snu.ac.kr

Hyun-Soo Lee
Professor, Ph.D., Dept. of Architecture, Seoul National Univ., Seoul, Korea
hunslee@snu.ac.kr

You-Sang Yoon
Researcher, Ph.D., Lean construction centre, Seoul National Univ., Seoul, Korea
nxt0824@snu.ac.kr

Abstract

All construction projects have unique characteristics that must be considered during cost estimating and checking activities. Especially in the conceptual and schematic design stages, owners require precise cost estimates than their information providing. However, information related to the project scope is more likely to change in the early design phases in response to ongoing scope changes. Also, as only minimal project scope information is available during the early stages of estimation, cost estimators require effective estimation strategies. In practice, parametric cost estimation—which utilizes historical cost data—is the most commonly used method in these initial phases. Therefore, compilation of historical data pertaining to appropriate cost variance governing parameters is a prime requirement. However, data mining (data preprocessing) for denoising internal errors or abnormal values must be performed before this compilation. To address these issues, this research utilizes applied statistical methods, and it develops an alternative cost model based on these. To construct the model, the building cost data of 124 apartment projects, which are selected for case study, in Korea are compiled.. The main objectives of the suggested model are to effectively prepare strategic and conceptual cost estimates, and to provide check and control functions during the conceptual and schematic design stages. The cost model is expected to provide more accurate and stable estimates than the conventional methods.

KEYWORDS: database, cost, estimate, preprocessing

INTRODUCTION

Generally, cost estimates are based on the estimator's experience, imaginative abilities, and a wide range of assumptions including appraisals of previously conducted projects that are similar in scope (Jarde & Alkass 2007). Practically, parametric cost estimates, developed by adopting regression analysis (Hegazy and Ayed 1998), are most likely to be carried out in the initial phases. One of the most common parametric estimation methods is the unit cost method (e.g., cost per square foot, cost per bed for a hospital), which utilizes either historical building cost data or cost books to obtain an estimate of a building's cost per square foot (Karshenas 1984, Kirkham 2007). Recently, Soutos and Lowe (2005) developed a parametric cost model adopting multiple regression equation based on buildings cost data and identified cost significant variables. In this way, the parametric method is based on historical data collected from similar past projects and scope reflecting parameters (Hendricson 2000). However, the intricate interactions among cost variance impact factors make negative influence on estimate accuracy and employment (Garza and Rouhnan 1995). As demonstrated by previous researches, the compilation of historical data with the cost variance impact parameters is a prime requirement for the preparation of liable and accurate cost estimating. However, despite their arguments on this issue, few approaches were made toward application of data preprocessing on the cost estimate of construction projects. Data preprocessing is a precedent practice of data mining for denoising internal errors or abnormal values. Data mining is the process of extracting useful information from database and looking for useful patterns that can aid decision making (Han & Kamber 2003). As an effort to deal with this issue, this research develops a statistical methodology that determines cost governing factors and performs data preprocessing. Thereafter, based on preprocessed data, parametric cost estimate equation is developed using multiple regression analysis. To develop this Statistically Preprocessed data Based Parametric (SPBP) cost model, the building cost data of 124 apartment projects in Korea are compiled. The main objectives of the SPBP model are to effectively prepare strategic and conceptual cost estimates, and to provide a check and control functions during the conceptual and schematic design stages. The SPBP cost model is expected to enhance cost estimation field by providing more accurate and stable result than conventional ones.

PREVIOUS RESEARCHES

Cost modeling is defined as a symbolic representation of a system in which the system's content is expressed in terms of the factors that influence the system's cost (Ferry and Brandon 2007). The fundamental objective of cost modeling is to simulate current or future situations in such a way that the solutions posed by the simulation will generate results that may be subsequently analyzed and utilized in the decision-making processes of design (Ashworth 1986). It is generally assumed that cost models can assist consultants and contractors in providing their clients with more reliable cost advice. Teicholz (1993) argued that a good forecasting method should not require input data that are expensive or difficult to collect. Instead, it should be simple, accurate, unbiased, timely, and stable enough to be easily integrated into the cost system. Ellsworth (1998), on the other hand, argued that the simplest method for determining a reasonable estimate of facility cost is to identify and compare the cost of one project to the cost of similar projects. Furthermore, Ellsworth's research indicates that parametric estimating methods—predominantly used by owners at the conceptual stage of construction—can be an effective cost estimating strategy in which the characteristics of similar projects are used as parameters. Traditionally, parametric cost estimating models are developed by applying regression analysis to historical project information. Yet, it is difficult to develop these models due to the inherent limitations of regression analysis (Hegazy and Ayed 1998). One major disadvantage of regression based techniques is that they require a defined mathematical form for the cost function that best fits to available historical data

(Creese and Li 1995). Another disadvantage is their unsuitability to account for a large number of variables involved in construction projects and knotty intricate interactions among these variables. These limitations contribute to the low accuracy of traditional models, and to their limited use in construction (Garza and Rouhnan 1995). However, previous research rarely examined data preprocessing as a viable component of cost estimating. In the literature, the cost estimating formula has been deduced using regression analysis based on survey results that rank parameters on nominal and ordinal scales (Chan & Park 2005), despite the fact that these parameters consist of real technical and qualitative data. Cheng et al. (2008) have developed a web-based conceptual cost estimate system by introducing techniques such as genetic algorithms, fuzzy logic, and neural networks. Nevertheless, as their system uses many impact factors and requires a complicated process, it is likely that the system will be regarded as a “black box.” Soutos and Lowe (2005), on the other hand, developed a parametric cost model using regression analysis based on a database of 360 buildings. However, their research did not deal with data preprocessing and did not address the presence of multicollinearity. Furthermore, Trost and Oberlender (2003), using multivariate regression analysis based on the results of factor analysis, identified five factors that exhibit a significant impact on estimate accuracy. Their research indicates that the drivers of estimate accuracy must be adequately identified and quantified. And alos, Hegazy and Ayed (1998) developed a parametric cost estimating model to address the lack of project scope information. Their model adopts the neural networks approach and consists of one or more functions, or cost estimating relationships, between cost (as a dependent variable) and the cost determining factors (as independent variables). This approach is appropriate because it can address difficult tasks that require intuitive judgment, and it can detect data patterns that elude conventional analytical techniques.

COMPONET METHODOLOGIES

Principle Component Analysis (PCA)

Principle Component Analysis (PCA) is the simplest eigenvector-based multivariate analyses. Its operation is often thought of as revealing the internal structure of data in a way that best explains variance in that data (Kwan 2008). PCA is a technique for forming new variables which are linear composites of the original variables. The new variables are called principle components and the values of the new variables are called principle components scores. The first principle component accounts for the maximum variance in the data. PCA, which supplies a lower-dimensional picture, is commonly referred as a data reduction technique without losing characteristics (Sharma 1996). If there might be high correlation among variables, or one to two high variance variables in the data set, most of the information of original data could be represented by the first components. Thus, the contribution of variables can be prioritized with principle component scores (Kwan 2008).

Multiple regression analysis

Regression analysis is used when the dependent variable and the multiple independent variables of the model are measured using a metric scale resulting in metric data. Regression analysis is generally used for prediction, hypothesis testing, and modeling of causal relationships. As multiple regression analysis contains many independent variables, it is more comprehensive, provides more accurate estimation, and also reduces estimate error. However, using multiple regression analysis usually involves the assumption that multicollinearity and interaction effect do not exist. More commonly, the issue of multicollinearity arises when multiple regression analysis is used which have a high degree of correlation (either positive or negative) between two or more independent variables. In the presence of multicollinearity in the data, stepwise approach may be appropriate for excluding the effect.

SPBP COST MODELING

To address the aforementioned problems, the Statistically Preprocessed Data Based Parametric (SPBP) cost model is developed in this research. First, the scope of the cost model is defined. Then, historical data are collected, and within the previously defined cost model scope, the collected data are normalized for time by using the historical cost index. Then, statistical preprocessing of data cleansing for denoising is conducted based on interval estimation sampling. Consequently, the statistically preprocessed data set is abstracted under the unit price, which is adjusted by the first cost variance governing (the dominant) parameter. Using the cleansed data, cost estimate relationships are then derived through a stepwise multiple regression analysis for excluding the effects of multicollinearity. Finally, the SPBP model is validated through a comparative study with other cases. The entire SPBP cost modeling process is diagramed in Fig. 1.

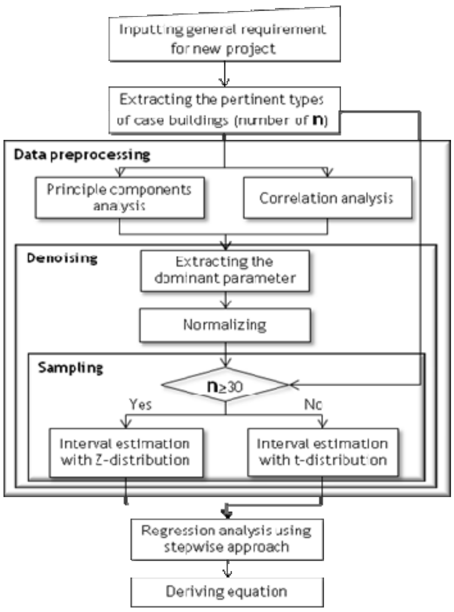

Figure 1: The SPBP Cost Modeling

Model Scoping

Cost estimation is crucial to construction contract tendering, providing a basis for establishing the likely cost of resource of the tender price for construction work (Akintoye 2000). The recent researches focus on estimates generated during the initial stages of a project. As such, the primary function of initial estimating is to produce a forecast of the probable cost of a future projects, before the building has been designed in detail and contract particulars are prepared (Seely 1996). Smith (1995) observes that the initial estimate is implemented for review has a particularly significant role because it is the basis for the release of funds for further studies of estimates, and because it becomes the marker against which all subsequent estimates are compared. Moreover, performance and overall project success are often measured by how well the actual cost compares to the early cost estimates (Oberlender & Trost 2001). In this context, the use of the SPBP cost model is to prepare strategic and conceptual estimates for budgeting that can provide the function of cost check and control for the conceptual and design stages.

Data Analysis

Currently, in Korea, different types of apartment households are typically developed and produced by unit gross area because these developments have been seriously affected by housing supply legislation. Other apartment building project data related to cost can also be analyzed and/or categorized by the same criteria (e.g, by unit gross area). Consequently, the results of a data analysis are expected to yield similar patterns. To develop a cost model, 124 apartment buildings data from 11 housing complex projects in Korea are collected. The collected data should be normalized in respect to escalation, regional location, and system specification. This normalization is performed using the historical cost index of 1.025 for converting data of 2005 year to 2007 year—published the Korean government—of each building type. Because of Korea's relatively small territory, there is little point in normalizing the data for regional location, and system specification. Then, a cost database is built. For practicality, some of the impact factors are chosen as parameters for regression equations. As seen in Table 1, all of this information is accumulated and analyzed in the cost database for every single area type: type 84(m2), 59(m2), and 49(m2). Through a comprehensive analysis of the plan of apartment buildings, it is determined that all the buildings are either the singular type of unit gross area or a combination of plural types. Type is a set of same unit area households. Thus, all historical cost data classified to make the singular type of unit gross area are classified independently.

Data Preprocessing

Data preprocessing is a commonly used preliminary data mining practice, includes any type of processing procedure that acts on and prepares raw data for another processing procedure. There are different methods used for data preprocessing. Sampling selects a representative subset from a large population; Transformation, which manipulates raw data to produce a single input. Denoising removes "noise" from data. Normalization organizes data for more

Table 1: Configuration of Cost Database

Category	Factors / Information	
Parameter	(X1)	Number of households
	(X2)	Gross floor area
	(X3)	Number of unit floor households
	(X4)	Number of elevators
	(X5)	Number of floors
	(X6)	Number of piloti with household scale
	(X7)	Number of households of unit floor per elevator
Cost information	Total cost and unit cost per each impact factor	
Other information	Building number, Structure type, Floor height, Number of pit floors, Pit height, Roof type, Piloti type, Piloti height, Top floor type, Shape of floor plan, and Year of design.	

efficient access. And Feature extraction pulls out specified data that are significant in a particular context (Han & Kamber 2003). Construction project's cost data normally have "noise," such as internal errors or abnormal values that can have a negative effect on the reliability confidence and accuracy of the estimating result. To address this problem, a data preprocessing process is developed. In this process, the dominant cost variance governing parameter is determined based on PCA, and then normalization is conducted with the dominant parameter, and denoising is performed with sampling applied by interval estimating. In observing the impact factors of the cost database, it is intuitive that some factors will be highly correlated to cost, while also demonstrating strong correlations amongst themselves. The correlation coefficient indicates the strength and direction of a linear relationship between two random variables. In fact, a high bivariate correlation is identified: gross floor area, number of households, and number of floors. A high correlation is also identified amongst the factors themselves which means there is probability of multicollinearity. Using the factor analysis function of SPSS with the correlation matrix extraction approach in the SPBP, two components are extracted from eight variables within type 49 and type 59, and three components are extracted for type 84. The first components of PCA can represent most of the information of the original data. The percentage of variance, explained by the first component, is over 57% for all types. More precisely, X1 and X2 have the highest coefficient (over 0.95) of the component matrix (Table 2). Thus, X2 (gross floor area) is selected for the dominant parameter for all three building types because its correlation coefficient to cost is higher than X1.

Based on confidence intervals, interval estimation is described with an upper and lower limit and used to indicate the accuracy of an estimate. With the assumption that unit cost data, normalized by gross floor area ($/m2), are approximated by the normal distribution, normalization and statistical sampling using interval estimation is conducted. As a result, it is found that there are 25 cases out of 49 with a 95% Confidence Level (CL) and 30 out of 49 with a 99% CL in type 84; 13 out of 25 with a 95% CL and 9 out of 25 with a 99% CL in type 59; and 7 out of 16 with a 95% CL and 9 out of 16 with a 99% CL in type 49. These cases are selected to be used in deriving the final cost estimating equations (Table 3). A common rule of thumb from the Central Limit Theorem (CLT) is that a normal distribution can be used for approximation when $n \geq 30$. Accordingly, there are over 30 cases for type 84 the Z – distribution was used for them, and in case of type 49 and 59 which are less than 30 cases respectively, the t - distribution was applied to these types.

Table 2: Principle Components Scores

	Components (Type 49)		Components (Type 59)		Components (Type 84)		
	1	2	1	2	1	2	3
X1	**0.968**	-0.190	**0.967**	0.166	**0.956**	0.201	-0.164
X2	**0.954**	-0.072	**0.978**	0.120	**0.962**	0.132	-0.202
X3	0.855	0.478	0.669	0.637	0.781	0.534	0.291
X4	0.161	0.936	-0.096	0.954	-0.176	0.946	-0.080
X5	0.152	-0.915	0.698	-0.361	0.542	-0.457	-0.594
X6	0.850	0.284	0.509	-0.530	0.493	-0.244	0.627
X7	0.768	-0.441	0.760	-0.355	0.807	-0.389	0.314
Cost	0.949	-0.074	0.967	0.286	0.952	0.118	-0.139
Eigen values	4.838	2.263	4.605	1.930	4.575	1.672	1.022
% of variance	60.475	28.284	57.567	24.230	57.187	20.902	12.776
Cumulative %	60.475	88.759	57.567	81.687	57.187	78.089	90.865

Table 3: Results of Interval Estimation for Each Confidence Level

Type	49			59			84		
Confidence level	100%	99%	95%	100%	99%	95%	100%	99%	95%
Mean	476.71	476.81	473.71	565.22	552.09	556.86	531.65	527.75	531.78
Standard deviation	14.69	4.63	3.10	57.74	16.32	14.07	66.89	11.48	8.81
Lower limit	457.20	464.63	467.49	450.31	529.52	537.60	470.94	507.04	512.92
Upper limit	508.38	488.78	485.93	708.28	600.91	592.83	928.17	556.26	550.38
Number of cases	16	9	7	25	17	13	49	30	25

Table 4: Parametric Equations Using Stepwise Multiple Regression

Type	Confidence level	Equations	Standard errors	t-statistics	Significance level	Adjusted R^2
49	100%	= – 52188 + 487.680 * X2	74108	16.484(X2)	0.000(X2)	0.948
	99%	= – 106914 + 495.272 * X2	21633	21.584(X2)	0.000(X2)	0.985
	95%	= – 91524 + 493.809 * X2	13384	34.735(X2)	0.000(X2)	0.995
59	100%	= 311396 + 547.368 * X2 – 26698.939 * X5	113245	26.359(X2) -2.878(X5)	0.000(X2) 0.009(X5)	0.980
	99%	= – 9972 + 484.533 * X2 + 68708.898 * X3	45159	25.492(X2) 3.016(X3)	0.000(X2) 0.010(X3)	0.996
	95%	= – 10372 + 492.361 * X2 + 63810.302 * X3	36666	26.483(X2) 2.687(X3)	0.000(X2) 0.023(X3)	0.998
84	100%	= 250069 + 470.656 * X2	182810	23.991(X2)	0.000(X2)	0.922
	99%	= – 6665 + 529.425 * X2	55567	60.087(X2)	0.000(X2)	0.992
	95%	= – 16091 + 531.382 * X2	42507	71.992(X2)	0.000(X2)	0.995

Deriving Cost Estimating Relationships

To determine cost estimating relationships, multiple regression analysis —a statistical process for analyzing the relationship between quantitative variables—is utilized. The multiple regression method is traditionally adopted for the purpose of developing parametric cost estimating equations, its main objective is to build a mathematical formula that will assist in accurately predicting the impact on a variable when a related variable changes. However, to prevent the presence of multicollinearity, the stepwise approach is applied using regression function of SPSS. Stepwise approach is a combination of forward and backward procedures. Forward begins with no independent variables in the equation. The high correlated variable is entered into the equation first. The rest are entered into the equation depending on their contribution. On the contrary backward begins with all variables in the equation and sequentially removes them (Park et al. 2004). Table 4 reports the regression equations using preprocessed data. These equations have different confidence levels and different estimated effects of the individual variable on building cost. Collectively, the variables of type 84 with a 100% CL account for 92.2% of the total variance of project cost; this figure rises to 99.5% when the data are preprocessed with a 95% CL, which is indicated by the adjusted R2. Type 49 and type 59 demonstrate the same results. In addition, t-values of absolute value is over 3.3257 with 99% CL and over 2.5098 with 95% CL. And more, significance levels of independent variables in these equations are below 1%. Hence, these regression equations are cured from probability of multicollinearity.

In terms of the preprocessing effect, when the case data is sufficient and relatively conformed to the central limit theorem (type 84), the mean of the difference between the actual and estimated value decreased, and the variation of the estimated value is stabilized relative to the degree of the preprocessing level (i.e., confidence level). However, when the case data is not sufficient (type 59), the variation in the estimated values using \$/GFA and \$/NH is at least twice that of the SPBP values. With the regression method, on the other hand, there is no significant change in the mean of difference and variation relative to the degree of the preprocessing level. As summarized Table 5, the relationship between the cost estimate equations (the SPBP equation with confidence levels, \$/GFA, and \$/NH) and the accuracy ratio dimensions are assessed using one-way ANOVA and post-hoc Tukey procedures. In the case of type 84, the results indicated a significant enhancing trend of accuracy ratio across the preprocessing level. The p-values of 0.001 (F-statistic is 5.884) advocate that means are significantly different at 1% significant level. Moreover, the post-hoc test results also denote the level of statistical significance found in the test at 5% significance level, support this. However, there is no significantly difference among means of type 59. Ultimately, the SPBP equations yield estimated values that are more reliable and stable than the values yielded from the traditional methods. Not only do the SPBP equations have a lower percentage of error than the traditional methods, they also demonstrate a more stabilized variation. Further, after more and more case data were inputted into the model, the SPBP model demonstrated the tendency of better stabilization of estimated values and variation (Table 6).

Table 5. Comparison of Accuracy Ratio (SPBP vs. Traditional methods)

Type	parameter							SPBP equation with CL of			$/GFA	$/NH
	X1	X2	X3	X4	X5	X6	X7	100%	99%	95%		
59	19	1597.1	2	1	10	0	2	126.21%	123.83%	124.15%	124.02%	123.93%
	24	1992.5	2	1	12	0	2	117.72%	118.94%	119.53%	122.57%	124.01%
	26	2188.4	2	1	14	2	2	118.19%	123.64%	124.36%	128.75%	128.48%
	28	2321.5	2	1	14	0	2	124.25%	128.77%	129.59%	134.93%	136.69%
	30	2492.3	2	1	15	0	2	116.21%	121.68%	122.52%	128.39%	129.81%
	38	3185.2	4	2	10	2	2	103.56%	104.74%	105.03%	104.29%	104.50%
	56	4420.3	4	1	15	2	4	97.61%	100.80%	101.42%	104.65%	111.36%
	58	4565.1	4	1	15	0	4	99.66%	102.43%	103.08%	106.71%	113.88%
	60	4687.7	4	1	15	0	4	102.55%	105.01%	105.71%	109.71%	117.95%
	60	4687.7	4	1	15	0	4	110.83%	113.48%	114.23%	118.56%	127.46%
	60	4703.3	4	1	15	0	4	100.34%	102.70%	103.38%	107.32%	115.00%
			Mean					**110.65%**	**113.28%**	**113.91%**	**117.26%**	**121.19%**
			Variance					**1.08%**	**1.09%**	**1.10%**	**1.24%**	**0.90%**
84	16	1809.1	2	1	9	2	2	122.65%	105.91%	105.25%	107.09%	102.96%
	18	2021.1	2	1	9	0	2	125.91%	111.46%	110.88%	112.62%	109.04%
	18	2024.8	2	1	9	0	2	128.22%	113.55%	112.96%	114.73%	110.88%
	20	2189.9	2	1	10	0	2	130.62%	117.57%	117.04%	118.74%	117.89%
	20	2244.7	2	1	10	0	2	126.22%	114.17%	113.68%	115.29%	111.67%
	22	2410.4	2	1	11	0	2	121.68%	111.57%	111.15%	112.62%	111.75%
	24	2627.2	2	1	12	0	2	122.70%	114.26%	113.90%	115.29%	114.50%
	24	2655.4	2	1	13	2	2	109.23%	101.90%	101.59%	102.81%	101.02%
	26	2879.9	2	1	14	2	2	126.99%	120.08%	119.77%	121.11%	118.86%
	26	2955.7	2	1	14	0	2	126.13%	119.75%	119.47%	120.76%	115.49%
	28	3091.4	2	1	15	2	2	106.66%	101.97%	101.76%	102.81%	101.23%
			Mean					**122.46%**	**112.02%**	**111.59%**	**113.08%**	**110.48%**
			Variance					**0.59%**	**0.41%**	**0.41%**	**0.42%**	**0.40%**

Table 6. Result of ANOVA and Tukey post-hoc test of Means of Accuracy Ratio

	SPBP 100%		SPBP 99%		SPBP 95%		$/GFA		$/NH	
	mean	variance	mean	variance	mean	variance	mean	variance	mean	variance
accuracy ratio										
type 59	110.65	1.08	113.28	1.09	113.91	1.1	117.19	1.24	122.46	0.90
type 84*	122.19	0.59	112.02	0.41	111.59	0.41	113.08	0.42	110.48	0.40

* Means with similar letter in a row are significantly different at p < 0.05 for all means, n = 11

CONCLUSIONS

For a construction project to progress smoothly, effective cost estimation is vital, particularly in the conceptual and schematic design stages. In these early phases, despite the fact that initial estimates are highly sensitive to changes in project scope, owners require accurate forecasts which reflect their supplying information. In this context, a parametric method was developed using regression analysis, an efficient and highly usable technique. Because parametric cost models primarily require historical data regarding appropriate cost variance governing parameters, this research proposed a statistical methodology for determining these parameters. Furthermore, this methodology included data preprocessing (data cleansing), and the development of the SPBP cost model, which is based on multiple regression analysis that excludes multicollinearity. This regression method was applied to the data of 90 apartment building projects in Korea. Then, the effect of preprocessing and the reliability of the cost model were tested using 22 cases. It was verified that data preprocessing has a positive impact on the enhancement of estimate accuracy, if the case data sufficiently conforms to the central limit theorem in deriving the cost estimating relationship. It was also found that, regardless of the number of case data used, the SPBP equations are more stable than their more conventional counterparts. While it must be noted that this research is based on limited case data, and that additional research and testing must be conducted to further validate the model and generalize the effect of data preprocessing, this research has immediate benefits. Indeed, the research findings confirm that the SPBP cost model can be an effective tool for preparing accurate strategic and conceptual cost estimates. Furthermore, the proposed model will contribute to enhancing cost estimating by increasing the value of historical data, improving the reliability of the cost estimating relationship, and by stabilizing estimates. Finally, this model is not limited to one construction project type; in fact, after customization, the findings of this study can also be usefully applied, domestically and internationally, to different types of construction projects.

ACKNOWLEDGMENT

This research was supported by a grant (R&D06CIT-A03) from the Innovative Construction Cost Engineering Research Center, and by a grant (05CIT-01) from the Construction Technology Innovation Program funded by the Ministry of Land, Transport and Marine Affairs (Government of Korea).

REFERENCES

Akintoye, A. (2000) “Analysis of Factor Influencing Project Cost Estimating Practice”, Construction Management and Economics, 18, pp. 77~89.

Ministry of Land, Transport and Marine Affairs of Korea (2007). Announcement of historical cost indices for public apartment project under price cap regulation, #2007-79.

Chan, L. S. & Park, M. (2005) “Project Cost Estimation Using Principle Component Regression”, Construction Management and Economics, 23, pp. 295~304.

Hegazy, T. & Ayed, A. (1998). “Neural Network Model for Parametric Cost Estimating of Highway Projects”, Journal of Construction Engineering and Management, 124(3), ASCE, pp. 210~218.

Jarde, A. & Alkass, S. (2007) “Computer-Integrated System for Estimating the Costs of Building Project”, Journal of Construction Engineering and Management, 13(4), ASCE, pp. 205~223.

Kirkham, R, J (2007). Ferry & Brandon‘s Cost Planning of Buildings - 8th ed.: Blackwell publishing.

Oberlender, G. D. & Trost, S. M. (2001). “Predicting Accuracy of Early Cost Estimates Based on Estimate Quality”, Journal of Construction Engineering and Management, 127(3), ASCE, pp. 173~182.

Paul Teicholz (1993). “Forecasting final cost and budget of construction projects”, Journal of Computing in Civil Engineering, ASCE, vol. 7(4), pp.511~529.

Sharma, S. (1996). Applied Multivariate Techniques: John Wiley & Sons, Inc.

Trost, S. M & Oberlender, G. D. (2003). “Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression”, Journal of Construction Engineering and Management, 129(2), ASCE, pp. 198~204.