PROJECT DATA WAREHOUSE MANAGEMENT WITH MULTIVARIATE ANALYSIS

Jui-Sheng Chou, P.E., Ph.D.
Associate Professor in Project Management, Department of Construction Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan
jschou@mail.ntust.edu.tw

Hsin Wang, Graduate Research Assistant
Project Management Division, Department of Construction Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan
m9705116@mail.ntust.edu.tw

Hsien-Cheng Tseng, MBA
Former Graduate Student

Abstract

Many studies have generated cost estimating relationships (CERs) for transportation projects via data analysis. Some studies collected data from databases, while others sourced data from conventional paper-based formats. When cost data were not in a consistent format, many studies failed to discuss the streamlining of pattern recognition. This study adopts a standard procedure for identifying CERs for transportation projects. A pavement maintenance and rehabilitation project type was selected as a case study for extracting data and concealed prediction rules. Linear and log-linear statistical approaches were employed to create optimal models. The resulting optimum estimation models via knowledge discovery in databases process can be then integrated into an expert system to facilitate information management and generate preliminary budgets for transportation agencies.

KEYWORDS: Data Warehouse, Multivariate Analysis, Estimation, Project Management, Transportation

INTRODUCTION

Construction of public infrastructure is critical in helping nations increase their global competitiveness and regional economic development. When the density of transportation network reach its peak level of saturation, few new roadways are constructed and in-service infrastructure must be maintained and rehabilitated to keep roadways safe and operating at a desirable level of service.

Maintenance operations have become increasingly complex due to rapid traffic growth, funding limitations, and a shortage of skilled workers. Government agencies have difficulty hiring and retaining qualified employees. To address these challenges, an urgent need exists for agencies to document project estimation practices, examine maintenance policies, and develop innovative budgeting schemes that improve preliminary cost estimates and control project costs.

With increased emphasis placed on taxpayer awareness of government image, the transportation agency requires reliable cost-estimation systems that provide quality and safety assurances to local residents and communities. This research proposes an effective framework that establishes the processes required for daily operations in the data warehouse, and transforms the wealth of data and engineering experience into effective models.

LITERATURE REVIEW

Accurately and timely estimating project costs is vital to successful project delivery. A wide range of accuracy patterns for estimations exists in the beginning stages of a project. The differences should be gradually narrowed down as a project progresses. Once the initial gap between estimated and actual costs can be reduced during the initiation stage, a project should proceed smoothly as additional input information is acquired (Chou 2009). Many cost estimation methods have been developed to increase the accuracy of project budgeting. Existing cost estimation methods can be categorized as analogous cost estimating, bottom-up estimating, computerized tools and artificial intelligence, and parametric modelling.

The parametric method estimates construction cost based on such parameters as project square footage or, say, number of beds in a hospital. This technique uses validated relationships between known technical and cost characteristics of a project obtained from historical data. Many studies have investigated the non-linear relationships between cost and project characteristics and have attained high levels of accuracy; however, the estimation accuracy depends on the quality of underlying data and sophisticated statistical techniques employed to construct the model (Harbuck 2002; Lowe et al. 2006; Phaobunjong and Popescu 2003).

A national cooperative research report indicated that most US departments of transportation (DOTs) used two parameters, i.e., number of lanes and project length, to estimate highway project cost during early project stages (Anderson et al. 2007). This approach is known as the lane-mile approach. Chou et al. (2006) proposed a quantity-based approach to estimate highway project costs (Chou et al. 2006). The advantages of this approach include segregation of unit price from estimation in the initial stage, which reduces uncertainties due to market conditions and time inflation, and early introduction of semi-detailed quantity estimates to continuously track quantity changes as the project proceeds to subsequent phases.

RESEARCH METHODOLOGY

Transportation agencies typically generate large amounts of detailed cost data for a project lifecycle, and this data is generally distributed across functional systems or saved in conventional paper-based formats. With timely access to and reuse of data for previous projects, one can create useful rules and apply these rules to future projects and decision-making processes (Mallach 2000; Turban and Aronson 2005). The knowledge discovered from historical data is considered paramount intangible business intelligence (BI) for an enterprise.

The Data Warehouse Institute defines BI as processes, tools, and technologies required to transform data into information, and then into knowledge and effective business plans. A data warehouse is defined as a subject-oriented, integrated, non-volatile, and time variant

collection of relevant data that ensures that historical data is consistent and easy to retrieve. A data warehouse also facilitates convenient access to and reuse of data to support of management decisions (Inmon 2002; Rujirayanyonga and Shi 2006).

The proposed research roadmap was divided into the following three phases: establishment of a dimensional data warehouse; statistical prediction modeling; and, development of the prototype expert system.

First, to construct a dimensional data warehouse that accommodates project data for long-term use, sample data was gathered from the five maintenance offices of the Ministry of Transportation and Communications, which are geographically distributed in Taiwan. The exemplary data warehouse for transportation construction projects contains such information as project type and description, project location, project length, number of lanes, traffic volume, terrain type, quantity calculation, unit price, estimated cost, actual cost, and completion time. The data warehouse has flexibility to scale-up depending on the project information that must be stored.

In the second stage, one selects one of the data marts and generates query tables of interest. The optimal statistical estimation models can thus be created by exploring non-volatile and cost- and subject-oriented data marts. Specifically, a data mart is a subset of a data warehouse covering a particular subject or department data for a specific purpose or personal query, visual presentation, and data mining.

For the final development stage of the proposed system, this study adopts case study method to demonstrate the framework of graphical user interfaces with embedded forecasting models. A parametric estimating technique was used to establish the relationships between project parameters (*i.e.*, characteristics, functions and features) and engineering quantity of work items. Item cost can then be derived by multiplying predicted quantity with the corresponding item unit price, which is readily available from the Taiwan Public Construction Commission (TPCC) database. All of these functions are integrated into the proposed expert system.

During project initiation, estimating the costs of all work items is unnecessary as changes can occur during a project's lifecycle. Therefore, based on observations of descriptive statistics of the extracted data, high-frequency and high-cost work items in previous projects were first identified and serve as the primary object in model development. Total project cost (TPC) can be derived via the summation of direct costs and indirect costs, as in Eq. (1). Indirect costs, such as business taxes, contractor and management overheads, are typically a fixed percentage of direct costs in public transportation projects.

$$TPC_j = DC_j + IC_j + Contingency_j$$

$$= \sum_{i=1}^{n} \frac{ItemQty_i \times UnitPrice_i}{CCPs_j} + IC_j + Contingency_j \quad (1)$$

where,

TPC_j: Total project cost of the j[th] project

DC_j: Direct costs of the j[th] project

IC_j: Indirect costs of the j[th] project

$Contingency_j$: Contingency of the j[th] project, represented as a percentage

$CCPs_j$: Cumulative cost percentage for standard work items of the j[th] project, which is equal to (Cumulative standard work item costs of the j[th] project) / (Direct costs of the j[th] project)

$ItemQty_i$: i[th] standard work item quantity of the j[th] project

$UnitPrice_i$: i[th] standard work item unit price of the j[th] project

ARCHITECTURE OF THE DIMENSIONAL DATA WAREHOUSE

Project documentation is generally stored in either electronic or paper-based files; however, these files are often stored in an unorganized and fragmented fashion (Weiser and Morrison 1998). Based on interviews with experienced engineers working in Taiwan's transportation agencies, such storage practices are common in the construction industry. Consequently, attempts to reuse stored information and generate beneficial rules are hindered. To extract valuable information from documentation efficiently, establishing an effective e-storage database and expediting information flow are important tasks. The database allows project estimators and managers to transform unprocessed data, which incorporates implicit intelligence, into explicit knowledge. A dimensional model-based data warehouse is proposed that performs such functions and accommodate mixed cost data related to transportation projects (Fig. 1).

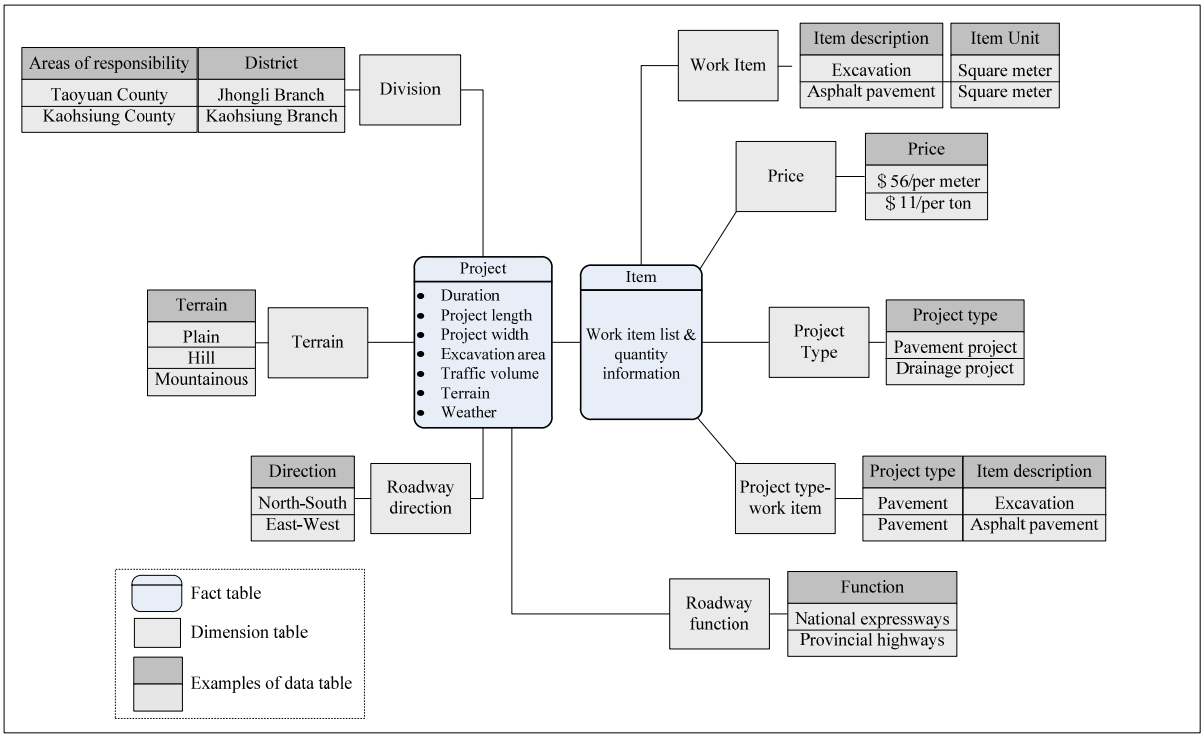

Figure 1: Fact constellation schema for transportation project data

The dimensional data model uses two tables, namely, fact tables and dimension tables. Fact tables store primary project-level and item-level data such as project duration, roadway

length, roadway width, daily traffic volume, engineering quantity of work item, and work item quantity at completion. Dimensional tables store descriptions of work items, terrain, roadway direction and roadway type, construction division in charge, historical item prices, and project type. All fact and dimensional tables are scalable and can be expanded to accommodate additional data fields of interest. The primary functions of the dimensional data model are to increase query efficiency and avoid data redundancy as the database expands. Table 1 shows the tables of the dimensional data warehouse.

Table 1: Data tables of the dimensional data warehouse

Type	Dimension	Description	Objectives	Table name
Fact table	1. Project	Parameters in the project.	Data storage	FactTbl_Project
	2. Item	Numerical data of the work item, including estimation and final accounts of work item quantities	Data storage	FactTbl_Item
Dimension table	1. Work Item	Description of work items and their units.	Reduce redundancy	DimTbl_ItemDescrip
	2. Price	Price listing of work items.	Query & analysis	DimTbl_ItemPrice
	3. Project type	Including retaining walls, pavement, bridge, tunnel, drainage, traffic control devices, landscape planting maintenance machine and others.	Query & analysis	DimTbl_ProjTypeDescrip
	4. Project type -work item	Work items listing with a specific project type.	Query & analysis	DimTbl_ProjItem
	5. Roadway function	Including national expressways, provincial highways, county highways, rural highways and exclusive highways.	Query & analysis	DimTbl_SysTypeDescrip
	6. Division	Contact information and location of division.	Reduce redundancy	DimTbl_SectionDescrip
	7. Terrain	Including terrain plain, hill and mountainous.	Query & analysis	DimTbl_TerrainDescrip
	8. Roadway direction	The description of roadway direction.	Query & analysis	DimTbl_DirectionDescrip

Data fields in the proposed data warehouse contain general project information. Most of the conceptual information can be acquired from an illustration of a typical roadway section. Data from the visual graph can be categorized as basic design parameters (*i.e.*, lane number, roadway width, and work area), environmental factors (*i.e.*, terrain type, level of precipitation, and number of rainy days), and traffic volumes (*i.e.*, traffic volume of various vehicle types). All data are easily accessible by estimators or engineers during the early project stages. Based on the dimensional data warehouse schema, the relationships among data sets can be drilled down or rolled up within a data cube. Dimensions of a work item, project type and roadway type, for instance, can be obtained from a populated database into a data mart. A data mart with increased dimensions can be constructed when necessary. In the following section, a particular project type is extracted into a sample data mart for subsequent statistical parametric modeling.

DATA ANALYSIS AND MODEL EVALUATION

This section describes a case study using a pavement maintenance project data mart extracted from a data warehouse for model development. Following data preprocessing, linear and log-linear regression models were constructed after carefully considering various measured response variables to establish an enhanced approach. Meaningful and interesting patterns were then verified through a hold-out sample.

Data description

The raw data were collected from the 1st to 5th engineering districts of Taiwan's Directorate General of Highways. Rehabilitation projects were randomly collected from the districts, including repair of road bases, slopes, and drainage, and pavement maintenance, landscaping and planting. Although obtaining a complete collection of data for all past projects is unlikely, the primary goal is to ensure that the data warehouse fits project data in practice, and can serve as a centralized data management system for the transportation agency. This study utilized the data mart for pavement maintenance operations (PMOs) which account for the largest number of projects and outnumber other project types by two or three times statistically, underscoring the importance and priority of PMOs in recent years.

Parametric formulation

The purpose of parametric modeling is to derive an empirical formula for estimating engineering quantity for major work items (response variables) with a set of significant parameters (predictor variables). A preliminary total project cost can be derived using Eq. 1 by multiplying item quantity by available item unit prices from the TPCC database. Based on literature findings, interviews with experienced estimators and availability of significant factors, 19 parameters from each project were obtained as possible input predictors for regression analysis. From descriptive statistics of original data, all data had skewed patterns. Therefore, the performance of the log-linear model was compared with that of the linear model. Furthermore, to assess the suitability of response variables, raw engineering quantity (REQ), engineering quantity per lane meter, engineering quantity per project length in meters, and engineering quantity per construction area in square meters were used to develop the linear and log-linear models.

As this study focuses on the early project phase, estimating the cost of all work items to determine total project cost is unnecessary for the following reasons: (1) an overview of project estimation is the key focus of this study; (2) emphasis should be placed on major work items (e.g., items accounting for over 1%, thereby increasing management effectiveness and reducing estimating burden at a conceptual stage); and, (3) detailed project and work information at this stage may be unavailable or difficult to obtain.

The following four major work items were identified out of 22 for a generic PMO project based on high cost percentage and high occurrence rate (cp%, or%): pavement of reclaimed densely graded asphalt concrete (65.7%, 92%); salvaging, hauling, and stockpiling reclaimable asphalt pavement (13.3%, 83%); reflective pavement markings (5.4%, 91%); and, sprinkling of the tack coat with emulsified asphalt (4.9%, 98%). Statistics show that the average cost percentage and occurrence rate for a generic project for the remaining 18 minor work items were less than 1% and 35%, respectively. The four major work items account for roughly 90% of the cumulative cost percentage for a typical project, and have an average occurrence rate of 91%. Hence, the major work items are considered standard items and are estimated first.

Model performance and validation

Nineteen data fields were extracted from the project data warehouse; most were identified as essential factors when estimating engineering costs in numerous studies (Akinci and Fischer

1998; Akintoye 1998; Al-Tabtabai et al. 1999; Baloi and Price 2003; Bell and Bozai 1987; Chou et al. 2006; Sanders et al. 1992). The selected parameters can be categorized as geometrical configurations of a roadway (roadway length, roadway width, roadway lane number, and construction area in lane meters and square meters), traffic volume (passenger car unit (PCU) and average daily traffic, including motorcycles, cars, and heavy trucks) and environmental factors (terrain, precipitation and average rainfall days near the construction area).

Analytical results indicate that the log-linear regression model had comparable predictive power and better calibration than the linear regression model. The non-standardized REQ of the log-linear models produces the best fit for data. Model results imply the predictors and response variables have a non-linear relationship, which can also be observed in advance based on the positive skewed distribution of variates.

Log transformation alleviates nonlinearity and multicollinearity of predictors, and improves explanatory power by 32% on average. Furthermore, mean absolute quantity prediction error of log-linear models was 10.9%, which was 24.2% less than that for the linear model using ten random hold-out test data samples.

Table 2: Randomly selected project samples for cost validation

Case No	Project Name	Cost Prediction Error (%)	
		Sum of Standard Work Item Costs	Total Project Cost via Eq. (1)
1	County highways No.102 30K+100~31K+333	22.0	16.0
2	Provincial highways No.4 0K+000~2K+140	13.8	9.0
3	Provincial highways No.1 16K+380~16K+760 & 20K+260~20K+940	8.2	10.6
4	Provincial highways No.3 75K+600~76K+800	2.3	1.9
5	Provincial highways No.1 51K+937~53K+000	9.1	6.5
6	Provincial highways No.3 36K+000~36K+500 & 40K+200~40K+600	11.0	1.5
7	Provincial highways No.7 14K+500~17K+020	17.3	5.7
8	Provincial highways No.7 11K+590~11K+920 & 14K+560~14K+690	10.6	19.1
9	Provincial highways No.3 415K+093~421K+372	3.9	25.9
10	County highways No.114 6K+750~7K+000 & 8K+500~9K+800	1.1	8.1
	Mean Absolute Prediction Error	9.9	10.4

Cost prediction accuracy is therefore evaluated using the log-linear quantity models by multiplying their mean unit costs and then comparing these predicted costs to actual costs. The mean absolute cost prediction errors of the summation of standard work item costs and total project cost (Table 2) are 9.9% and 10.4%, respectively, which in practice is satisfactory.

SYSTEM DEVELOPMENT AND ESTIMATING FLOW

Rapid application development (RAD), a process that emphasizes immediate deployment of a system while simultaneously maintaining quality and reducing development costs (Marakas 2006; Touran and Suphot 1997), offers a series of techniques for compressing analysis, design, construction, and test phases into a series of short, iterative development cycles (Fig. 2). Upon defining the model and prototype requirements using feedback from experienced engineers, the system was generated by defining the scope and analyzing requirements followed by deployment of the iteratively rapid approach.

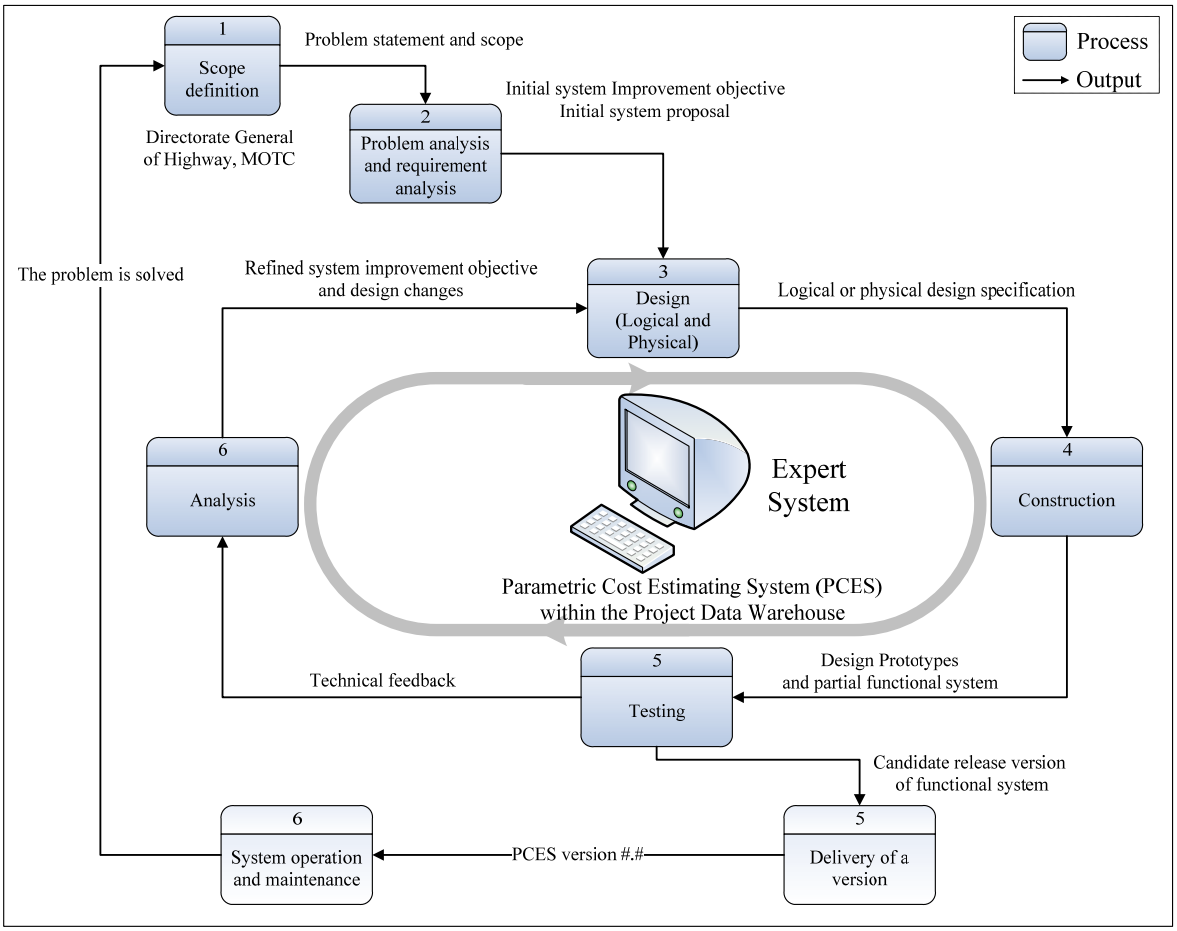

Figure 1: RAD approach for the expert system

CONCLUSIONS

This study provides researchers and practitioners with further insight into the relationships between various engineering quantity measurements and conceptual information (features, functions and characteristics) during the early stages of a project. A practical data-mining process is presented starting from collection of data, preprocessing, and construction of a novel project cost data warehouse. A parametric prediction technique is applied to establish useful estimation models. The derived response functions are constructed in linear and log-linear forms with a set of predictors. Through comparison of modeling results, the best estimation model is the natural logarithm of the quantity model with the original measurement unit; this model had a mean adjusted R-square of 0.668 and mean absolute

quantity prediction error of 10.9%. Validation results suggest that the natural logarithmic transformation reduces mean absolute quantity prediction error by 24.2% when compared to those of linear models. For total project cost estimation, the transformed models generated a mean absolute project cost prediction error of 10.4%, which satisfies generally acceptable level of accuracy (15–35%) during the early stages of a project. Future research should focus on model refinement for full-scale system implementation and an automatic link to the most recent unit price database to improve continuous estimation accuracy.

ACKNOWLEDGMENT

This research project was financially supported by the National Science Council, Taiwan under the contract No. 95-2221-E-194-001.

REFERENCES

Akinci, B., and Fischer, M. (1998). "Factors affecting contractors' risk of cost overburden." *Journal of Management in Engineering*, 14(1), 67-76.

Akintoye, A. (1998). "Analysis of factors influencing project cost estimating practice." *Construction Management and Economics*, 18, 77-89.

Al-Tabtabai, H., Alex, A. P., and Tantash, M. (1999). "Preliminary cost estimation of highway construction using neural networks." *Cost Engineering*, 41(3), 19-24.

Anderson, S., Molenaar, K., and Schexnayder, C. (2007). "Guidance for Cost Estimation and Management for Highway Projects During Planning, Programming, and Preconstruction." *NCHRP Report 574*, National Cooperative Highway Research Program, Transportation Research Board, WASHINGTON, D.C.

Baloi, D., and Price, A. D. F. (2003). "Modelling global risk factors affecting construction cost performance." *International Journal of Project Management*, 21, 261-269.

Bell, L. C., and Bozai, G. A. "Preliminary cost estimating for highway construction projects." *1987 AACE Transactions*, C.6.1-C.6.4.

Chou, J.-S. (2009). "Generalized linear model-based expert system for estimating the cost of transportation projects." *Expert Systems with Applications*, 36(3, Part 1), 4253-4267.

Chou, J.-S., Peng, M., Persad, K. R., and O'Connor, J. T. (2006). "Quantity-based approach to preliminary cost estimates for highway projects." *Transportation Research Record*(1946), 22-30.

Harbuck, R. H. (2002). "Using Models in Parametric Estimating for Transportation Projects." *AACE International Transactions*, EST.05(ES51), EST.05.1-EST.05.09.

Inmon, W. H. (2002). *Building the data warehouse*, Wiley Computer Publishing.

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

Mallach, E. G. (2000). *Decision support and data warehouse systems*, Boston : Irwin/McGraw-Hill.

Marakas, G. M. (2006). *Systems Analysis & Design*, McGraw-Hill.

Phaobunjong, K., and Popescu, C. M. (2003). "Parametric Cost Estimating Model for Buildings." *AACE International Transaction*, EST.13.1-EST.13.11.

Rujirayanyonga, T., and Shi, J. J. (2006). "A project-oriented data warehouse for construction." *Automation in Construction* 15(6), 800-807

Sanders, S. R., Maxwell, R. R., and Glagola, C. R. (1992). "Preliminary estimating models for infrastructure projects." *Cost Engineering*, 34(8), 7-13.

Touran, A., and Suphot, L. (1997). "Rank correlation in simulating construction costs." *Journal of Construction Engineering and Management*, 123, 297-301.

Turban, E., and Aronson, J. E. (2005). *Decision support systems and intelligent systems*, Upper Saddle River, NJ: Prentice-Hall.

Weiser, M., and Morrison, J. (1998). "Project memory: Information management for project teams." *Journal of Management Information Systems*, 14(4), 149-166.