COMPARISON OF PARAMETRIC MODELS FOR CONCEPTUAL DURATION ESTIMATION OF BUILDING PROJECTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AZİZ HELVACI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
CIVIL ENGINEERING

AUGUST 2008

Approval of the thesis:

COMPARISON OF PARAMETRIC MODELS FOR CONCEPTUAL DURATION ESTIMATION OF BUILDING PROJECTS

Submitted by **AZİZ HELVACI** in partial fulfillment of the requirements for the degree of **Master of Science in Civil Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _______________
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Güney Özcebe _______________
Head of Department, **Civil Engineering**

Assist. Prof. Dr. Rifat Sönmez _______________
Supervisor, **Civil Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Talat Birgönül _______________
Civil Engineering, METU

Assist. Prof. Dr. Rifat Sönmez _______________
Civil Engineering, METU

Assoc. Prof. Dr. İrem Dikmen Toker _______________
Civil Engineering, METU

Assoc. Prof. Dr. Murat Gündüz _______________
Civil Engineering, METU

Ali Gökçe Yaşar, M.S. _______________
Civil Engineer, YAPI KREDİ KORAY

Date: _______________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Aziz Helvacı

Signature :

ABSTRACT

COMPARISON OF PARAMETRIC MODELS FOR
CONCEPTUAL DURATION ESTIMATION
OF BUILDING PROJECTS

Helvacı, Aziz
M.S., Department of Civil Engineering
Supervisor: Assist. Prof. Dr. Rifat Sönmez

August 2008, 85 pages

Estimation of construction durations is a very crucial part of project planning, as several key decisions are based on the estimated durations. In general, construction durations are estimated by using planning and scheduling techniques such as Gannt or bar chart, the Critical Path Method (CPM), and the Program Evaluation and Review Technique (PERT). However, these techniques usually require detailed design information for estimation of activity durations and determination of the sequencing of the activities. In some cases, pre-design duration estimates may be performed by using these techniques, however, accuracy of these estimates mainly depends on the experience of the planning engineer.

In this study, it is aimed to develop and compare alternative methods for conceptual duration estimation of building constructions with basic data information available at the early stages of projects. Five parametric duration estimation models are developed with the data of 17 building projects which were constructed by a contractor in United States. Regression analysis and artificial neural networks are used in the development of these five duration estimation models. A parametric cost estimation model is developed using regression analysis for cost estimations to be used in calculating the prediction performances of cost based duration estimation models. Finally, prediction performances of all parametric duration estimation models are determined and compared. The models provided reasonably accurate estimates for construction durations. The results also indicated that construction durations can be predicted accurately without making an estimate for the project cost.

Keywords: Construction Duration, Regression, Artificial Neural Networks, Conceptual Estimation

ÖZ

BİNA PROJELERİNDE KEŞİF ÖNCESİ SÜRE TAHMİNİ İÇİN PARAMETRİK MODELLERİN KARŞILAŞTIRILMASI

Helvacı, Aziz
Yüksek Lisans, İnşaat Mühendisliği Bölümü
Tez Yöneticisi: Yrd. Doç. Dr. Rifat Sönmez

Ağustos 2008, 85 sayfa

Proje başarısı için gerekli bazı kararlar süre tahminlerine dayanılarak verildiğinden, inşaat sürelerinin tahmini proje planlamasının çok önemli bir parçasıdır. İnşaat süreleri genelde çubuk diyagramları, CPM ve PERT gibi planlama ve programlama teknikleriyle tahmin edilmektedir. Ancak bu teknikler aktivite sürelerinin tahmini ve aktivitelerin sıra ilişkilerinin belirlenmesi için çoğunlukla detaylı mühendislik ve mimari proje bilgilerine gerek duyarlar. Bazı durumlarda, bu teknikler kullanılarak dizayn öncesi süre tahminleri yapılabilir, ancak bu tahminlerin kesinliği çoğunlukla planlama mühendisinin tecrübesine bağlı olmaktadır.

Bu çalışmada, projelerin ilk aşamalarında elde edilebilir temel verilerle bina inşaatları için alternatif keşif öncesi süre tahmin yöntemlerinin geliştirilmesi ve karşılaştırılması amaçlandı. Bir yüklenici tarafından Amerika Birleşik Devletlerin'de inşa edilmiş 17 bina projesinin verileriyle beş tane parametrik süre tahmin modeli geliştirildi. Bu beş süre tahmin modelinin geliştirilmesinde regrasyon analizi ve yapay sinir ağları kullanıldı. Maliyet üzerine kurulan süre tahmin modellerinin tahmin performanslarının hesaplanmasında kullanılacak olan maliyet tahminleri için regrasyon analizi ile parametrik bir maliyet tahmin modeli geliştirildi. Sonunda, bütün parametrik süre tahmin modellerinin tahmin performansları belirlendi ve karşılaştırıldı. Modeller inşaat süreleri için makul derecede doğru tahminler sağladı. Sonuçlar ayrıca inşaat sürelerinin proje maliyetleri için tahmin yapılmadan da kestirilebileceğini gösterdi.

Anahtar Kelimeler: İnşaat Süresi, Regrasyon, Yapay Sinir Ağları, Keşif Öncesi Tahmin

TO MY FAMILY,

ACKNOWLEDGMENTS

I want to express my gratitude to Dr. Rifat Sönmez for his unlimited guidance and supervision at each step of this research. His encouragements and sincere approach throughout the preparation of this thesis made everything easier.

I would also like to thank my uncle and his wife for their helps and moral supports throughout my studies.

Finally, I would like to thank my family for their endless love and support to achieve my goals in life.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AD	Australian Dollar
ANN	Artificial Neural Network
Area/Unit	Area per Unit
BPT	Breakpoint
BTC	Bromilow's Time-Cost
CPM	Critical Path Method
CCRC	Continuing Care Retirement Community
GFA	Gross Floor Area
MAPE	Mean Absolute Percentage Error
Mas	Masonry
MLR	Multiple Linear Regression
NoF	Number of Floors
Per(C+H)	Combined Percent Area of Commons and Health Center
Per(P)	Percent Area of Structured Parking
PERT	Program Evaluation and Review Technique
PE	Percentage Error
Pre	Precast
RC	Reinforced Concrete
St	Steel
USD	United States Dollar
W	Wood

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Construction duration is identified as one of the most important criteria along with cost and quality for measuring the overall success of construction projects. Therefore, completion of projects within the planned time is very critical for achieving successful projects. However, projects may not be completed in durations that they are initially estimated. Consequently, contractors may loose money and reputation. Construction delay, which is a common problem in the sector, may occur due to some factors such as poor productivity, variation of orders, unforeseen weather conditions or simply due to underestimated construction duration. From the contractors' point of view, underestimation of construction duration leads to reorganization and reallocation of resources which was not planned initially. On the other hand, overestimation of construction duration can also be as bad as underestimation of construction duration. It may lead to spending extra money on the allocation of extra resources which is unnecessary, in fact. Hence, a reliable estimation of construction duration is very important and critical for contractors.

In practice, construction durations are generally decided on the basis of client's deadline or are estimated by making use of planning and scheduling techniques such as bar charts, the critical path method (CPM) and the program evaluation and review technique (PERT). Ng et al. (2001) stated that "Many contractors simply assume that the contract duration set by the client is realistic and prepare their bids accordingly". However, this is risky and may increase contractor's cost if the project can not be completed in time. On the other hand, contractors may not want to use scheduling techniques (Bar chart, CPM and PERT) in order to avoid spending large amounts of time and money on the detailed estimation of construction duration (Ng et al, 2001). These scheduling techniques usually require detailed analysis of all activities to complete the project. Planner should identify the completion times of each activity and their sequencing. However, this is possible mostly when the detailed design information and specifications are available. Planner may also use alternative techniques to estimate the duration of construction with limited data available at the early stages. Nevertheless, this is a subjective method and the reliability of the estimation depends on the estimator's individual experience on similar projects and judgment to interpret the new project. In order to minimize this subjectivity, there is a need to develop a systematic and quantitative methodology for estimation of duration at the early stages of the project.

Conceptual duration estimate can be defined as the forecast of construction project's duration based only on general information available at the early stages of the project. In conceptual estimates, methodologies are usually developed with historical information. Historical information may include basic elements of the projects defined as parameters. Then, methodology developed to forecast the construction duration becomes parametric method. The parametric method uses project characteristics to form a model to forecast project duration. The model is developed by establishing relationships between the parameters and the project duration. Several methods can be used to develop the model such as artificial neural network (ANN) and linear regression. The model may be simple with having only one parameter, or it may be complex with multiple parameters.

The importance of the construction duration estimation, shortcomings of the current practices relating to conceptual construction duration estimation, the desire to develop inexpensive, quick and reasonably accurate estimation methodology and the limited research in the area of conceptual duration estimation are the main motivations of this study.

1.2 OBJECTIVES

A detailed estimate based on material quantities can not be made at the early stages of the construction project, because detailed design information and specifications are not available during the early phases. Alternative methodologies are required to estimate the duration of the project during these early phases.

The main objective of this study is to develop and compare reasonably accurate and practical methodologies for conceptual duration estimation of building projects. Based on the main objective, the following sub-objectives are identified:

- To develop a parametric model for conceptual cost estimation since cost estimates are also required in the assessment of prediction performances of the time-cost models.

- To test the time-cost model proposed by Bromilow (1974).

- To develop time-cost models (models where cost is used to estimate the duration of the projects).

- To develop parametric models for conceptual duration estimation.

- To compare all the models developed in terms of their predictive abilities.

1.3 SCOPE

The scope of this research is limited to the development and comparison of parametric estimating models for conceptual duration estimation of building construction projects. The data used in the model developments and validations are based on the historical project data collected from the 17 continuing care retirement community (CCRC) projects. All these 17 CCRC building projects are constructed by the same contractor in the United States. Regression analysis and artificial neural networks are used to develop the parametric duration estimation models by establishing the relationships between the building parameters and the building duration.

1.4 RESEARCH ORGANIZATION

This study is organized into five chapters:

- Chapter I - Introduction - This chapter includes the importance of the subject, applications in the current practice, brief overview of the conceptual estimates, objectives and scope of the study.

- Chapter II – Literature Review – This chapter presents the research background. It focuses on the factors affecting construction project duration, time-cost models and other parametric models for construction duration estimations.

- Chapter III – Methodology – This chapter presents the data used in this study. Brief information about the analysis techniques and parametric model development are also presented.

- Chapter IV – Data Analysis and Results – This chapter focuses on the data analysis performed in this study. The developments and the determination of prediction performances of the estimation models are clearly presented.

- Chapter V – Conclusions and Recommendations – A summary, conclusions and recommendations are discussed in this chapter.

CHAPTER 2

LITERATURE REVIEW

Construction duration is considered to be one of the most important criteria for construction project's success along with the cost and quality. Therefore, several studies have been made related to construction project duration. However, few studies focused on formulating construction duration predictive models at the early stages of the construction projects. Previous studies related to identification of important factors influencing construction duration are summarized in this literature review.

2.1 TIME-COST MODELS

Bromilow (1974) modeled the relationship between duration and cost of building projects in Australia, and formulated this relationship in the form:

$$T=KC^{B} \tag{2.1}$$

where;

T = duration of construction period from possession of site to practical completion measured in working days;

C = final cost in millions of Australian Dollar (AD) adjusted to a cost index;

K = a constant describing the general level of duration performance for a million of AD project; and

B = a constant describing how the duration performance is affected by project size as measured by value.

329 building projects which were constructed in Australia during the period between June 1964 and June 1967 were analyzed. K was found to have a value of 350 working days and B had a value of 0,30.

Bromilow et al. (1980) re-studied the relationship between the cost and duration of building projects in order to determine whether equation 2.1 holds or not. This time 408 building projects completed between 1970 and 1976 in Australia were analyzed. The results showed that the relationship established by equation 2.1 still holds.

Several authors used Bromilow's time-cost (BTC) model and validated the relationship with different sets of data (e.g. Chan, 2001). Some authors made comparisons with classifying the data into set of factors and analyzing them to see the effect of each factor on construction duration (e.g. Kaka and Price, 1991). Some of them tried to improve the model by introducing other factors to the model or by suggesting an alternative function for the proposed equation (e.g. Ng et al, 2001).

Albert P.C. Chan (1999) attempted to validate BTC model for the 110 building projects in Hong Kong. In order to validate the BTC model by regression analysis, equation 2.1 is rewritten in the natural logarithmic form as:

$$\ln T = \ln K + B \ln C \qquad [2.2]$$

Regression results and constants found are shown in Table 2.1.

Table 2.1 Regression results of BTC model (Chan, 1999)

	All Projects	Public Projects	Private Projects
ln K	2,182	2,221	2,078
K	152,082	166,257	119,569
B	0,292	0,281	0,337
R	0,922	0,954	0,854
R^2	0,850	0,911	0,729
Adjusted R^2	0,846	0,906	0,715

According to the regression analysis results shown in Table 2.1, relationship between time and cost can be expressed by the following equations for all building projects, for public projects and for private projects in Hong Kong:

All building projects:

$$T_{\text{all}} = 152C^{0.29} \quad [2.3]$$

Public projects:

$$T_{\text{public}} = 166C^{0.28} \quad [2.4]$$

Private projects:

$$T_{\text{private}} = 120C^{0.34} \quad [2.5]$$

Chan (1999) pointed out that equations 2.3, 2.4 and 2.5 can be used to predict construction durations of buildings by project managers and clients. On the other hand, Chan (1999) compared publicly funded projects with privately funded projects in terms of their construction duration performances. BTC model represented by equation 2.1 verifies T=K when C=1. Hence, K value gives the expected construction duration in days for a project of 1 million Hong Kong Dollars when C=1. Chan classified the data for publicly funded projects and privately funded projects, and compared them in terms of their expected construction durations for a project of 1 million Hong Kong Dollars. Privately funded projects had shorter completion time (120 days) then publicly funded projects (166 days). This can be explained by private sector's more concern on time criteria than public sector to start the business and to get returns on their investments as mentioned by Kaka and Price (1991).

Albert P.C. Chan (2001) similarly validated BTC model, but this time for 51 public building projects in Malaysia. Relationship explained by BTC model was rewritten as shown in equation 2.2, and the data was analyzed by using simple linear regression technique.

Table 2.2 Regression results of BTC model (Chan, 2001)

ln K	5,596
K	269,4
B	0,315
R	0,638
R^2	0,407
Adjusted R^2	0,395

According to the regression analysis results shown in Table 2.2, the time and cost relationship for public building projects in Malaysia may be represented by the following equation:

$$T = 269C^{0.32} \quad [2.6]$$

Chan (2001) suggested the equation 2.6 to be used as an alternative and objective method for estimating construction durations of public buildings in Malaysia.

Ng et al. (2001) tested the BTC model with a new set of data compiled from building projects completed between 1991 and 1998 in Australia. The data was partitioned into subgroups according to the client sector, the method of contractor selection, the type of project and the type of contractual arrangements. Equation 2.2 was given in the form of

$$y = \alpha_0 + \alpha_1 x \qquad [2.7]$$

letting y = ln T, x = ln C, α_0 = ln K and α_1= B to simplify the linear regression equation. Results obtained by the regression analysis of the partitioned data are shown in Table 2.3.

Table 2.3 Regression results (Ng, 2001) s: significant, ns: not significant

		# of project	α_1	R	R_2	R_2(adj)	α_0
All		*93*	*0,3105*	*0,767*	*0,588*	*0,583*	*ns*
Sector	Public	31	0,3276	0,8242	0,679	0,668	ns
	Private	62	0,3007	0,7346	0,540	0,532	ns
Contractor selection	Selective tender	59	0,2882	0,714	0,510	0,501	ns
	Open tender	15	0,3289	0,884	0,782	0,765	ns
	Negotiation	19	0,3879	0,895	0,800	0,788	ns
Project type	Recreational	9	0,2529	0,792	0,627	0,574	ns
	Industrial	26	0,3617	0,900	0,810	0,802	ns
	Educational	15	0,4238	0,702	0,493	0,454	ns
	Residential	11	0,1563	0,486	0,236	0,151	ns
	Other	32	0,3299	0,849	0,720	0,711	ns
Contract	Lump sum	61	0,3239	0,780	0,609	0,602	ns
	Design & construct	16	0,2108	0,676	0,457	0,418	s
	Construction management	8	0,3607	0,980	0,960	0,954	ns
	Other	8	0,4333	0,785	0,616	0,561	ns

This study revealed that client sector, contractor selection method and type of contractual arrangements do not affect construction duration significantly. The finding that the client sector does not affect construction duration significantly contradicted with the study of Chan (1999) made for buildings in Hong Kong. On the other hand, significant differences were found between project types (industrial projects vs. educational and residential projects). Two separate models, one for industrial projects and one for non-industrial projects were analyzed and it was shown that smaller industrial projects take less time to complete than the smaller educational and residential projects. Also, functions (i.e. C, C^2, $\sqrt{C}$, $1/C$) alternative to Bromilow's log-log model were tested as independent variables (no changes made for dependent variable $\ln T$); however, none of them found to be better than $\ln C$ as shown in Table 2.4.

Table 2.4 R^2 (adjusted) results (Ng, 2001)

Independent variable	R^2 (adjusted)
C	0,180
C^2	0,064
$\sqrt{C}$	0,375
$\ln C$	0,583
$\log_{10} C$	0,583
$1/C$	0,473

Ng et al. (2001) also compared the K constants observed in previous studies for buildings in Australia. Results of the study indicated improvement in the construction speed over the last 40 years.

Similarly, Ogunsemi and Jagboro (2006) tried to formulate a time prediction model in the form expressed by equation 2.1 for 87 building projects completed in Nigeria. In order to analyze the data by linear regression, equation 2.1 was rewritten in the form of equation 2.2. Summary of regression results for BTC model was shown in Table 2.5.

Table 2.5 Regression results of BTC model (Ogunsemi and Jagboro, 2006)

Parameters	All Projects	Public	Private
ln*K*	4,138	4,001	4,230
K	63	55	69
B	0,262	0,255	0,312
R	0,453	0,443	0,567
R^2	0,205	0,196	0,322
R^2 (adjusted)	0,193	0,177	0,293

According to results shown in Table 2.5, construction duration prediction models for all projects, public projects and private projects in Nigeria can be written as follows:

All projects:

$$T = 63C^{0.262} \tag{2.8}$$

Public projects:

$$T = 55C^{0.255} \tag{2.9}$$

Private projects:

$$T = 69C^{0.312} \tag{2.10}$$

However, Ogunsemi and Jagboro (2006) stated that equations 2.8, 2.9 and 2.10 can not be used, since coefficient of determination (R^2) values of each equation were very low. Hence, it can be concluded that the BTC model was not valid for Nigeria. So, Ogunsemi and Jagboro (2006) suggested a piecewise linear model with breakpoint (BPT) for the analysis of 87 Nigerian building projects' data. Piecewise linear model with breakpoint was expressed as a type of nonlinear model which is linearized by introducing a breakpoint between two linear models. As an example; a simple model consisting of two variables was given by

$$T = a_0 + a_1 C\ (C \leq \mathrm{BPT}) + a_2 C\ (C > \mathrm{BPT}) \qquad [2.11]$$

where ($C \leq$ BPT) and ($C >$ BPT) denote logical conditions that evaluate to 0 if false, and 1 if true. This implies that the model becomes

$$T = a_0 + a_1 C \quad \text{if } (C \leq \mathrm{BPT}) \qquad [2.12]$$

or

$$T = a_0 + a_2 C \quad \text{if } (C > \mathrm{BPT}) \qquad [2.13]$$

Analysis of data revealed that piecewise model with a breakpoint can be used to explain the relationship with a high predictive ability. R^2 values of 0.7656, 0.7762 and 0.8306 were found for all projects, public projects and private projects respectively. And the resulting models were expressed as follows:

For all projects:

$$T = 118.563 + 0.401C\ (C \leq 408) \text{ or } T = 603.427 + 0.610C\ (C > 408) \qquad [2.14]$$

For public projects:

$$T = 98.010 + 0.357C\ (C \leq 353) \text{ or } T = 567.967 + 0.283C\ (C > 353) \qquad [2.15]$$

For private projects:

$$T = 168.895 + 0.491C\ (C \leq 557) \text{ or } T = 709.66 + 0.884C\ (C > 557) \qquad [2.16]$$

Ogunsemi and Jagboro (2006) believed that use of these models will provide an alternative assessment for comparison of traditional construction duration estimation methods in Nigeria.

Although the time-cost model proposed by Bromilow was accepted as reliable and practical by many researchers, it fails to consider factors other than cost when establishing the construction duration (Walker, 1995). Reliability of using contract cost to predict the construction duration is another problem, since it may be considerably different than the actual cost which can only be known after the project is finished.

2.2 FACTORS AFFECTING CONSTRUCTION DURATION

The construction project duration is a function of many variables. The first problem that may be encountered during parametric model development is the identification and selection of variables that may be used in forecasting the construction durations for building projects. Then, the second problem is the determination of the relationships between the variables and the project duration.

Kaka and Price (1991) identified the factors affecting construction durations by making use of Bromilow’s time-cost relationship. Firstly, they investigated the validity of Bromilow’s time-cost relationship for 661 building projects and 140 road projects in United Kingdom. Similar empirical relationships were achieved for both types. Then, they classified the data according to type of client (public, private), type of project (building, civil engineering), type of tender (open competition, selected competition, negotiated competition) and form of tender (fixed price tender, fixed adjusted tender). Regression analysis was performed to see the effects of these factors on the duration. It was concluded that the form of tender, type of client and type of project influenced the duration while the type of tender has no effect on the duration. Construction durations of projects with adjusted price contracts generally took longer than projects with fixed price contracts. This is reasonable since adjusted price contracts were usually chosen for projects which are expected to take relatively long time to eliminate the inflation risk. On the other hand, construction durations of public buildings were shown to be longer than that of private buildings. This can be explained by the implications of private clients related to construction time in order to get returns on their investments as soon as possible. It was also shown that construction durations of building projects took generally longer than that for civil engineering projects of similar value. The logic behind this is the involvement of high transportation costs of heavy plant and equipment in civil engineering projects compared to building projects.

Kumaraswamy and Chan (1995) illustrated the factors affecting construction duration (See Figure 1). They stated that illustrated factors are based on the general international literature, observed common construction practice, and the survey results. These factors include both qualitative and quantitative contributors. The construction duration can be regarded as a function of all these factors.

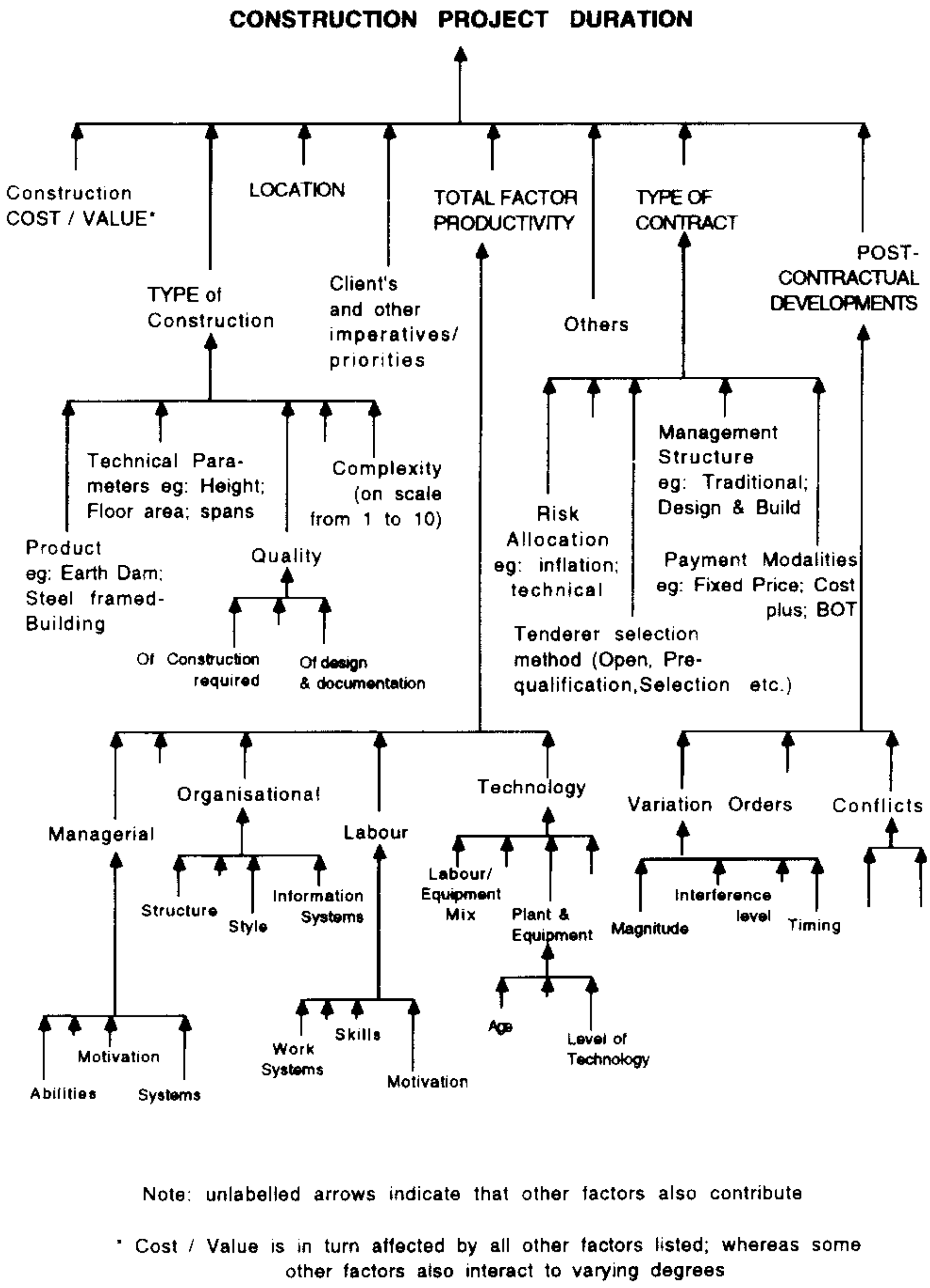

Figure 2.1 Factors affecting construction project duration (Kumaraswamy and Chan, 1995)

According to Kumaraswamy and Chan (1995), reliability of construction duration estimates depends on the skill and experience of the planning engineer. They proposed to formulate and test empirical time-cost models, time-floor area models, and time-number of floors models in order to minimize the subjective effect of planning engineer on construction duration estimates. They argued that, the duration of construction duration can be predicted by putting the significant characteristic variables into the proposed models. Firstly, they chose to test Bromilow's time-cost model as a first approximation for predicting duration, since the BTC model includes only cost out of the factors listed in Figure 1. BTC model was found significant within the each category of 111 projects' data completed in Hong Kong. Then, they proposed that gross floor area (GFA) can be replaced with the cost (C) in BTC model which is explained by equation 2.1. GFA was hypothesized to be the principal contributor to the building cost. The new model which is described as:

$$T=LA^{M} \tag{2.17}$$

where;

A = GFA in m^2

L, M = The coefficients corresponding to the constants K and B in the equation 2.1

was found to be more fundamental and GFA was confirmed to be a significant quantitative factor affecting construction duration.

In the second phase of the study Kumaraswamy and Chan (1995) investigated the relationship between the duration and the number of floors of a building. This time the following model was proposed:

$$T=FS^{G} \tag{2.18}$$

where;

S = the number of floors of a building with one single block only

F, G = the coefficients corresponding to the constants K and B in the equation 2.1.

Number of floors was found to be useful, however, it was confirmed that number of floors was not as significant as cost and GFA for construction duration. Therefore, the model

$$T=KC^{B}A^{M} \quad [2.19]$$

was derived to develop an equation relating the construction cost and GFA of a building with the construction duration by multiple regression method. The model represented by equation 2.19 was tested and confirmed to be significant. The authors claimed that use of all three models explained by the equations 2.17, 2.18 and 2.19 is possible. However, use of model including both the construction cost and the floor area (eqn. 2.19) appeared to be better when the variables are known at the feasibility stage. In addition to the three project macro variables considered (construction cost, GFA and number of floors), Kumaraswamy and Chan (1995) found out that the micro factors that affect productivity (plant utilization level and the efficiency of site laborers) also influences the construction duration significantly. The investigation through a case study on construction site of a new building demonstrated that productivity is a significant parameter influencing the construction duration. "The case study included field investigations into

(1) Plant utilization levels such as tower cranes and truckmixers;

(2) A comparison of the average productivity of different concrete placing methods such as pump and crane and skip;

(3) The activity analysis profiles of construction workers such as formwork riggers, steel bar benders, steel-fixers and concreters on site".

The field study revealed that such factors relating to productivity affect construction durations significantly. Kumaraswamy and Chan (1995) ended up their study noting that factors such as project complexity, quality level required, management style, overall organizational structure of project team, communications between parties and type of contract should also be taken into account, since they also influence the construction durations.

2.3 OTHER PARAMETRIC MODELS

Some authors suggested new approaches to identify factors influencing duration and to formulate construction duration predictive models alternative to time-cost models reviewed previously. Multiple linear regression (MLR) technique was applied generally (e.g. Nkado, 1992; Kumaraswamy and Chan, 1999). Bhokha and Ogunlana (1999) used artificial neural network (ANN) to forecast construction duration of buildings.

Nkado (1992) has shown that by using multiple linear regression analysis, durations of activity groups including substructure, superstructure, cladding, finishes, services and their sequential start-start lag times can be predicted from 12 variables:

(1) End use (function) of project (office, retail, other).

(2) Type of structural frame (concrete, steel, other).

(3) Location (London, elsewhere).

(4) Accessibility to site (poor, not poor).

(5) Type of cladding (prefabricated panels, curtain wall, brick).

(6) Is atrium featured? (yes, no)

(7) Intensity of services (low, medium, heavy)

(8) Number of floors excluding basement floor.

(9) Height from ground to eaves levels (m).

(10) Area of ground floor (m^2).

(11) Gross floor area (GFA) (m^2).

(12) Approximate volume of excavation (m^3).

Work content of activity groups can be summarized as follows:

- *Substructure*: All activities necessary to complete the ground works up to and including the ground floor slab, foundations, underslab drainage, basement, etc.

- *Superstructure*: All activities necessary to erect the load-bearing frame, including structural roof members.

- *Cladding*: All activities necessary to render the building watertight and weathertight, including external walls, roofing, windows and external doors.

- *Finishes*: All activities necessary to decorate the works including internal non-loadbearing partitions.

- *Services*: All activities necessary to erect the mechanical and electrical work including plumbing.

The data of 29 privately funded commercial building projects were used in the study. The results of regression analysis for the five activity groups and four sequential lag times were summarized in Table 2.6.

Table 2.6 Summary of regression analysis (Nkado, 1992)

Work package	R^2	R^2 (Adj.)	Mean	Standard error
SUBSTRUCTURE	0,92	0,9	23,8	3,4
SUPERSRTUCTURE	0,79	0,77	25,7	8,1
CLADDING	0,87	0,83	28,5	5,1
FINISHES	0,77	0,67	39,5	12,2
SERVICES	0,86	0,81	38,7	8,5
Lag times				
B	0,74	0,72	11,6	4,8
C	0,77	0,7	8,5	4,1
D	0,84	0,78	4,8	4,8
E	0,87	0,84	2,0	2,4

Out of 45 variables which were entered into the regression equations, 29 variables appeared in the accepted equations. Based on the frequencies of occurrence in the regression analysis, the eight most prominent and possibly most important variables for estimating construction times were the gross floor area, height, type of cladding, number of floors, location, predominant frame (steel or concrete), storey-height and approximate volume of building. Accuracy of the model was tested by comparing the model's predictions with estimates produced by nine construction planners for three office building projects which were not used in developing the model. Nkado (1992) suggested the model can be used for estimating construction durations and producing outline construction plan of buildings in the early design stages, as the models provided reasonably accurate results.

After the two phased study (1995) which aimed to find the significant factors contributing to construction duration, Kumaraswamy and Chan (1999) developed construction duration prediction model in the fourth phase of their study. The data for 56 standard 'Harmony' type domestic blocks in Hong Kong were used for analysis. 'Harmony' type block is an average quality public housing block ranging from 30 to 40 storeys and containing about 16 residential units on each floor. Kumaraswamy and Chan (1999) separated the primary work packages in the building process. The primary work packages were classified as site set-up, piling, pile caps/raft, superstructure, electrical & mechanical services, finishes and external works, similarly as they were classified in a study conducted by Nkado (1992). Pile caps/raft was described as the "activities necessary to construct either the pile caps in the case of pile foundations or the raft foundation, including the ground floor slab". Although site set-up and external works were classified as primary work packages, durations for them were not included in the analysis since they were not critical in determining the overall construction duration. 94 variables entered into the regression model. 84 of them were identified through questionnaire survey, and the remaining 10 were derived by authors to be used in regression analysis. The results of regression analysis for the durations of work packages, their sequential lag times, and the estimated overall construction duration stipulated in the contract were summarized in Table 2.7.

Table 2.7 Summary of regression analysis results (Kumaraswamy and Chan, 1999)

Work package	R^2	R^2 (Adj.)	Mean (months)	Standard deviation (months)
PILING	0,8935	0,8612	4,97	0,93
PILE CAPS/RAFT	0,7576	0,7155	3,83	0,87
SUPERSTRUCTURE	0,8453	0,8354	17,40	2,65
SERVICES	0,7990	0,7522	23,63	3,44
FINISHES	0,8038	0,7788	18,93	2,82
LAG2	0,9043	0,8892	4,76	0,88
LAG3	0,8936	0,8777	3,78	0,96
LAG4	0,7571	0,7307	1,64	2,01
LAG5	0,8652	0,8479	4,36	1,68
ESTIMATED TIME	0,8731	0,8572	33,21	2,93

According to frequencies of occurrence in the regression models, the seven most significant variables in forecasting construction duration were found as:

(1) Area of external cladding

(2) Height of building

(3) Ratio of total GFA to the number of storeys

(4) Type of foundations

(5) Information flows between architect/engineer and contractor

(6) Presence/absence of precast façades

(7) Type of scheme (rental/purchase)

Reliability of the model was tested for the data of nine similar public housing projects which were not included in the derivation of the model, since they were still under construction. Although high percentage error values were obtained for durations of 5 work packages and 4 lag times, overall durations were predicted with ±7% accuracy compared to planners' estimates.

Then, Albert P.C. Chan and Daniel W.M. Chan (2004) focused on a prediction model for actual overall construction duration (direct construction duration prediction, i.e. work packages and their lag times was not taken into account) by applying MLR analysis with the same data of 56 standard 'Harmony' type residential buildings in Hong Kong. A stepwise selection method with a significance level of 5% was used to select statistically significant variables to be incorporated into the model. Data variables were added and deleted one at a time and, the regression model was re-run, noting at each step the changes in the coefficient of determination (R^2) value and in essence the significance level of variables. Only those variables with a significance level (p-value) of less than 5% were retained for inclusion in the final regression model equations. The final equation for actual overall construction duration of 56 standard 'Harmony' type public building with a coefficient of variation (R^2) value 0,7769 was found to be:

$\log_e$ *ACT-TIME* = 3,0264 + 0,1236 log *ACT-COST* + *TYPESCH* (- 0.0544 for purchase; 0 for rental) + *FAÇADE* (0 for with facades; 0,0666 for without facades) + 1,3E-06 *VOLTOTAL* – 0,0003 *GFA/NOSTOREY* [2.20]

with five critical variables which were:

(1) actual total construction cost in HK$M

(2) type of housing scheme (rental or purchase)

(3) presence or absence of precast facades

(4) total volume of building in m^3

(5) the ratio of GFA in m^2 to number of floors.

The validity of the model was tested by comparing the actual values with predicted values. Mean absolute percentage error (MAPE) method was used to test the reliability of the model. MAPE was defined as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|(\text{predicted duration})i - (\text{actual duration})i\right|}{\left|(\text{actual duration})i\right|}\times 100 \qquad [2.21]$$

The MAPE value of 3.97% confirmed the reliability of the model. Thus, Chan and Chan (2004) suggested the model for predicting the overall construction durations of similar projects.

More recently, Love et al. (2005) proposed a new model using multiple regression technique of weighted least squares for 126 projects including building projects, factories, warehouses, and airports completed in Australia. Four different project types were used in this study. These were:

(1) New build

(2) Refurbishment/renovation

(3) Fit out

(4) New build/refurbishment

The new model which generated to forecast construction duration of projects was in the following form:

$$\log(T) = 3.178 + 0.274 \log(\text{GFA}) + 0.142 \log(\text{Floor}) \quad [2.22]$$

GFA and number of floors were suggested as the factors influencing construction duration. Love et al. (2006) stated that “In practice, the GFA and number of floor levels are known before a project commences. Therefore, project managers, cost planners, estimators, and the like can use this model to predict project time before a project commences so that they can better plan their projects and so that clients can arrange their funding requirements”. On the other hand, cost was indicated to be poor predictor for construction duration since it could not be known before the project is completed. However, Love et al. (2006) suggested use of traditional BTC model as well for comparison with equation 2.22’s prediction to obtain a more reasonable judgment.

Apart from the studies which were mainly based on regression techniques, Bhokha and Ogunlana (1999) proposed use of ANN to forecast the construction duration of buildings at the predesign stage. In a study conducted by Bhokha and Ogunlana (1999), buildings which were higher than 23 m and with a functional area not less than 10000 m^2 were taken into consideration. The data of 136 building projects which were constructed in Greater Bangkok were used. The samples were divided into two parts for training and testing. Eleven independent variables were used as inputs. The eleven inputs which were selected to be used in ANN model were:

(1) building function (two nodes)

(2) structural system (two nodes)

(3) functional area (one node)

(4) height index (one node)

(5) complexity of foundation works (one node)

(6) exterior finishing (two nodes)

(7) decorating quality (one node)

(8) site accessibility (one node).

Out of these 11 variables, only the functional area was a real-value variable. The remaining ten were binary. For example, the two nodes for building function could have the values (1, 0) for residential construction; (0, 1) for an office; (1, 1) for a building with dual functions; and (0, 0) for other buildings. Variables and building features were shown in Table 2.8.

Although it was stated by the authors that there was no exact rule for determining the optimum number of hidden nodes, six hidden nodes were found to be the most satisfactory number for the network by trial-and-error. The output variable of the neural network was construction duration as intended. Validation of the model using the 68 test samples verified that results obtained by the final network for duration forecasting are satisfactory with an average error of 18,2%. On the other hand, average error for the 136 projects was 13,6%.

Table 2.8 Inputs and building features (Bhokha and Ogunlana, 1999)

Description	Input node	Building feature
Building function	1 2	Residence only Office only Dual (residence+office) Others
Structural system	3 4	Cast-in-place RC RC frame + PC slab Others
Functional area ($\times 10^6$ m^2)	5	
Height index	6	# of floor >25 # of floor ≤25
Foundation index	7	Complex Simple
Exterior finishing (walls)	8 9	Brick/cement block Curtain wall/glass Others
Decorating quality	10	Excellent Normal
Accessibility to site	11	Difficult Easy

Literature review revealed that although several studies have focused on determination of construction durations for building projects, only a few of them concentrated on the estimation of construction durations at the early design stages of the project. The time-cost model proposed by Bromilow (1974) formed a basis for many studies conducted in different parts of world. It enabled many authors to identify important factors affecting construction duration and to estimate construction duration from cost. The model may be practical and easy, since it only uses contract final sum in determining the duration. However, reliable estimation of cost is required if this model is to be used at the predesign stage. The other studies conducted for estimation of construction durations in the early stages made use of MLR and ANN as summarized. Understanding of these techniques contributed to the development of the models and comparison of results of this study.

CHAPTER 3

METHODOLOGY

3.1 DESCRIPTION OF THE DATA

Development of parametric duration estimation models depends on the availability of the historical data. Data of building characteristics and the associated durations must be collected and analyzed to identify the significant variables influencing construction duration and to establish the relationship for parametric model development.

The historical project data used in this study consists of 17 building projects constructed by a contractor in the United States. The project data is collected from continuing care retirement community (CCRC) projects built in 14 different states from 1975 to 1995. A CCRC can be defined as an establishment which provides housing and health care services to people of retirement age. A CCRC generally includes residential, health center and commons buildings. Some CCRCs may also have structured parking (Sonmez, 2004).

In this study, adjusted cost data is used. Variability in the project cost due to the location differences and the time differences are quantified with city cost index and historical cost index, respectively (Waier et al. 1996). Reliability of using these detailed cost data is ensured since the detailed cost data used in the contracts were very close to the actual costs. On the other hand, data of actual durations are used in this study. Table 3.1 lists the projects and the actual durations of the projects.

Table 3.1 Projects and durations

NO	PROJECT NAME	DURATION (in months)
1	Project 1	21
2	Project 2	12
3	Project 3	16
4	Project 4	14
5	Project 5	13
6	Project 6	16
7	Project 7	14
8	Project 8	16
9	Project 9	18
10	Project 10	10
11	Project 11	18
12	Project 12	20
13	Project 13	17
14	Project 14	17
15	Project 15	16
16	Project 16	20
17	Project 17	14

The other project characteristics data that will be used for data analysis are as follows:

1) Total building area (Area)

2) Number of floors (NoF)

3) Area per unit (Area/unit)

4) Combined percent area of commons and health center (Per(C+H))

5) Percent area of structured parking (Per(P))

6) Type of structural frame of the building (steel (St), masonry (Mas), reinforced concrete (RC), precast (Pre), wood (W))

Total building area is obtained by adding the building areas of residential, health center, commons and structured parking facilities. The area per unit is calculated by dividing the total building area by the total numbers of all residential units including studios, one-bedroom, two-bedroom and three-bedroom apartment units. The combined percent area of commons and health center is determined by dividing the sum of commons and health center area by the total building area. Similarly, percent area of structured parking is determined by dividing the structured parking area by the total building area. The values of the variables are tabulated in Table 3.2. For the types of structural frames, the "1" means that type of structural method is used. All these variables that will be used in developing the parametric estimation models and their abbreviations are shown in Table 3.3.

Table 3.2 Data of independent variables affecting construction cost and duration of building projects

NO	Area (in m^2)	NoF	Area/Unit	Per(C+H)	Per(P)	St	Mas	RC	Pre	W
1	29357	5	131	0,161	0,000	1	1	1	1	0
2	20954	4	105	0,163	0,000	0	0	0	1	0
3	16812	4	124	0,234	0,000	0	1	0	1	0
4	17655	2	111	0,219	0,035	0	0	0	0	1
5	13930	3	97	0,104	0,000	0	0	0	0	1
6	16249	3	107	0,198	0,000	0	0	0	0	1
7	25373	3	157	0,220	0,000	1	0	0	0	1
8	11100	3	137	0,118	0,000	0	1	0	0	1
9	22927	3	156	0,184	0,066	0	0	0	0	1
10	6735	3	160	0,132	0,046	0	0	0	0	1
11	39495	6	258	0,158	0,143	0	0	1	0	0
12	30584	6	173	0,175	0,042	1	0	1	0	0
13	25961	3	177	0,180	0,119	1	0	1	1	1
14	24380	4	181	0,180	0,120	1	0	1	0	1
15	19826	2	134	0,140	0,000	1	1	0	0	1
16	30673	3	142	0,162	0,049	1	1	0	1	1
17	27937	2,5	168	0,176	0,000	1	1	1	0	1

Table 3.3 Variables and their abbreviations

Variable	Abbreviation
Total building area	Area
Number of floors	NoF
Area per unit	Area/Unit
Combined percent area of commons and health center	Per(C+H)
Percent area of structured parking	Per(P)
Steel	St
Masonry	Mas
Reinforced Concrete	RC
Precast	Pre
Wood	W

3.2 DESCRIPTION OF ANALYSIS TECHNIQUES

In this study, the techniques of regression analysis and artificial neural networks are used to develop parametric estimation models. Regression analyses are performed utilizing the Microsoft Office Excel 2003. Artificial neural network models are developed by using BrainMaker Professional 3.10.

3.2.1 Regression Analysis

Regression Analysis is a statistical tool for investigating the functional relationship between a dependent variable and one or more independent variables. Different types of regression models can be developed including a simple linear, a multiple linear or a nonlinear polynomial regression. The basic formulation for multiple regression analysis is:

$$Y = \alpha_0 + \alpha_1 X_1 + \ldots\ldots + \alpha_n X_n \qquad [3.1]$$

where:

Y = dependent variable

α_0 = constant

α_1 = partial regression coefficient for X_1

X_1 = independent variable 1

α_n = partial regression coefficient for X_n

X_n = independent variable n

n = number of variable

The partial regression coefficient (α) measures the amount of change in the dependent variable for one unit change in the independent variable. All the partial regression coefficient values is determined by the method of least squares which ensures that the selection of the coefficients in the resulting model produces the smallest sum of squared differences between the actual and modeled values of the dependent variable.

The overall objectives of regression analysis can be summarized as follows:

1) To determine whether a relationship exists between the variables or not

2) To describe the relationship in terms of a mathematical equation

3) To evaluate the accuracy of prediction achieved by the regression equation

4) To evaluate the relative importance of independent variables in terms of their contribution to variation in the dependent variable

3.2.2 Artificial Neural Networks

Artificial neural network (ANN) is a branch of artificial intelligence in which the main structure is based on the biological nervous system. It can exhibit a surprising number of the human brain's characteristics. For example, it can learn from experience and generalize from previous examples to new problems. ANN can provide meaningful answers even when the data to be processed include errors or are incomplete, and can process information extremely rapidly when applied to solve real world problems (Smith, 1993).

ANN consists of many computational elements called nodes. The nodes are arranged in layers. A typical network arrangement consists of an input layer, an output layer and a hidden layer or a number of hidden layers. An ANN with three layers is shown in Figure 3.1.

ANN can be prepared by following five basic steps:

1) Defining the problem by deciding what information to use and what the network will do

2) Architecture determination

3) Learning process determination

4) Training the network

5) Testing the trained network

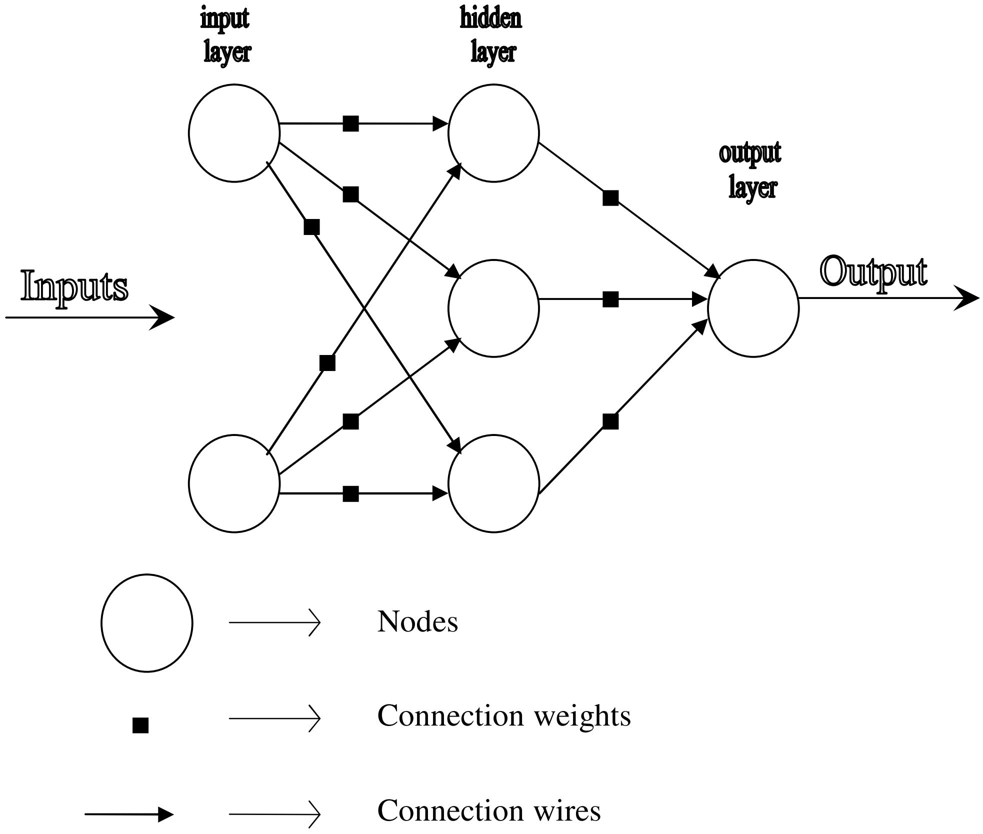

Figure 3.1 An artificial neural network with three layers

Input nodes receive the values of the input variables which then propagate through the network layer by layer. During training, the input layer broadcast a pattern to all the hidden nodes. The system is then asked to calculate an output value in a feed-forward way. The hidden nodes broadcast their results to all output nodes. Each output node then calculates a weighted sum and passes it to the output node to generate a result. The result is compared with the target value, which the trainer has established at the beginning of a training session. The difference yields the system output error. At this stage the system has to decide whether further learning is required or not. This is achieved by comparing the obtained total difference with a specified acceptable error given by the system developer. If the decision is to continue, the output nodes calculate the derivatives of the error with respect to the weights and the result is sent back through the system to all the hidden nodes. Each hidden node calculates the weighted sum of the error. Then, each hidden-layer node and output-layer node change their weights to compensate for the corrections. Once the weights have been changed, the feed-forward computation starts again. New output values are obtained and the cycle continues until a desired result is obtained. At this stage training of the system is complete and the testing phase can start. The system can be used to predict the outcome of an input not previously seen by the ANN (Boussabaine, 1996).

3.2.3 BrainMaker Professional Basics

BrainMaker Professional 3.10 is used to develop neural network models. BrainMaker reads three kinds of neural network files. These are definition files, fact files and network files.

A definition file describes everything about the network to BrainMaker, such as the number of neurons (nodes) in each layer and the type of data being used. BrainMaker uses the definition file to create the neural network. The default extension for the definition file is .Def.

A fact file gets the data into BrainMaker. There are fact files for training, testing and running. The default extension for the training fact file is .Fct, for testing it is .Tst, and for running it is .In.

A network file is created by BrainMaker during training using the data in the training fact file and the instructions in the definition file. The network file contains the actual connection information as well as training parameter information. The default extension for a network file is .Net.

NetMaker, which is included in the BrainMaker Professional 3.10 software, can create the BrainMaker files. However, NetMaker is not a word processor or spreadsheet. So, data file should be prepared with another program such as Lotus, Excel, or an ASCII editor. Then, NetMaker can read the data file and create BrainMaker files.

The data should be arranged as rows and columns of numbers or symbols. The first row should be words representing the column headings. Each column is a separate category of input and output information. Columns for reference use, such as the date of the fact or the case number of the fact, can also be included. Reference columns are labeled annotations in NetMaker. Annotations may be displayed in BrainMaker along with each fact, and written to output files. They are very useful in keeping track of the data during all phases of design. Each row in the data file should contain one fact and every fact should contain a value in every column. There is no such thing as a null value in neural networks (BrainMaker Professional User's Guide and Referance Manual, 1993).

3.3 DEVELOPMENT OF PARAMETRIC MODELS

Parametric method in conceptual duration estimation involves the identification of the significant building parameters and development of the parametric model. Parametric model defines a dependent variable as a function of one or more independent variables. It can be used to understand the relative importance of the parameters included in the model. Since the model is developed with relevant historical project data, the success of the model depends on the ability of the model to establish appropriate relationship between the independent and dependent variables. In this way, the parametric model developed can also be used to estimate the duration of a future project.

In this study, five duration estimation models and a cost model will be developed. Three of models will be developed with the historical cost, and duration data shown in Table 3.1. The other two duration estimation models will be developed by analyzing the duration data and data of other project characteristics shown in Table 3.2. Since the construction costs of building projects can not be known at the early stages, duration estimation models developed with the historical cost data can not be used for estimating the duration of a new project. Therefore, a parametric cost estimation model will be developed with the available data. Cost estimates obtained from this model will also be used in determination of the prediction performances of time-cost models. The six models developed, the data used and the analysis technique utilized in the development of these models are summarized in Table 3.4 as follows:

Table 3.4 Models Developed

MODEL NAME	HISTORICAL DATA	ANALYSIS TECHNIQUE
COST MODEL	cost data – project characteristics data (Table 3.2)	Multiple Linear Regression
MODEL 1	duration data (Table 3.1) - cost data	Simple Linear Regression
MODEL 2	duration data (Table 3.1) - cost data	Simple Linear Regression
MODEL 3	duration data (Table 3.1) - cost data	Artificial Neural Network
MODEL 4	duration data (Table 3.1) - project characteristics data (Table 3.2)	Multiple Linear Regression
MODEL 5	duration data (Table 3.1) - project characteristics data (Table 3.2)	Artificial Neural Network

While “model 1”, “model 2” will be developed by simple linear regression analysis, “model 3” will be developed by artificial neural network technique with only cost parameter having been used to estimate the construction duration of building projects. For “cost model”, “model 4” and “model 5” which have multiple variables, parsimonious models will be developed to determine the best relationship between the dependent variable and the most significant variables. A parsimonious model can be defined as a model that fits the data adequately without using any unnecessary parameters (Sonmez, 2004). All of the independent variables will be considered at first and a backward elimination procedure will be used to select and eliminate the insignificant variables one at a time. In regression models (cost model and model 4), elimination of insignificant variables will be established by using the significance level (P-value) and coefficient of determination (R^2). The P-value indicates the significance of the variables included in the model. The R^2 indicates how much variation in the dependent variable is explained by a group of independent variables. In neural network models, insignificant variables will be eliminated by sensitivity analysis. This analysis determines the sensitivities of each input parameter with output results observed by varying each parameter incrementally while holding the values of remaining parameters constant.

3.4 PREDICTION PERFORMANCE TEST

Prediction performance test will be conducted by calculating the mean absolute percentage error (MAPE) method and prediction performances will be compared for “cost model”, “model 1”, model 2”, “model 3”, “model 4” and “model 5”. Cross-validation technique will be used to assess the prediction performances of the models. The procedure can be summarized as follows:

1) The data of 17 CCRC projects will be divided into five groups. The first three groups will consist of three project data and the remaining two groups will consist of four project data. The projects in the groups will be

selected randomly but they will not be changed in the prediction performance calculations of different models for consistency.

2) Data of first group including three projects will be selected as the test sample and the regression or neural network model will be developed again with the remaining data.

3) The new model will be used to calculate the predictions for the test sample which are not used in developing the model. For time-cost models, predicted cost values of test sample observed from the cost model will be used to predict durations instead of detailed cost values.

4) Steps 2 and 3 will be repeated for the data of second, third, fourth and fifth group as well.

5) MAPE will be calculated after predictions for all 17 projects are completed.

Test groups explained in "step 1" and the projects involved in these groups are shown in Table 3.5.

Table 3.5 Test groups and projects involved

Test Groups	Projects
Group 1	Project 5 Project 12 Project 15
Group 2	Project 6 Project 8 Project 11
Group 3	Project 14 Project 16 Project 17
Group 4	Project 1 Project 3 Project 4 Project 10
Group 5	Project 2 Project 7 Project 9 Project 13

CHAPTER 4

DATA ANALYSIS AND RESULTS

4.1 COST MODEL

As it has been stated previously, construction cost of a project can not be known unless the detailed estimation with the quantities is made. However, detailed cost estimation may not be possible since the detailed design information and specifications may not available at the early stages of the project. On the other hand, a quick and reasonably accurate cost estimation model is required in order to estimate the construction cost of a new project and use this estimated cost to estimate construction duration of the project by using the time-cost models developed in this study. Also, use of estimated costs in the determination of prediction performances of time-cost models will be proper instead of use of detailed costs available, since the duration estimations will be employed with cost estimates obtained from this cost model. In view of the fact that developing duration estimation models of building projects at the early stages of project is the main objective of this study, only this cost estimation model is developed in this study.

4.1.1 Data Analysis

The technique of multiple regression analysis is employed to determine the mathematical function that establishes the relationship between the building parameters shown in Table 3.2 and the cost of the building. To measure the amount of change in cost for one unit change in the type of structural frame, the

“1” values representing the structural frame type of the project is multiplied by the total area of the project as shown in Table 4.1.

The form of the first regression model (RM1) including all the variables shown in Table 4.1 was as follows:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \alpha_7 X_7 + \alpha_8 X_8 + \alpha_9 X_9 + \alpha_{10} X_{10} \quad [4.1]$$

where Y = detailed cost (C)

α_0 = regression constant

$\alpha_{1,2,3,...10}$ = partial regression coefficient of $X_{1,2,3,...,10}$

X_1 = Area

X_2 = NoF

X_3 = Area/unit

X_4 = Per(C+H)

X_5 = Per(P)

X_6 = St multiplied by the area

X_7 = Mas multiplied by the area

X_8 = RC multiplied by the area

X_9 = Pre multiplied by the area

X_{10} = W multiplied by the area

As explained in part 3.3, elimination of insignificant variables in regression models is established by using the significance level (P-value) and coefficient of determination (R^2) to achieve parsimonious model. The highest P-value 0,980 for the regression coefficient corresponding to the Per(C+H) indicated that inclusion of Per(C+H) in the RM1 may not be required since it probably does not have a significant contribution to the RM1. Therefore, it is dropped from the regression model.

Table 4.1 Data of independent variables with type of structural frames multiplied by area

NO	Area (in m^2)	NoF	Area/Unit	Per(C+H)	Per(P)	St	Mas	RC	Pre	W
1	29357	5	131	0,161	0,000	29357	29357	29357	29357	0
2	20954	4	105	0,163	0,000	0	0	0	20954	0
3	16812	4	124	0,234	0,000	0	16812	0	16812	0
4	17655	2	111	0,219	0,035	0	0	0	0	17655
5	13930	3	97	0,104	0,000	0	0	0	0	13930
6	16249	3	107	0,198	0,000	0	0	0	0	16249
7	25373	3	157	0,220	0,000	25373	0	0	0	25373
8	11100	3	137	0,118	0,000	0	11100	0	0	11100
9	22927	3	156	0,184	0,066	0	0	0	0	22927
10	6735	3	160	0,132	0,046	0	0	0	0	6735
11	39495	6	258	0,158	0,143	0	0	39495	0	0
12	30584	6	173	0,175	0,042	30584	0	30584	0	0
13	25961	3	177	0,180	0,119	25961	0	25961	25961	25961
14	24380	4	181	0,180	0,120	24380	0	24380	0	24380
15	19826	2	134	0,140	0,000	19826	19826	0	0	19826
16	30673	3	142	0,162	0,049	30673	30673	0	30673	30673
17	27937	2,5	168	0,176	0,000	27937	27937	27937	0	27937

Since Per(C+H) which has the highest P-value is dropped from the RM1, the new regression model (RM2) included nine independent variables. Then, analysis of RM2 revealed that coefficient for the Area/unit has the highest P-value of 0,913. Therefore, the variable Area/unit is dropped from the RM2. The application of this procedure has been continued until all the partial coefficients corresponding to variables with significance level of higher than 10% are dropped from the regression models, one by one at each time. Summary of the regression results including R^2 values and the highest P-values for partial coefficients together with the corresponding variables for each model are shown in Table 4.2. It is noted that, decrease in the values of R^2 was very small as variables dropped. The R^2 value of 0,973 for the final regression model indicated that the majority of variability in cost is explained by the remaining five independent variables.

Table 4.2 Summary of regression results for cost models

Model	Independent variables in the regression equation	R^2	Variable corresponding to the coefficient with highest P-value	P-value of the coefficient
RM1	Area, NoF, Area/unit, Per(C+H), Per(P), St, Mas, RC, Pre, W	0,981	Per(C+H)	0,980
RM2	Area, NoF, Area/unit, Per(P), St, Mas, RC, Pre, W	0,981	Area/unit	0,913
RM3	Area, NoF, Per(P),St, Mas, RC, Pre, W	0,981	Per(P)	0,785
RM4	Area, NoF, St, Mas, RC, Pre, W	0,981	Mas	0,177
RM5	Area, NoF, St, RC, Pre, W	0,977	W	0,235
RM6	Area, NoF, St, RC, Pre	0,973	Pre	0,085

Since the significance level of the partial regression coefficient of variables in the equation has been selected to be 10%, only those variables with a P-value of less than 10% are retained in the final regression model equation (RM6). The independent variables retained in RM6, partial coefficients and the P-values of the coefficients corresponding to these variables are demonstrated in Table 4.3.

Table 4.3 P-values for regression model RM6

Independent variable	Partial coefficients	P-value of the coefficient
Area	821	0,000
NoF	921565	0,075
St	166	0,001
RC	-112	0,030
Pre	-66	0,085

Regression constant for RM6 is found to be -3014070. Then, the final regression equation from which the construction cost can be estimated is written as follows:

$$C = -3014070 + 821\times Area + 921565\times NoF + 166\times St - 112\times RC - 66\times Pre \qquad [4.2]$$

4.1.2 Prediction Performance Test

The validity of the final model is assessed in terms of predictive accuracy. That is, the predicted cost values are compared with the detailed cost values to verify the predictive efficiency of the cost model. Mean absolute percentage error (MAPE) is used to assess the forecasting performance of the model. Percentage Error (PE) and MAPE for the final cost model are defined as follows:

$$\text{PE} = \frac{\text{predicted cost} - \text{detailed cost}}{\text{detailed cost}} \times 100 \qquad [4.3]$$

$$\text{MAPE} = \sum_{i=1}^{17} \frac{\left|\text{PE}_i\right|}{17} \qquad [4.4]$$

Steps described in Part 3.4 are followed. The PEs and MAPE are shown in Table 4.4.

Table 4.4 PEs and MAPE for cost model

Project No	PE	$\lvert PE \rvert$
1	8,6	8,6
2	12,8	12,8
3	0,8	0,8
4	12,0	12,0
5	8,9	8,9
6	-15,1	15,1
7	19,4	19,4
8	-18,4	18,4
9	-2,2	2,2
10	6,7	6,7
11	3,4	3,4
12	-11,1	11,1
13	-10,1	10,1
14	5,7	5,7
15	1,4	1,4
16	-14,7	14,7
17	-0,3	0,3
	MAPE	**8,9**

4.2 MODEL 1

Literature review revealed that the time-cost model proposed by Bromilow (1974) is widely used by many researchers with different sets of data. In this part of the study, the Bromilow's time-cost (BTC) model will be used as a basis to develop a formulation to estimate construction duration of a project from cost and to verify whether such a relationship holds true or not for the data of 17 CCRC building projects. Formulations discussed in the literature review will be written in this part again.

4.2.1 Data Analysis

The equation proposed by Bromilow was

$$T=KC^{B} \qquad [4.5]$$

where;

T = duration of construction period from possession of site to practical completion measured in working days;

C = final project value in millions of AD adjusted to a cost index;

K = a constant describing the general level of duration performance for a million of AD project; and

B = a constant describing how the duration performance is affected by project size as measured by value.

A simple linear regression technique is used to analyze the data. The nonlinear model represented by equation 4.5 is rewritten in the natural logarithmic form for statistical verification of the time cost relationship as:

$$\ln(T) = \ln(K) + B\ln(C) \tag{4.6}$$

By letting $Y = \ln(T)$, $X = \ln(C)$, $\alpha_0 = \ln(K)$ and $\alpha_1 = B$; simple linear regression equation is obtained. Cost and duration data is used in the analysis. Firstly, cost data is expressed in thousand United States Dollars (USDs). Then, duration data is expressed in days by multiplying month values with 30. After that, ln values of cost and duration is analyzed as X and Y in the regression equation with 95% confidence level. Regression results are summarized in Table 4.5.

Table 4.5 Regression results of Model 1

α_0	3,0622
α_1	0,3170
Multiple R	0,77
R^2	0,59
Adjusted R^2	0,56

ln(K) is represented by α_0 in the simple linear regression equation. On the other hand, the constant B is represented by α_1. Significance levels (P-values) observed for regression constant (α_0) and regression coefficient of ln(C) were 0,00035 and 0,00031, respectively. Therefore, significance of regression constant and independent variable ln(C) is verified and they are included in the final model. Then, it is concluded that the time-cost relationship for the 17 CCRC building can be expressed as follows:

$$T = 21C^{0,32} \qquad [4.7]$$

4.2.2 Prediction Performance Test

To evaluate the performance of duration predictions, steps developed in Part 3.4 are used. Predicted cost values observed from the cost model are used to predict durations of each projects in the test groups. In all of the five new models which are developed with the data excluding one test group each time, duration (T) data is expressed in terms of days and cost (C) data is expressed in terms of thousands USD. Therefore, durations are estimated in terms of days. Then, these estimates are divided by 30 to obtain values in months. Percentage Error (PE) for the comparison of actual durations and predicted durations is defined as follows:

$$PE = \frac{\text{predicted duration} - \text{actual duration}}{\text{actual duration}} \qquad [4.8]$$

MAPE is calculated by the equation 4.4. Actual durations and predicted durations are compared in Table 4.6.

Table 4.6 Actual duration vs. predicted duration (Model 1)

Project No	Actual Duration (in months)	Predicted Duration (in months)	PE	|PE|
1	21	17,3	-17,7	17,7
2	12	16,0	33,0	33,0
3	16	14,7	-8,2	8,2
4	14	14,6	4,4	4,4
5	13	13,8	6,1	6,1
6	16	13,8	-13,8	13,8
7	14	18,5	32,3	32,3
8	16	11,7	-26,6	26,6
9	18	16,7	-7,2	7,2
10	10	11,7	16,9	16,9
11	18	19,2	6,8	6,8
12	20	18,0	-10,2	10,2
13	17	16,6	-2,3	2,3
14	17	17,4	2,1	2,1
15	16	16,0	-0,1	0,1
16	20	18,1	-9,7	9,7
17	14	17,7	26,6	26,6
			MAPE	**13,2**

4.3 MODEL 2

Model 2 is developed in order to establish the linear relationship between the cost and duration of building projects by using simple linear regression analysis. The cost and duration data of 17 CCRC projects is analyzed and a linear regression model is established to predict the construction project duration from cost.

4.3.1 Data Analysis

In order to determine the model that establishes the relationship between the duration and the cost, simple linear regression analysis is employed. Firstly, a simple linear equation is written as follows:

$$Y = \alpha_0 + \alpha_1 X \tag{4.9}$$

Where;

Y = actual duration (T)

α_0 = regression constant

α_1 = partial regression coefficient of detailed cost (C)

X = detailed cost (C)

Regression results obtained by the analysis of cost and duration data are summarized in Table 4.7.

Table 4.7 Regression results of Model 2

α_0	10,46614
α_1	2,91E-07
Multiple R	0,74
R^2	0,55
Adjusted R^2	0,52

Regression analysis results revealed that regression constant (α_0) and regression coefficient of cost (α_1) were significant with P-values 1,66E-06 and 0,00064 for α_0 and $\alpha_{1,}$ respectively. The resulting linear equation was in the following form:

$$T = 10,47 + 2,91\text{E-}07\ C \tag{4.10}$$

4.3.2 Prediction Performance Test

Steps described for prediction performance assessment in Part 3.4 are followed again. Five new models similar to equation 4.10 are developed excluding the test samples. Then, predicted cost values obtained from the cost model are used in the duration models to predict the durations of each project in the test groups. Actual durations, predicted durations and prediction performance measures are given in Table 4.8.

Table 4.8 Actual duration vs. predicted duration (Model 2)

Project No	Actual Duration (in months)	Predicted Duration (in months)	PE	|PE|
1	21	17,5	-16,6	16,6
2	12	15,7	30,5	30,5
3	16	14,6	-8,8	8,8
4	14	14,5	3,7	3,7
5	13	13,9	6,6	6,6
6	16	13,6	-15,3	15,3
7	14	18,5	32,3	32,3
8	16	12,0	-25,2	25,2
9	18	16,4	-8,9	8,9
10	10	12,7	26,6	26,6
11	18	20,2	12,2	12,2
12	20	18,2	-8,8	8,8
13	17	16,3	-4,1	4,1
14	17	17,3	1,7	1,7
15	16	15,8	-1,1	1,1
16	20	18,2	-9,2	9,2
17	14	17,7	26,7	26,7
			MAPE	**14,0**

4.4 MODEL 3

A neural network with a back-propagation algorithm is utilized to capture the relationship between the construction cost and duration of building projects in this model.

4.4.1 Neural Network Models

Two feed-forward artificial neural networks are developed (Figure 3.1). In the neural network models, the input layer consisted of one variable which was construction cost. Similarly, the output layer consisted of construction duration. Hidden layer included 3 nodes in Model 3a and 6 nodes in Model 3b. Since there is no exact rule for determining the optimum number of hidden nodes, 3 (# of input variables + # of output variables + 1) and 6 nodes are used to determine the number of nodes.

Similar to other time-cost models which are developed by regression analysis, the cost and duration data of 17 CCRC projects are used to develop a model with artificial neural network technique. The steps are summarized as follows:

1) The data is prepared in Microsoft Excel including the project numbers, construction costs and durations of each project. Then, the data information is saved as Microsoft Excel Worksheet 3.0 which is transferred to the NetMaker.

2) NetMaker is executed and 'Manipulate Data' option in NetMaker is selected to open the data information saved in Excel 3.0 format. The column including the project numbers is labeled as annotation, the column including the construction cost data is labeled as input, and the column including the duration data is labeled as pattern (output). Then,

BrainMaker files including training fact file, testing file and definition file are created.

3) By default, NetMaker chooses a number of projects and incorporates the data information of them into testing file. Data information of projects which are chosen randomly to be used in testing file is not included in the training fact file created by NetMaker. However, it is aimed to develop a neural network model in which all the cost and duration data of 17 CCRC building projects are taken into consideration in this study. Testing of the model will be then performed separately through prediction performance test. Therefore, training fact file and testing file are opened with Microsoft Excel and data information included in the testing file is added to training fact file.

4) Definition file is opened in the BrainMaker. Firstly, number of hidden nodes is assigned from the 'Connections' menu. Then, automatic heuristic learn rate is chosen for learning rate tuning at 'Learning Setup' window in the 'Parameters' menu. Default values are used for training control flow. Finally, the network is trained by choosing 'Train Network' from the 'Operate' menu.

After the training operation is completed, network model can be used to predict the durations from cost. Resulting model can be saved as a network file which is created by BrainMaker. Two different network models are developed as described. Same procedure is followed in developing the two models. The only difference was the number of hidden nodes.

4.4.2 Prediction Performance Test

Prediction performance tests are performed with the procedure described in Part 3.4 for both Model 3a and Model 3b. Cost and duration data regarding the projects involved in each test group is eliminated from training fact file created by NetMaker. The eliminated data is added to testing file. As described in step 3 of part 4.4.1, NetMaker chooses a number of projects and incorporates the data information of them into testing file. Since data information of these projects is not included in the training fact file, they are added to training fact file again manually. After the data information to be included in the training fact file and testing file is arranged properly, procedure described in step 4 of Part 4.1.1 is performed through BrainMaker. Since there are five test groups, five new test models are developed with BrainMaker for each network model 3a and network model 3b. BrainMaker can write the predicted durations of all projects in a test group as an output file, once the test model is developed. This can be accomplished if running facts are read from testing file and the trained network is run. However, this method is not utilized to predict durations since detailed cost information is included in the testing files. Predicted cost values should be used to predict the construction durations. Therefore, predicted cost values regarding the projects in each test group are used to predict the durations in new models developed for that test groups. Cost values are inputted manually and predicted duration values corresponding to these cost values are read from the display screen. Slightly smaller MAPE value is obtained for Model 3b as comparison made with the MAPE value of Model 3a. So, Model 3b will be used as the final neural network model. Prediction performance test results of Model 3 are shown in Table 4.9.

Table 4.9 Actual duration vs. predicted duration (Model 3)

Project No	Actual Duration (in months)	Predicted Duration (in months)	PE	|PE|
1	21	17,4	-17,0	17,0
2	12	16,7	38,8	38,8
3	16	14,5	-9,5	9,5
4	14	14,5	3,2	3,2
5	13	13,9	6,6	6,6
6	16	13,3	-16,6	16,6
7	14	18,6	32,7	32,7
8	16	11,9	-25,6	25,6
9	18	17,3	-3,6	3,6
10	10	13,9	39,5	39,5
11	18	19,3	7,0	7,0
12	20	18,1	-9,5	9,5
13	17	17,3	1,6	1,6
14	17	17,5	3,2	3,2
15	16	15,9	-0,5	0,5
16	20	18,3	-8,7	8,7
17	14	17,9	28,1	28,1
			MAPE	**14,8**

4.5 MODEL 4

This model is developed to predict the construction duration of building projects by using multiple linear regression technique and the parameters (excluding cost) that are impacting the duration.

4.5.1 Data Analysis

The technique of multiple linear regression analysis is utilized to determine the relationship between the building parameters shown in Table 4.1 and the construction duration of the building.

The first regression model (RM1) including all the variables shown in Table 4.1 is described by the following equation:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \alpha_7 X_7 + \alpha_8 X_8 + \alpha_9 X_9 + \alpha_{10} X_{10} \quad [4.10]$$

where;

Y = actual duration (T)

α_0 = regression constant

$\alpha_{1,2,3,\ldots10}$ = partial regression coefficient of $X_{1,2,3,\ldots,10}$

X_1 = Area

X_2 = NoF

X_3 = Area/unit

X_4 = Per(C+H)

X_5 = Per(P)

X_6 = St multiplied by the area

X_7 = Mas multiplied by the area

X_8 = RC multiplied by the area

X_9 = Pre multiplied by the area

X_{10} = W multiplied by the area

Equation 4.10 is very similar to equation 4.1 which is formed for the cost model described in part 4.1. The only difference is the dependent variable 'Y'. The procedure performed in the data analysis of cost model is followed to develop parametric duration estimation model as well. Elimination of insignificant variables in regression models is again established by using the significance level (P-value) and coefficient of determination (R^2) to achieve parsimonious model. The relationship between the duration and real value independent variables are also determined by correlation coefficient. The equation for the correlation coefficient was as follows (x and y are the sample means):

$$\text{CORREL(X,Y)} = \frac{\sum_{i=1}^{17}(x - x_i)(y - y_i)}{\sqrt{\sum_{i=1}^{17}(x - x_i)^2 \sum_{i=1}^{17}(y - y_i)^2}} \qquad [4.11]$$

Correlation between duration and Area, NoF, Area/unit, Per(C+H), Per(P) were 0,67, 0,49, 0,30, 0,12, 0,29, respectively. Regression results and the independent variables eliminated from the regression models are summarized in Table 4.10.

Table 4.10 Summary of regression results for duration models

Model	Independent variables in the regression equation	R^2	Variable corresponding to the coefficient with highest P-value	P-value of the coefficient
RM1	Area, NoF, Area/unit, Per(C+H), Per(P), St, Mas, RC, Pre, W	0,721	St	0,790
RM2	Area, NoF, Area/unit, Per(C+H), Per(P), Mas, RC, Pre, W	0,717	Per(C+H)	0,763
RM3	Area, NoF, Area/unit, Per(P), Mas, RC, Pre, W	0,713	RC	0,647
RM4	Area, NoF, Area/unit, Per(P), Mas, Pre, W	0,705	Area	0,548
RM5	NoF, Area/unit, Per(P), Mas, Pre, W	0,692	Pre	0,553
RM6	NoF, Area/unit, Per(P), Mas, W	0,681	Per(P)	0,242
RM7	NoF, Area/unit, Mas, W	0,636	Area/unit	0,359

Firstly, the independent variable St is eliminated from RM1. Regression model 2 (RM2) is developed with the remaining nine independent variables including Area, NoF, Area/unit, Per(C+H), Per(P), Mas, RC, Pre and W. Then, the independent variable Per(C+H) is eliminated from RM2. This elimination of independent variables is continued until all the P-values of partial regression coefficients corresponding to independent variables included in the regression model were below 10%. The independent variables retained in RM8, regression coefficients and the P-values of the coefficients corresponding to these variables are demonstrated in Table 4.11.

Table 4.11 P-values for regression model RM8

Independent variable	Partial coefficients	P-value of the coefficient
NoF	2,313	0,002
Mas	9,93E-05	0,039
W	1,52E-04	0,031

Regression constant (α_0) for RM8 is found to be 4,935. Then, the resulting linear equation for estimating construction project duration from independent variables NoF, Mas and W is defined as follows:

$$T = 4{,}935 + 2{,}313 \times NoF + 9{,}93E\text{-}05 \times Mas + 1{,}52E\text{-}04 \times W \qquad [4.12]$$

4.5.2 Prediction Performance Test

The validity of the regression model is assessed in terms of predictive accuracy. Similarly, prediction performance test is performed with the steps described in Part 3.4. Five new models are developed without including the data of NoF, Mas and W which were selected in the test sample. Then; values of NoF, Mas and W in the test sample are used in the regression models to predict the construction durations. The prediction performance test results are shown in Table 4.12.

Table 4.12 Actual duration vs. predicted duration (Model 4)

Project No	Actual Duration (in months)	Predicted Duration (in months)	PE	|PE|
1	21	18,2	-13,6	13,6
2	12	14,7	22,4	22,4
3	16	15,3	-4,1	4,1
4	14	12,9	-8,1	8,1
5	13	14,0	7,5	7,5
6	16	14,1	-12,0	12,0
7	14	15,1	8,1	8,1
8	16	14,3	-10,4	10,4
9	18	14,9	-17,3	17,3
10	10	13,3	32,8	32,8
11	18	19,2	6,8	6,8
12	20	18,5	-7,3	7,3
13	17	15,2	-10,6	10,6
14	17	19,5	15,0	15,0
15	16	14,5	-9,5	9,5
16	20	24,1	20,7	20,7
17	14	21,5	53,3	53,3
			MAPE	**15,3**

4.6 MODEL 5

Neural network models using back-propagation algorithm are utilized to capture the relationships between the construction duration and the parameters given in Table 3.2. The main advantage of the neural network models is their capability to capture the non-linear relations as well as linear relations. Twenty network models are developed, and prediction performances of each model are assessed.

4.6.1 Neural Network Models

The first two network models are developed by including all of the independent variables given by Table 3.2 as inputs. Ten independent variables including Area, NoF, Area/unit, Per(C+H), Per(P), St, Mas, RC, Pre, and W are represented by ten nodes in the input layer. Construction duration is represented by one node in the output layer. Hidden layer included 12 nodes in one model (NM1a) and 6 nodes in the other model (NM1b). The numbers of input variables used in developing the models are decreased by sensitivity analysis. The number of hidden nodes are changed accordingly. There is no exact rule for determining the optimum number of hidden nodes. Two different numbers of hidden nodes are used for the models developed with same input variables with the purpose of comparing prediction performances of each model. The following equations are used to determine the number of hidden nodes to be used in the network models:

$$\text{Number of hidden nodes (a)} = \text{number of inputs} + \text{number of outputs} + 1 \quad [4.13]$$

$$\text{Number of hidden nodes (b)} = \frac{(\text{Number of hidden nodes (a)})}{2} \quad [4.14]$$

The results given by equation 4.14 are rounded to the upper value. Number of outputs is always '1' since construction duration is the only output in all of the network models. Network models, number of input variables and the number of hidden nodes used are listed in Table 4.13.

Table 4.13 Number of hidden nodes used in network models

Network Model	Number of input nodes	Number of hidden nodes	
		a	b
NM1	10	12	6
NM2	9	11	6
NM3	8	10	5
NM4	7	9	5
NM5	6	8	4
NM6	5	7	4
NM7	4	6	3
NM8	3	5	3
NM9	2	4	2
NM10	1	3	2

The procedure performed in developing the first two models (NM1a and NM1b) are summarized as follows:

1) The data given by Table 3.2 is listed in a Microsoft Excel file. Then, the data information is saved as Microsoft Excel Worksheet 3.0 which is recognizable by the NetMaker.

2) NetMaker is opened and 'Manipulate Data' option is selected to open the data information saved in Excel 3.0 format. The column including the project numbers is labeled as annotation, the column including the construction duration data is labeled as pattern (output). All the

remaining columns including the data information of ten independent variables are labeled as input. Then, BrainMaker files including training fact file, testing file and definition file are created.

3) Training fact file and testing file are opened with Microsoft Excel and data information incorporated into the testing file is added back to the training fact file through copy-paste option offered by Microsoft Excel. Thus, inclusion of all 17 CCRC projects' data into the training fact file is achieved.

4) Definition file created by NetMaker is opened with BrainMaker after the procedures described in the first three steps are performed. With the BrainMaker, number of hidden nodes is assigned from the 'Connections' menu. Then, automatic heuristic learn rate is chosen for at 'Learning Setup' window in the 'Parameters' menu. Default values are used for training control flow. Finally, the network is trained by choosing 'Train Network' from the 'Operate' menu and the developed model is saved as network file by BrainMaker.

The only difference encountered in the application of procedure described above for developing the NM1a and NM1b was the number of hidden nodes assigned. The NM1a is developed with 12 hidden nodes. On the other hand, the NM1b is developed with 6 hidden nodes. The resulting models saved as network file can be used to predict the durations.

Same procedure summarized above in the four steps are followed to develop the network models NM2, NM3, NM4, NM5, NM6, NM7, NM8, NM9 and NM10.

As described, the first neural network model (NM1) developed to predict construction duration of building projects included all of the independent variables. Then, the subsequent network models are developed by eliminating the least sensitive independent variable from the preceding network model. The least sensitive variable in the network model is identified by sensitivity analysis. Influence of each variable on the behavior of the selected network is evaluated by varying the value of each variable incrementally while keeping the values of remaining variables constant and observing the output durations. Increment values for the real-value variables including Area, NoF, Area/unit, Per(C+H) and Per(P) are calculated by dividing the difference of maximum and minimum values of each variable with 20. So, 20 increments are made for each real-value variables during sensitivity analysis. Values of variables corresponding to types of structural frames are not varied incrementally. Only '0' and '1' is tested to see the effect of structural frame types on the duration of building project while values of real-value variables are kept constant to their averages. RC is chosen to be the reference frame. Thus, as value of one of the real-value variable is varied, the values of the remaining four real-value variables are kept constant to their averages, and only value for RC is inputted as '1' while values for other structural frame types are inputted as '0'. All the input values, either increments for real-value variables or '0' and '1' values for variables corresponding to structural frame types, are inserted as inputs manually. Consequently, resulting duration outputs are read from the BrainMaker's display screen. Sensitivities of real-value variables are calculated with the maximum and minimum durations which are chosen from the twenty one output duration results observed for each increments corresponding to each variable. Two output duration results were observed for each structural frame type. One of them is observed for input '0' and the other is observed for input '1'. Therefore, sensitivities of structural frame types are calculated by using these two output duration values. The following equation is used for calculating sensitivity (S):

$$S = \frac{(\text{maximum output duration}) - (\text{minimum output duration})}{(\text{minimum output duration})} \times 100 \qquad [4.15]$$

Sensitivities of independent variables for each network model are shown in Table 4.14.

Table 4.14 Sensitivities of independent variables for each network model

Independent variables	Sensitivities								
	NM1b	NM2b	NM3a	NM4a	NM5a	NM6b	NM7b	NM8b	NM9a
Area	59,1	61,7	74,8	54,7	13,8	72,1	50,1	66,0	55,3
Per(P)	27,8	1,1	13,8	27,2	7,9	26,5	3,3	72,5	**9,6**
Per(C+H)	10,2	11,4	12,3	11,7	21,1	17,1	23,4	**60,8**	
St	20,0	16,8	39,9	19,0	37,5	35,1	**1,5**		
W	20,0	15,2	20,1	15,6	26,9	**3,5**			
RC	18,6	12,1	36,0	16,3	**0,2**				
NoF	5,3	12,5	47,8	**10,1**					
Pre	9,3	47,8	**1,9**						
Mas	52,3	**0,0**							
Area/unit	**3,9**								

It should be noted that; out of the two network models including same variables but developed with different number of hidden nodes, the one which offers a better prediction performance is chosen for sensitivity analysis.

4.6.2 Prediction Performance Test

Prediction performances of network models are also tested through MAPE described by equation 4.4. Since there were five test groups (See Table 3.5), five new test models are developed for each twenty network models. The procedure followed for the development of test models was similar to the procedure performed for the network models developed with the data of all projects. The main difference was the forming of testing file and training fact file which are created by NetMaker. As mentioned, NetMaker automatically chooses a number of projects and incorporates the data of them into testing file. Data of projects which are chosen randomly to be used in testing file is not included in the training fact file created by NetMaker. Training fact file and testing file which were created with the data of 17 building projects for test models are opened with Microsoft Excel. Firstly, data information of projects incorporated into the testing file is cut from the testing file, and pasted to the training fact file. Then, data information of projects involved in one of the test groups is cut from training fact file and pasted to the testing file. For example, data information of project 5, project 12 and project 15 are included in the testing files of first test models of any network model. After the cut-paste operation, empty rows in the training fact file are deleted and both the training fact file and testing file are saved. Then, BrainMaker is opened with definition file of the test model and number of hidden nodes corresponding to the network model to be tested is assigned. Automatic heuristic learn rate is adjusted and the data information of projects included in the training fact file is trained. Thus, development of the test model is finalized and saved as a network file which is created by BrainMaker.

Durations of projects included in a test group are predicted from the test model developed without using the data of projects corresponding to that test group. Predicted values of 17 projects are obtained by collecting the predicted durations of each project, one by one, from the five test models. This collection of predicted durations of 17 projects from the test models is repeated for each twenty network models similarly. Since the data of independent variables included in the testing file of test models is used to predict the durations, predicted durations of projects in a test group are obtained by an output file which is created by BrainMaker with the application of following procedure:

1) Test model which was saved as a network file is opened by BrainMaker.

2) From the 'File' menu, 'Select Fact Files' is chosen.

3) In the 'Fact Files' screen displayed, 'Read Running Facts' option is selected. Then, name of the testing file including the projects of which predicted durations are to be observed is printed into the dialog box. Thus, reading of running facts from the corresponding test file is achieved.

4) From the 'File' menu, 'Write Facts to File' is chosen. 'Open File' is selected after the name of the output file is chosen.

5) From the 'Operate' menu, 'Run Trained Network' is chosen.

6) Finally, output file is opened with Microsoft Excel and predicted duration values are read.

Smallest MAPE value (15,2%) is observed for NM10a. Therefore, NM10a is considered to be the best parametric network model and is consequently chosen to represent Model 6. Network models, independent variables used in the models and MAPEs calculated for each of the twenty model are shown in Table 4.15.

Table 4.15 MAPE results for twenty network models developed

Model	Independent variables in the network model	Prediction Performances	
		MAPE a	MAPE b
NM1	Area, NoF, Area/unit, Per(C+H), Per(P), St, Mas, RC, Pre, W	20,7	**19,3**
NM2	Area, NoF, Per(C+H), Per(P), St, Mas, RC, Pre, W	21,6	**19,0**
NM3	Area, NoF, Per(C+H), Per(P), St, RC, Pre, W	**28,8**	31,1
NM4	Area, NoF, Per(C+H), Per(P), St, RC, W	**18,5**	22,5
NM5	Area, Per(C+H), Per(P), St, RC, W	**20,9**	26,6
NM6	Area, Per(C+H), Per(P), St, W	20,1	**19,6**
NM7	Area, Per(C+H), Per(P), St	26,5	**22,0**
NM8	Area, Per(C+H), Per(P)	19,2	**16,2**
NM9	Area, Per(P)	**18,8**	20,9
NM10	Area	**15,2**	15,4

Actual durations, predicted durations and prediction performance measures observed for NM10a are listed in Table 4.16.

Table 4.16 Actual duration vs. predicted duration (Model 5)

Project No	Actual Duration (in months)	Predicted Duration (in months)	PE	|PE|
1	21	17,3	-17,8	17,8
2	12	16,6	38,7	38,7
3	16	14,6	-8,8	8,8
4	14	14,7	4,9	4,9
5	13	13,9	7,0	7,0
6	16	13,3	-16,9	16,9
7	14	17,9	28,0	28,0
8	16	12,5	-22,1	22,1
9	18	17,3	-4,0	4,0
10	10	13,9	39,1	39,1
11	18	20,4	13,2	13,2
12	20	17,9	-10,6	10,6
13	17	18,0	6,1	6,1
14	17	17,0	-0,1	0,1
15	16	15,2	-4,8	4,8
16	20	18,4	-7,8	7,8
17	14	17,9	28,0	28,0
			MAPE	**15,2**

4.7 DISCUSSION OF RESULTS

In this chapter, data analyses performed for cost model, model 1, model 2, model 3, model 4 and model 5 are described in detail and results are presented. In this section, the predictive accuracy of the models is compared along with the discussion of the results.

- *Cost Model* - Regression analysis performed with the data of 17 continuing care retirement community (CCRC) building projects revealed that construction cost values of these buildings can be predicted with an accuracy range of ± 20%. The independent variables included in the final cost model were Area, NoF, St, RC and Pre. The R^2 value of 0,973 for the final regression model indicated that the majority of variability in the construction cost is explained by these five independent variables. Inclusion of area into the regression model as a significant variable was not unexpected since area affects construction cost significantly. The minus signs observed for the partial coefficients of RC and Pre indicated that contributions of these variables to the cost explained by this cost model are negative. The sign observed for the partial coefficient of St was positive. This is reasonable since steel structures are generally expected to be more expensive than reinforced concrete and precast structure, especially for the buildings with a few number of floors.

- *Model 1* – Analysis of 17 CCRC building projects confirmed that the time and cost have a relationship in the form of '$T=KC^B$'. The final model expressed by $T=21C^{0,32}$ had a prediction performance of 13,2%. Durations are estimated within a range of ± 33% accuracy.

- *Model 2* – Linear model established by simple regression analysis had a prediction performance of 14%. Duration estimations are varied within an accuracy range of ± 33%.

- *Model 3* – The average absolute accuracy which is defined by MAPE was 14,8% for this network model. Duration estimations observed from this model are varied within an accuracy range of ± 40%.

- *Model 4* – The independent variables included in the final regression model were NoF, Mas and W. Unexpectedly, it was found that the area variable did not have a significant influence on project duration. One of the reasons behind this could be due to the fact that some of the relations between the area and duration may have been included by the interaction variables (including area), that were used to model the structure type. On the other hand, inclusions of variable masonry and variable wood with partial coefficients of positive sign indicated that buildings constructed with masonry and wood take longer as compared to buildings with reinforced concrete, steel and precast. However, this can be explained by the longer durations required to complete villa type projects with wood as compared to building project of same serviceable area with reinforced concrete or steel. This regression model had a prediction performance of 15,2%. Duration estimations are varied within an accuracy range of ± 33% for 16 projects.

- *Model 5* – The independent variable that remained in the final network model through sensitivity analysis was 'Area'. This outcome is reasonable since gross floor area is one of the most important factors that affect construction duration of building project. The best network model in terms of prediction performance was that final network with an absolute average accuracy of 15,2%. Duration estimations observed from this network model are varied within an accuracy range of ± 40%. Duration estimates observed from this model for each project were very similar with the estimates observed from network model 3 which is developed with cost parameter only. This can be attributed to the significant correlation between the area and cost.

Prediction performances for all conceptual duration estimation models are shown in Table 4.17.

Table 4.17 Prediction performances of duration models

Model	Explanation	Analysis Technique	MAPE
Model 1	BTC Model	Simple Linear Regression	13,2
Model 2	Time-Cost Model	Simple Linear Regression	14,0
Model 3	Time-Cost Model	Artificial Neural Network	14,8
Model 4	Parametric Model	Multiple Linear Regression	15,3
Model 5	Parametric Model	Artificial Neural Network	15,2

Paired t-test is performed in order to test whether the differences between MAPEs of duration estimation models are significant or not. For the first test, predictions corresponding to Model 1 and Model 4 are paired and compared. For the second test, predictions corresponding to Model 2 and Model 4 are paired and compared. The hypothesis that there is no difference in the MAPEs of these paired models is tested for α=0.05. Ott (1988) defined t by the following equation:

$$t = \frac{d - D_o}{s_d / \sqrt{n}} \qquad [4.16]$$

where d and s_d are the sample mean and standard deviation of the n differences, and D_o is the difference between MAPEs of the paired models. "n" is the number of observations which is 17 for both paired test models.

Calculations are performed automatically through Microsoft Excel and shown in Table 4.18 and Table 4.19.

Table 4.18 t-Test: Paired Model 1-Model 4 for means

	Model 1	Model 4
Mean	13,1674	15,2730
Variance	113,6709	146,8806
t Stat	-0,7007	
t Critical one-tail	1,7459	

For df = n-1 = 16 and α=0.05, the critical *t-value* for a one-tailed test is 1,7459. Since the observed value of *t*, -0,7007, is smaller than the critical *t-value* for a one-tailed test, the hypothesis that there is no difference in MAPEs of Model 1 and Model 4 is not rejected.

Table 4.19 t-Test: Paired Model 2-Model 4 for means

	Model 2	Model 4
Mean	14,0135	15,2730
Variance	109,1400	146,8806
t Stat	-0,4452	
t Critical one-tail	1,7459	

Similarly, the observed value of *t*, -0,4452, is smaller than the critical *t-value* for a one-tailed test. Therefore, the hypothesis that there is no difference in MAPEs of Model 2 and Model 4 is also not rejected. Thus, it is claimed by the results of paired t-test that the differences between MAPEs of duration estimation models are not significant.

CHAPTER 5

CONCLUSIONS

The main objective of this study was to develop reasonably accurate and practical methodologies for conceptual duration estimation of building projects through regression and artificial neural network techniques. Within this context, five different modeling approaches were used to predict the construction durations at the early stages of the projects. The first model consisted of a power regression model that expressed the duration in terms of cost. The second model also expressed the duration in terms of cost but was a linear regression model. In the third model, neural networks were used to capture the relation between the duration and cost. In the fourth and fifth models, cost was not used as an independent variable, but other parameters were used to predict the construction durations. Fourth model was a linear regression model and fifth model was a neural network model. Based on the model results, the following conclusions are identified:

- This study has shown that the models developed provide alternative methodologies for estimating the construction durations of buildings at the early stages of the projects. All the models can be used to predict durations with reasonable accuracy. The models developed in this study provide an alternative to the current duration estimation practice which is largely based on the individual experience of the estimator.

- No significant differences were observed between the time-cost models and the parametric models in terms of predictive accuracy. All the models provided reasonably accurate duration estimates. Predictive performances of time-cost models were slightly better than the prediction performances of the parametric models. However, use of parametric models can also be encouraged for duration estimations since parametric models remove the necessity of cost estimation required for time-cost models.

- In this study, regression analysis and artificial neural network techniques were utilized for the developments of parametric estimation models. The findings also revealed that there are no significant differences in the prediction accuracy of the regression and neural network models. Therefore, it can be concluded that linear regression analysis provides an adequate and pragmatic methodology for duration estimation of construction projects.

- The data used in the development of parametric duration estimation models was limited to the historical project data of 17 CCRC buildings. Sample size of 17 can be regarded as sufficient to develop duration estimation models since most of the contractors do not have large sets of historical data available. Thus, constructors can develop their own estimation models with the methodologies presented in this study. Consequently, contractors can evaluate whether the contract periods stipulated by the clients are reasonable or not, without performing detailed scheduling.

- This study showed that by using the data 17 building projects, an accuracy of 13-15% can be achieved for conceptual estimation of construction durations.

Following recommendations are offered for future studies:

- The estimation models were developed using a sample of 17 CCRC buildings. Therefore, further testing is recommended with the data from more building projects of similar type to improve the reliability the models.

- The parametric model performances can also be improved by considering other factors that may influence the duration of building projects such as project environment, productivity of the workforce and quality level required. However, it should be noted that factors which may influence the duration should be quantified at the early stages of the project.

- The methods presented in this study can be used to develop for other type of projects such as; industrial, process, highway, etc.

REFERENCES

Bhokha, S. and Ogunlana, S.O. (1999). Application of artificial neural network to forecast construction duration of buildings at the predesign stage. *Engineering, Construction and Architectural Management, 6*, 133-144.

Boussabaine, A.H. (1996). The use of artificial neural networks in construction management: a review. *Construction Management and Economics, 14*, 427-436

BrainMaker Professional (1993). *User's Guide and Reference Manual*. 4th ed. California Scientific Software, Nevada City, CA.

Bromilow, F.J. (1974). Measurement and scheduling of construction time and cost performance in the building industry. *The Chartered Builder, 10*, 57.

Bromilow, F.J., Hinds, M.F., and Moody, N.F. (1980). AIQS survey of building contract time performance. *Building Economics, 19*, 79–82.

Chan, A.P.C. (1999). Modelling building durations in Hong Kong. *Construction Management and Economics, 17*, 189-196.

Chan, A.P.C. (2001). Time-cost relationship of public sector projects in Malaysia. *International Journal of Project Management, 19*, 223-229.

Chan, A.P.C. and Chan, D.W.M. (2004). Developing a benchmark model for project construction time performance in Hong Kong. *Building and Environment, 39*, 339-349.

Kaka, A. and Price, A. D. F. (1991). Relationship between value and duration of construction projects. *Construction Management and Economics, 9*, 383–400.

Kumaraswamy, M.M. and Chan, D.W.M. (1995). Determinants of construction duration. *Construction Management and Economics, 13*, 209-217.

Kumaraswamy, M.M. and Chan, D.W.M. (1995). A study of the factors affecting construction durations in Hong Kong. *Construction Management and Economics, 13*, 319-333.

Kumaraswamy, M.M. and Chan, D.W.M. (1999). Modelling and predicting construction durations in Hong Kong public housing. *Construction Management and Economics, 17*, 351-362.

Love, P.E.D., Tse, R.Y.C. and Edwards, D.J. (2005). Time-cost relationships in Australian building construction projects. *Journal of Construction Engineering and Management, 131*, 187-194.

Ng, T., Mak, M.M.Y., Skitmore, M., and Varnam, M. (2001). The predictive ability of Bromilow's time–cost model. *Construction Engineering and Management, 19,* 165-173.

Nkado, R.N. (1992). Construction time information system for the building industry. *Construction Managements and Economics, 10*, 489-509.

Ogunsemi, D.R. and Jagboro, G.O. (2006). Time-cost model for building projects in Nigeria. *Construction Management and Economics, 24*, 253-258.

Ott, L. (1988). *An introduction to statistical methods and data analysis*. 3rd ed. PWS-KENT, Boston.

Smith, M. (1993). *Neural Networks for Statistical Modeling*. Van Nostrand Reinhold, New York.

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

Waier, P.R., Akins, T.J., Chiang, J.H., Ferguson, J.H., Kaplan, M.H., Jr., Mewis, R.W., Mossman, M.J., Moylan, J.J., Murphy, J.D., and Page, J.R. (*Editors*). (1996). *Building construction cost data 1995*. 54th ed. R.S. Means Company, Kingston, Mass.

Walker, D.H.T. (1995). An investigation into construction time performance. *Construction Management and Economics, 13,* 263-274.